Comparing Predictive Accuracy in the Presence of a Loss Function Shape Parameter*

Sander Barendse
Oxford University

Andrew J. Patton
Duke University

November 6, 2019

Abstract

We develop tests for out-of-sample forecast comparisons based on loss functions that contain shape parameters. Examples include comparisons using average utility across a range of values for the level of risk aversion, comparisons of forecast accuracy using characteristics of a portfolio return across a range of values for the portfolio weight vector, and comparisons using a recently-proposed "Murphy diagrams" for classes of consistent scoring rules. An extensive Monte Carlo study verifies that our tests have good size and power properties in realistic sample sizes, particularly when compared with existing methods which break down when then number of values considered for the shape parameter grows. We present three empirical illustrations of the new test.

Keywords: Forecasting, model selection, out-of-sample testing, nuisance parameters.

J.E.L. codes: C53, C52, C12.

*We thank Dick van Dijk, Erik Kole, Chen Zhou, and seminar participants at Oxford University for valuable discussions and feedback. The first author also acknowledges financial support from the Erasmus Trustfonds. All errors remain our own. Correspondence to: Sander Barendse, Nuffield College, New Road, Oxford, UK. Email address: `sander.barendse@economics.ox.ac.uk`.

1 Introduction

Forecast comparison problems in economics and finance invariably rely on a loss function or utility function, and in many cases these functions contain a shape parameter, for example, when comparing two forecasting models based on average utility. In many cases, there is no single specific value for the shape parameter that is of interest, rather there is a *range* of values that are of interest to the researcher. In this case the null hypothesis of equal predictive accuracy has a continuum of testable implications, but most existing work instead tests equal predictive accuracy at a few ad hoc values of the shape parameter. This paper combines work on forecast comparison tests, see Diebold and Mariano (1995) and Giacomini and White (2006), with bootstrap theory for empirical processes, see Bühlmann (1995), to provide forecast comparison tests that allow for inference across a range of values of a loss function shape parameter.

A leading example of an application that depends on some arbitrary parameter is a test of equal expected utility in which the utility function is parameterized by a risk aversion parameter. Given that economists have not converged on what value of risk aversion is appropriate (see, e.g., Bliss and Panigirtzoglou, 2004, for discussion) it is desirable to consider a null hypothesis of equal expected utility over a range of reasonable risk aversion values, instead of testing at some single value. Current practice usually evaluates the hypothesis of equal expected utility at one or a select few risk aversion parameter values, see Fleming et al. (2001), Marquering and Verbeek (2004), Engle and Colacito (2006) and DeMiguel et al. (2007) for example.

Another application that involves a continuum of testable implications is when one evaluates forecasts from multivariate models on the basis of their implied forecasts of univariate quantities, for example, evaluating a multivariate volatility model through its forecasts of Value-at-Risk for portfolios of the underlying assets. In practice, comparisons of quantile forecasts of portfolio returns, as generated from multivariate models, often only consider the equal weighted portfolio or some other fixed combination of portfolio constituents, see McAleer and Da Veiga (2008), Santos et al. (2012) and Kole et al. (2017) for example. Considering only a single weight vector can fail to reveal the sensitivity, or the robustness, of the ranking of two models to the choice of weight vector.

A final, recent, example is the comparison of forecasts using "elementary scoring rules," see Ehm et al. (2016). These authors show that the family of loss functions (or "scoring rules") that are consistent for a given statistical functional (e.g., the mean, a quantile, an expectile) can be represented as a convex combination of elementary scoring rules. The plot of the elementary scoring rules is called a "Murphy diagram," and Ehm et al. (2016) note that joint testing of the Murphy diagram is not yet fully developed. Ziegel et al. (2017) introduce tests for Murphy diagrams based on controlling the family-wise error rate but such corrections can perform poorly in large-scale multiple testing problems, see Hand (1998) and White (2000). Moreover, their tests consider only a finite subset of the testable implications instead of a continuum.

This paper develops new out-of-sample tests for multiple testing problems over a continuum of shape parameters, including the above three examples, which do not rely on bounds such as the Bonferroni correction, and which take into account the time series nature of the data used in most forecasting settings. To the best of our knowledge, such tests have not been considered in the literature to date. We consider tests of equal predictive ability (two-sided tests) and superior predictive ability (one-sided tests), and provide a framework for forecast dominance between two models based on one-sided tests that use opposing null hypotheses. We derive our tests using the supremum or average of Diebold-Mariano test statistics for each value of the shape parameter in its range, and obtain critical values using the moving blocks bootstrap of Bühlmann (1995), which is applicable to weakly dependent empirical processes indexed by classes of functions, and is general enough to cover our cases of interest: loss functions parameterized by a vector that can take values in a bounded subset of Euclidian space.

Our tests build on the out-of-sample testing framework of Diebold and Mariano (1995) and West (1996). Similar to Giacomini and White (2006) we consider evaluating the forecasting *method*, which, in addition to the forecasting model, includes the estimation scheme and choices of in-sample and out-of-sample periods. Our framework has some similarities with the work of White (2000) and Hansen (2005), who present predictive ability tests to compare a benchmark model with (finitely) many alternative models using a single loss function. In contrast, we consider only two models but we compare them using a continuum of loss functions. As such, our testing problem cannot be addressed using the work of those papers.

Recently, Jin et al. (2017) and Post et al. (2018) introduced multiple comparison tests for general loss functions over relatively large classes of loss functions (although the loss function should be minimized at zero), by translating the multiple hypothesis into hypotheses about stochastic dominance and "nonparametric forecast optimality," respectively. This allows the authors to develop tests based on the empirical distribution function or empirical likelihood. Our paper differs from this strand of research in two ways. First, we consider classes of loss functions that are economically motivated, whereas the previous authors abstract away this choice. In many settings, such as the ones considered in this paper, it is realistic to assume the researcher has at least some knowledge of what constitutes a relevant loss, and thus what values of the shape parameter are relevant. Ruling out economically uninteresting loss functions from the general set of loss functions can improve test power. Second, our paper applies to scenarios in which the loss function is not minimal at zero forecast error whereas the previous papers do not. Such loss functions include, for instance, many utility functions as well as the Fissler and Ziegel (2016) loss function for Value-at-Risk and Expected Shortfall.

We show via an extensive simulation study that the proposed testing methods have good size and power properties in realistic sample sizes. These positive results stand in stark contrast to the two most familar existing methods used to compare forecasts across a range of loss function parameter values: we find that the Wald test has finite-sample size as high as 50% for a 5% level test, while tests based on a Bonferroni correction suffer, unsurprisingly, from low power.

We consider three empirical applications of the proposed new tests. Firstly, in a comparison of expected utility of equal weighted and minimum-variance portfolio strategies (see, e.g., DeMiguel et al., 2007) we show that our tests are able to reject the null of equal average utility when existing alternative methods cannot. Secondly, we consider a tests of portfolio quantile forecasts generated by multivariate GARCH-DCC and RiskMetrics modelsand we find that our tests again are able to detect violations of the null hypothesis where existing methods cannot. Finally, we consider tests based on the Murphy diagrams for quantile forecasts generated by GARCH and RiskMetrics models, an application where existing methods simply cannot be applied due to the nature of the testing problem.

The remainder of the paper is structured as follows. In Section 2 we discuss our three illustrative examples. In Section 3 we present the general testing framework and develop our tests. In Section 4 we use Monte Carlo experiments to study the small sample properties of our tests in settings close to our illustrative examples. In Section 5 we explore these settings empirically. Section 6 concludes.

2 Loss function shape parameters in practice

We consider three representative examples of forecast comparison scenarios. In all of these examples we consider loss differences defined as

$$L_{t+1}(\gamma) = S(Y_{t+1}, g_t^{(1)}(\gamma); \gamma) - S(Y_{t+1}, g_t^{(2)}(\gamma); \gamma) \tag{1}$$

where S is the loss function (or scoring rule), $\gamma \in \Gamma \subset \mathbb{R}^d$ is the shape parameter of the loss function, Y_{t+1} is the target variable, and $g_t^{(i)}(\gamma)$ is the forecast of Y_{t+1} made using model i, which may depend on the shape parameter γ.

A test of uniform equal predictive accuracy across all values of the shape parameter is:

$$H_0 : E\left[L_{t+1}(\gamma)\right] = 0 \ \forall \ \gamma \in \Gamma \tag{2}$$

$$\text{vs. } H_1 : E\left[L_{t+1}(\gamma)\right] \neq 0 \text{ for some } \gamma \in \Gamma \tag{3}$$

We will also consider tests of uniform superior predictive ability, which consider the hypotheses:

$$H_0' : E\left[L_{t+1}(\gamma)\right] \leq 0 \ \forall \ \gamma \in \Gamma \tag{4}$$

$$\text{vs. } H_1' : E\left[L_{t+1}(\gamma)\right] > 0 \text{ for some } \gamma \in \Gamma \tag{5}$$

2.1 Comparisons based on expected utility

Two forecasting models each generate forecasts of optimal portfolio weights and we seek to compare them in terms of out-of-sample average utility from the resulting portfolio returns. The portfolio returns are obtained as $Y_{t+1}' g_t^{(i)}(\gamma)$, where Y_{t+1} is a vector of returns on the underlying assets, and $g_t^{(i)}(\gamma)$ is the forecasted optimal portfolio weights from model i assuming a preference parameter γ, and per-period utility is computed using some utility function $u(\cdot;\gamma)$. For instance, when $u(\cdot;\gamma)$ is the exponential utility function γ denotes the (scalar) risk aversion parameter, and $\Gamma = [a, b]$, for some $0 < a < b < \infty$. In this case we set

$$S(Y_{t+1}, g_t^{(i)}(\gamma); \gamma) = -u(Y_{t+1}' g_t^{(i)}(\gamma); \gamma) \tag{6}$$

so that lower values of the scoring rule indicate better performance.

2.2 Multivariate forecast comparison based on portfolio characteristics

Let Y_{t+1} denote some vector of returns, and compute portfolio returns as $\tilde{Y}_{t+1}(\gamma) = \gamma' Y_{t+1}$ for some weight vector γ. We are interested in forecasting some statistic ψ_t of $\tilde{Y}_{t+1}(\gamma)$ for all portfolio weights $\gamma \in \Gamma$. When we consider all long-only portfolios with weights summing to one, the parameter space Γ is the unit simplex. If we consider α-quantile forecasts of the portfolio return, then we may use the "tick" loss function to measure forecast performance and so set

$$S(Y_{t+1}, g_t^{(i)}(\gamma,\alpha);\gamma,\alpha) = (\mathbf{1}\{\gamma' Y_{t+1} \leq g_t^{(i)}(\gamma,\alpha)\} - \alpha)(g_t^{(i)}(\gamma,\alpha) - \gamma' Y_{t+1}). \tag{7}$$

where $\mathbf{1}\{\mathcal{A}\}$ equals one if $\mathcal{A}$ is true and zero otherwise.

2.3 Forecast comparisons via Murphy diagrams

Let Y_{t+1} denote some scalar return, and let ξ_t denote some statistic of Y_{t+1}, such as a mean or quantile. If ξ_t is elicitable, see Gneiting (2011a), then there exists a family of scoring rules (loss functions), $\mathcal{S}$ such that for any scoring rule $S^* \in \mathcal{S}$ it holds

$$E[S^*(Y_{t+1},\xi_t)] \leq E[S^*(Y_{t+1},x)],\ \forall x \in \mathcal{X} \tag{8}$$

where $\mathcal{X}$ is the support of ξ_t. That is, ξ_t has lower expected loss than any other forecast, x, for all scoring rules $S^* \in \mathcal{S}$. The scoring rule S^* is then said to be "consistent" for the statistic ξ_t. Many statistics, such as the mean, quantile, and expectile, admit *families* of consistent scoring functions, see Gneiting (2011a). For example, the mean is well-known to be elicitable using the quadratic loss function, and more generally it is elicitable using any "Bregman" loss function. The α-quantile is elicitable using the tick loss function in equation (7), and more generally using any "generalized piecewise linear" (GPL) loss function, see Gneiting (2011b). Comparisons of forecasts are usually done using a *single* consistent scoring rule (e.g., using mean squared error to compare estimates of the mean), however Patton (2018) shows that in the presence of parameter estimation error or misspecified models the ranking of forecasts can be sensitive to the specific scoring rule used.

In a recent paper, Ehm et al. (2016) show that any scoring rule that is consistent for a quantile or an expectile (the latter nesting the mean as a special case) can be represented as:

$$S^*(Y_{t+1}, x) = \int_{-\infty}^{\infty} \tilde{S}(Y_{t+1}, x; \gamma) dH(\gamma) \tag{9}$$

for some non-negative measure H, and $\gamma \in \Gamma \subset \mathbb{R}$, where $\tilde{S}$ is an "elementary" scoring rule, defined below. A plot of the average elementary scores across all values of γ is called a "Murphy diagram" by Ehm et al. (2016). If one forecast's Murphy diagram lies below that of another forecast for all values of γ, then it has lower average loss for *any* consistent scoring rule.

With the above representation of a consistent scoring rule as a mixture of elementary scoring rules, we can consider rankings across *all* consistent scoring rules, overcoming the sensitivity discussed in Patton (2018). For example, to compare α-quantile forecasts we set

$$S(Y_{t+1}, g_t^{(i)}(\alpha); \gamma, \alpha) = (\mathbf{1}\{Y_{t+1} < g_t^{(i)}\} - \alpha)(\mathbf{1}\{\gamma < g_t^{(i)}(\alpha)\} - \mathbf{1}\{\gamma < Y_{t+1}\}) \tag{10}$$

and then test for forecast equality/superiority across all $\gamma \in \mathbb{R}$. The right-hand side of the above equation is the quantile elementary scoring rule from Ehm et al. (2016).

3 Forecast comparison tests in the presence of a loss function shape parameter

Consider the stochastic process $W = \{W_t : \Omega \to \mathbb{R}^{N+s}, N \in \mathbb{N}_+, s \in \mathbb{N}, t = 1, 2, \ldots\}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$. We partition the observed vector W_t as $W_t = (Y_t, X_t)$, where $Y_t : \Omega \to \mathbb{R}^N$ is a vector a variables of interest and $X_t : \Omega \to \mathbb{R}^s$ is a vector of explanatory variables. We define $\mathcal{F}_t = \sigma(W_1, \ldots, W_t)$.

To fix notation, we let $|A| = (\mathrm{tr}(A'A))^{1/2}$ denote the Euclidean norm of a matrix A, and $\|A\|_q = (E|A|^q)^{1/q}$ denote the $\mathcal{L}_q$ norm of a random matrix. Finally, $\Rightarrow$ denotes weak convergence with respect to the uniform metric.

We denote the total sample size by T and the out-of-sample size by n. We consider moving or fixed window forecasts generated with in-sample periods of size m, such that the forecast for period $t+1$ is obtained using observations at periods $t - m + 1, \ldots, t$ with the moving scheme, and $1, \ldots, m$ with the fixed scheme, respectively.

We consider some (scalar) measurable loss (difference)

$$L_{t+1}(\gamma) = L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma) \tag{11}$$

that takes as arguments $m + 1 < \infty$ elements of W and some parameter vector $\gamma \in \Gamma \subset \mathbb{R}^d$ that is independent of W, with Γ a bounded set. As in Giacomini and White (2006), our assumption that $m < \infty$ imposes a limited memory condition on the forecasting methods, which precludes methods with model parameters estimated over expanding windows, but allows for those estimated over fixed and moving windows of finite length.

Below we describe the two-sided tests of equal predictive accuracy, and the one-sided tests of superior predictive accuracy.

3.1 Equal predictive ability tests

Here we consider tests of the following null and alternative hypotheses:

$$H_0 : E\left[L_t(\gamma)\right] = 0 \ \forall \ \gamma \in \Gamma \tag{12}$$

$$\text{vs. } H_1 : \left|E\left[L_t(\gamma^\dagger)\right]\right| \geq \Delta > 0 \text{ for some } \gamma^\dagger \in \Gamma \tag{13}$$

To develop a test of H_0 we consider the Diebold and Mariano (1995) test statistic as a function of $\gamma \in \Gamma$:

$$t_n(\gamma) \equiv \sqrt{n}\frac{\bar{L}_n(\gamma)}{\hat{\sigma}_n(\gamma)}, \tag{14}$$

with $\bar{L}_n(\gamma) \equiv \frac{1}{n}\sum_{t=m}^{T-1} L_{t+1}(\gamma)$, and $\hat{\sigma}_n^2(\gamma)$ denoting a (strongly) consistent estimator of $\sigma^2(\gamma) \equiv E[L_{t+1}(\gamma)^2]$.

It should be noted that when autocorrelation is present in $L_{t+1}(\gamma)$, $t_n(\gamma)$ does not converge in distribution to a standard normal limit, because $\hat{\sigma}_n^2(\gamma)$ is not a heteroskedasticity and autocorrelation consistent (HAC) estimator of the asymptotic covariance matrix of $\sqrt{n}\bar{L}_n(\gamma)$, see, e.g. Newey and West (1987)). We are not aware of strong uniform law of large numbers results for HAC estimators, which are required in our theoretical results below. As noted by Hansen (2005), $\hat{\sigma}_n^2(\gamma)$ is not required to be a consistent estimator of the variance of $\sqrt{n}\bar{L}_n(\gamma)$, because the bootstrap accounts for time series features in the data to obtain critical values of our tests. Indeed, in some scenarios it might be better not to studentize at all, and fix $\hat{\sigma}_n^2(\gamma) = 1$ instead.

Such scenarios include those for which $\hat{\sigma}_n^2(\gamma)$ may be close to zero in small samples.

To test H_0 we employ the test statistics

$$\sup t_n^2 \equiv \sup_{\gamma \in \Gamma} t_n^2(\gamma) \tag{15}$$

$$\operatorname{ave} t_n^2 \equiv \int_\Gamma t_n^2(\gamma) dJ(\gamma) \tag{16}$$

where J is a weight function on Γ. In the simplest case, we can set J to be uniform on Γ. Each of the above test statistics can be written as functions $v(t_n)$, where v maps functionals on Γ to $\mathbb{R}$ and we write $t_n = \{t_n(\gamma) : \gamma \in \Gamma\}$ as a random function on Γ. Each function v is continuous with respect to the uniform metric, monotonic in the sense that if $Z_1(\gamma) \leq Z_2(\gamma)$ for all γ then $v(Z_1) \leq v(Z_2)$, and has the property that if $Z(\gamma) \to \infty$ for γ for some subset of Γ with positive mass under weight function J, then $v(Z) \to \infty$.

We derive the asymptotic distribution of the test statistics above using the following assumptions.

Assumption 1. *$\{W_t\}$ is stationary and β-mixing (absolutely regular), with $\beta(t) = c_\beta a^t$, for some finite constant c_β, and $0 < a < 1$.*

Assumption 2. *$E[\sup_{\gamma \in \Gamma} |L_{t+1}|^{4r}] < \infty$, for some $r > 1$, and for all t.*

Assumption 3. *$\|L_{t+1}(\gamma) - L_{t+1}(\gamma')\|_{4r} \leq B|\gamma - \gamma'|^\lambda$, for some $B < \infty$, $\lambda > 0$, and for all $\gamma, \gamma' \in \Gamma$, and t.*

Assumption 4. *$\hat{\sigma}_n^2(\gamma) \xrightarrow{a.s.} \sigma^2(\gamma)$ uniformly over $\gamma \in \Gamma$. Moreover, $\inf_{\gamma \in \Gamma} \sigma^2(\gamma) > 0$.*

The β-mixing condition in Assumption 1, which is stronger than α-mixing, but weaker than ϕ-mixing, is usually assumed when deriving functional CLTs for time series data with unbounded absolute moments. Bühlmann (1995) notes that the mixing rate is satisfied for ARMA(p, q) processes with innovations dominated by the Lebesgue measure. Boussama et al. (2011) provides conditions under which multivariate GARCH models satisfy geometric ergodicity, and Bradley et al. (2005, Thm. 3.7) shows that geometric ergodicity implies β-mixing with at least exponential rate, satisfying Assumption 1. Assumption 2 is a standard moment condition, and Assumption 3 requires a Lipschitz condition to hold for these moments. Assumption 4 requires that $\hat{\sigma}_n^2(\cdot)$ satisfies a strong Uniform Law of Large Numbers (see, e.g., Andrews, 1992), and also imposes uniform non-singularity of $\sigma^2(\cdot)$.

Theorem 1. *Let Assumptions 1 to 4 be satisfied. Under H_0 it follows that $\sqrt{n}\bar{L}_n(\cdot) \Rightarrow Z(\cdot)$, for some Gaussian process $Z(\cdot)$ with covariance kernel $\Sigma(\cdot,\cdot) \equiv \lim_{n\to\infty} Cov\big(\sqrt{n}\bar{L}_n(\cdot), \sqrt{n}\bar{L}_n(\cdot)\big)$. Moreover, $v(t_n) \xrightarrow{d} v(\tilde{t})$, with $\tilde{t}(\cdot) \equiv Z(\cdot)/\sigma(\cdot)$.*

The following result establishes inference under the null and alternative hypotheses given in equations (12) and (13) above.

Theorem 2. *Let Assumptions 1 to 4 be satisfied, and let $\Sigma(\cdot,\cdot)$ be nondegenerate. Under H_0 it follows that $P\left(v(t_n) > c(1-\alpha)\right) \to \alpha = P\left(v(\tilde{t}) > c(1-\alpha)\right)$, where $c(1-\alpha)$ is chosen such that $P\left(v(\tilde{t}) > c(1-\alpha)\right) = \alpha$. Moreover, let $\Gamma^\dagger = \{\gamma : |\gamma - \gamma^\dagger|^\lambda < \Delta/B\}$ have positive J-measure. Under H_1 it follows that $v(t_n) \xrightarrow{d} \infty$, and $P\big(v(t_n) > c(1-\alpha)\big) \to 1$.*

We establish the consistency of the block bootstrap for general empirical processes of Bühlmann (1995) for $v(t_n)$. The block bootstrap was first studied by Künsch (1989) for general stationary observations.

The bootstrap counterpart of $\sqrt{n}\bar{L}_n(\gamma)$ is given by

$$\sqrt{n}\bar{L}_n^*(\gamma) \equiv \sqrt{n}\frac{1}{n}\sum_{t=m}^{T-1}\big(L_{t+1}^*(\gamma) - \mu_n^*(\gamma)\big), \tag{17}$$

where $\mu_n^*(\gamma) \equiv \frac{1}{n-l+1}\sum_{i=1}^{n-l+1}\frac{1}{l}\sum_{t=m+i}^{m+i+l-1} L_t(\gamma)$ denotes the expectation of $\frac{1}{n}\sum_{t=m}^{T-1} L_{t+1}^*(\gamma)$ conditional on the original sample, l denotes the block length, and $L_{t+1}^*(\gamma)$ denotes the bootstrap counterpart of $L_{t+1}(\gamma)$. Similarly, let $t_n^*(\gamma) \equiv \sqrt{n}\bar{L}_n^*(\gamma)/\hat{\sigma}_n(\gamma)$, and let $c_n^*(1-\alpha)$ denote the α-quantile of $t_n^*(\gamma)$.

We impose the following condition on the rate that $l = l(n) \to \infty$, as $n \to \infty$.

Assumption 5. *The block length l satisfies $l(n) = O(n^{1/2-\varepsilon})$, for some $0 < \varepsilon < 1/2$.*

The following result establishes consistency of the bootstrap.

Theorem 3. *Let Assumptions 1 to 5 be satisfied. Under H_0 it follows that $\sqrt{n}\bar{L}_n^*(\cdot) \Rightarrow Z(\cdot)$ almost surely. Moreover, $P(v(t_n) > c_n^*(1-\alpha)) \to \alpha$.*

Theorem 3 shows that we can estimate $c_n^*(1-\alpha)$ through simulation: let $c_n^B(1-\alpha)$ denote the $\alpha \cdot 100\%$ percentile of the test statistics $v(t_n^{*(1)}(\gamma)), \ldots, v(t_n^{*(B)}(\gamma))$ obtained from B bootstrap samples. As $B \to \infty$, $c_n^B(1-\alpha)$ becomes arbitrarily close to $c_n^*(1-\alpha)$.

3.2 Superior predictive ability tests

The results of the previous section allow us to additionally consider tests of uniform superior predictive ability, which consider the hypotheses:

$$H_0' : E\left[L_{t+1}(\gamma)\right] \leq 0 \ \forall \ \gamma \in \Gamma \tag{18}$$

$$\text{vs. } H_1' : E\left[L_{t+1}(\gamma^\dagger)\right] \geq \Delta > 0 \text{ for some } \gamma^\dagger \in \Gamma \tag{19}$$

Notice that H_0 in Equation (12) is the element of H_0' least favorable to the alternative. We can thus construct a valid test of H_0' from

$$\sup t_n \equiv \sup_{\gamma \in \Gamma} t_n(\gamma). \tag{20}$$

The results under H_0 in Theorems 1, 2, and 3 hold for H_0' for a demeaned version of $\sup t_n$, defined as

$$\sup \tau_n \equiv \sqrt{n} \frac{\bar{L}_n(\gamma) - E[\bar{L}_n(\gamma)]}{\hat{\sigma}_n(\gamma)}. \tag{21}$$

We must use $\sup \tau_n$ because H_0' allows for $E[\bar{L}_n(\gamma)] < 0$. Note that we can obtain valid critical values under H_0' from Theorem 3. Further, the result in Theorem 2 under H_1 also holds under H_1' for $\sup t_n$ (not just for $\sup \tau_n$), which implies that $\sup t_n$ has asymptotic power against fixed alternatives in this setting as well.

The above framework can also be used to derive a test based on $\inf t_n$. In this case the alternative hypothesis is that the benchmark model is worse than the competing model uniformly on Γ, i.e., under the alternative hypothesis the competing model dominates the benchmark uniformly. We do not include this test here, but, related, below we consider a framework for forecast dominance based on the employment of two one-sided tests.

3.3 Testing for forecast dominance

A rejection of the null hypothesis H_0' in equation (18) constitutes evidence against the benchmark model at some values of γ in Γ. However, it may be that the benchmark model is superior to the competing model for other points in Γ, and thus that neither model uniformly dominates the other.

To resolve this ambiguity we propose a framework using two one-sided tests in which we switch the role of benchmark and competing models between tests. In the following we denote the benchmark model as $\mathcal{B}$ and the competing model as $\mathcal{A}$. The first test is standard and applies to H_0', whereas the second test is applied to the *opposite* null hypothesis

$$H_0^{''} : E\left[-L_{t+1}(\gamma)\right] \leq 0 \;\forall\; \gamma \in \Gamma$$

and which is implemented simply by using test statistic $\sup t_n$.

Similar to forecast "encompassing" tests, see Chong and Hendry (1986), employing both tests results in one of the following four outcomes:

1. Reject H_0', fail to reject $H_0^{''}$: $\mathcal{A}$ significantly beats $\mathcal{B}$ for some values of γ, and is not significantly beaten by $\mathcal{B}$ for any γ. Thus $\mathcal{A}$ dominates $\mathcal{B}$.
2. Fail to reject H_0', reject $H_0^{''}$. Similar to outcome 1, but $\mathcal{B}$ dominates $\mathcal{A}$.
3. Fail to reject both H_0' and $H_0^{''}$. Neither model significantly beats the other for any value of γ. Thus the models have statistically equal performance across γ.
4. Reject both H_0' and $H_0^{''}$: There are values of γ for which $\mathcal{A}$ significantly beats $\mathcal{B}$, and values for which $\mathcal{B}$ significantly beats $\mathcal{A}$. Thus there is no ordering of the models across all γ.

Outcomes 1 and 2 clearly reveal a preferred model. Outcome 3 indicates a lack of power to distinguish between the competing models (or actual equality of forecast performance across γ). Outcome 4 reveals an important sensitivity in the ranking of the models to the choice of shape parameter.

3.4 Feasible implementation of the tests

When $\Gamma \in \mathbb{R}^d$ contains infinitely many elements, as in our main examples, it is not possible to evaluate the test statistics in equations (15), (16) and (20). Here we provide two numerical approximations to these test statistics for which Theorems 1 and 2 remain valid.

We first consider discretizations of Γ that become increasingly dense. Consider a grid of Γ with K_n elements Γ_n^i, such that $\sup_{\gamma,\gamma' \in \Gamma_n^i} |\gamma - \gamma| < \delta_n$, for all $i = 1, \ldots, K_n$, and let $\gamma_{n,i}$ be some point in Γ_n^i. We have the following approximations to our test statistics:

$$\widehat{\text{ave}\, t_n^2} \equiv \sum_{i=1}^{K_n} t_n^2(\gamma_{n,i}) \int_{\Gamma_n^i} dJ(\gamma), \tag{22}$$

$$\widehat{\sup t_n^2} \equiv \max_{i=1,\ldots,K_n} t_n^2(\gamma_{n,i}), \text{and} \tag{23}$$

$$\widehat{\sup t_n} \equiv \max_{i=1,\ldots,K_n} t_n(\gamma_{n,i}). \tag{24}$$

If Γ is a hyperrectangle in $\mathbb{R}^d$, a particularly convenient choice of J, K_n, and $\{\Gamma_n^i\}_{i=1}^{K_n}$ derives from partitioning each dimension of Γ in v_n equal parts, which results in $K_n = v_n^d$. Choosing J to be uniform over Γ gives $\int_{\Gamma_n^i} dJ(\gamma) = v_n^{-d}$, which is simple to implement. Other choices of J may require more involved algebra or simulations to obtain $\int_{\Gamma_n^i} dJ(\gamma)$.

The condition that $\delta_n \to 0$ implies that K_n grows quickly with large d. As a result the calculation of $\widehat{\text{ave}\, t_n^2}$, $\widehat{\sup t_n^2}$, and $\widehat{\sup t_n}$ becomes problematic for large d. In such cases one can instead use Monte Carlo simulations from J to obtain approximations of the test statistics. Consider S_n independent draws $\gamma^{(i)}$ from J, $i = 1, \ldots, S_n$, and the approximations:

$$\widetilde{\text{ave}\, t_n^2} \equiv \frac{1}{S_n} \sum_{i=1}^{S_n} t_n^2(\gamma^{(i)}), \tag{25}$$

$$\widetilde{\sup t_n^2} \equiv \max_{i=1,\ldots,S_n} t_n^2(\gamma^{(i)}), \text{and} \tag{26}$$

$$\widetilde{\sup t_n} \equiv \max_{i=1,\ldots,S_n} t_n(\gamma^{(i)}). \tag{27}$$

Proposition 1. *Let the assumptions of Theorem 1 hold. Moreover, let the weight function J be absolutely continuous. For some $K_n \to \infty$, such that $\delta_n \to 0$, as $n \to \infty$, $\widehat{v(t_n)} \xrightarrow{p} v(t_n)$. Moreover, $\widetilde{v(t_n)} \xrightarrow{p} v(t_n)$ for $S_n \to \infty$ as $n \to \infty$.*

3.5 Benchmark testing methods

For comparison with the proposed new tests, we consider standard joint Wald tests and the Bonferroni multiple-comparison correction as benchmarks to the tests proposed above. It should be noted that these tests can only be applied at a finite number of points, M, in Γ, and therefore cannot generally test over all Γ.

Consider some discrete parameter set $\Gamma_M = \{\gamma_1, \ldots, \gamma_M\} \subset \Gamma$. We define the standard Wald test statistic as

$$\hat{Q}_n^h \equiv n\tilde{L}_n(\Gamma_M)'\hat{\Omega}_{M,n}^{-1}\tilde{L}_n(\Gamma_M), \tag{28}$$

where $\tilde{L}_n(\Gamma_M) \equiv \left(\bar{L}_n(\gamma_1)', \ldots, \bar{L}_n(\gamma_M)'\right)'$, and $\hat{\Omega}_{M,n}$ is some HAC estimator of the asymptotic covariance matrix of $\sqrt{n}\tilde{L}_n(\Gamma_M)$, e.g., the estimator of Newey and West (1987).

A two-sided α-level test rejects the null hypothesis when $\hat{Q}_n^h > \chi^2_{M,1-\alpha}$, where $\chi^2_{M,1-\alpha}$ denotes the $(1-\alpha)$-quantile of a χ^2 distribution with M degrees of freedom. As M increases $\hat{\Omega}_{M,n}$ becomes near-singular, which can lead to erratic behavior in the test statistic; we observe such behavior in our simulation study below.

A two-sided α-level test using the Bonferroni correction rejects the null hypothesis if, for at least one $\gamma \in \Gamma_M$, we find $n\bar{L}_n^2(\gamma)/\tilde{\sigma}_n^2(\gamma) > \chi^2_{1,1-\alpha/M}$, with $\tilde{\sigma}_n^2(\gamma)$ a HAC asymptotic covariance estimator of $\sqrt{n}\bar{L}_n(\gamma)$. Tests using the Bonferroni correction are conservative for large M, and also whenever the test statistics are positively correlated.

A one-sided α-level test using the Bonferroni correction rejects the null hypothesis if, for at least one $\gamma \in \Gamma_M$, we find $n\bar{L}_n(\gamma)/\tilde{\sigma}_n(\gamma) > z^{-1}_{1-\alpha/M}$, with $z^{-1}_{1-\alpha/M}$ denoting the $(1-\alpha/M)$-quantile of the standard normal distribution. As usual for Bonferroni-corrected tests, the critical value is much larger than for the individual tests ($z^{-1}_{1-\alpha/M}$ rather than $z^{-1}_{1-\alpha}$) which can lead to low power; we observe this in our simulation study below. We do not consider one-sided Wald tests as a benchmark here, although they may be constructed using the methods of Wolak (1989).

4 Simulation studies

In this section we evaluate the finite-sample performance of our proposed tests in the three applications described in Section 2. In Sections 4.1–4.3 below we describe the simulation designs for each of these applications, and in Section 4.4 we present the results.

4.1 Comparisons based on expected utility

We consider the difference in expected utility of two commonly-used portfolio management strategies: the equal-weighted portfolio and the minimum-variance portfolio. For an in-depth analysis of these and other portfolio strategies see DeMiguel et al. (2007). These strategies can be defined in terms of portfolio weight vectors:

$$
\begin{aligned}
w_{t+1}^{\mathrm{eq}} &= \frac{1}{N}\iota, \\
w_{t+1}^{\mathrm{mv}} &= \frac{1}{\iota'\Sigma_{t+1}^{-1}\iota}\Sigma_{t+1}^{-1}\iota,
\end{aligned}
\tag{29}
$$

where Y_{t+1} is an $(N \times 1)$ vector of monthly excess returns with conditional covariance matrix Σ_{t+1}, and ι is an $(N \times 1)$ vector of ones. Denote the portfolio returns as $Y_{t+1}^{\mathrm{eq}} = w_{t+1}^{\mathrm{eq}}{}'Y_{t+1}$, and $Y_{t+1}^{\mathrm{mv}} = w_{t+1}^{\mathrm{mv}}{}'Y_{t+1}$. The feasible counterpart of w_{t+1}^{mv} depends on an estimate of Σ_{t+1}. In our simulation study we consider a simple rolling window estimate of the covariance matrix based on the most recent 120 observations, corresponding to 10 years of monthly returns data.

We test whether the equal weighted and minimum-variance portfolio returns have equivalent expected utility across a range of levels of risk aversion. We model utility using the exponential utility function $u(y;\gamma) = -\exp\{-\gamma y\}/\gamma$. A wide range of values of the risk aversion parameter have been reported in the literature, ranging from near zero to as high as 60, see Bliss and Panigirtzoglou (2004). These authors estimate this parameter as being between 2.98 to 10.56, while DeMiguel et al. (2007) perform comparisons for investors with risk aversion ranging between 1 and 10. Based on this, we test for equal expected utility over $\Gamma = [1, 10]$. We draw uniformly over this range when implementing the ave-test.

We simulate excess returns Y_{t+1} according to a one-factor model, based on the DGP in the simulation study of DeMiguel et al. (2007). Let $Y_{t+1} = (Y_{t+1}^f, Y_{1,t+1}, \ldots, Y_{N-1,t+1})'$, where Y_{t+1}^f denotes the excess return on the factor portfolio and $Y_{i,t+1}$ denotes the $N-1$ excess returns, generated as:

$$
\begin{aligned}
Y_{i,t+1} &= \alpha_i + \beta_i Y_{t+1}^f + \eta_{i,t+1}, \\
\nu_{i,t+1} &\sim \mathrm{iid}\,\mathcal{N}(0, \sigma_{\eta,i}^2), \\
Y_{t+1}^f &\sim \mathrm{iid}\,\mathcal{N}(\mu_f, \sigma_f^2).
\end{aligned}
\tag{30}
$$

We follow the parameterization of DeMiguel et al. (2007), which resembles estimates that are commonly found in empirical studies. We set $\alpha_i = 0$, and $\beta_i = 0.5 + (i-1)/(N-1)$, for all $i = 1, \ldots, N-1$. Moreover, we set $\mu_f = 8\%$, and $\sigma_f = 16\%$. Finally, we let the idiosyncratic volatilities vary between 10% and 30%. However, unlike DeMiguel et al. (2007), who draw from the uniform distribution on $[10\%, 30\%]$, we opt for deterministic cross-sectional variation between 10% and 30% by setting $\sigma_{\eta,i} = 10\% + 20\% \cdot \sin(\pi(i-1)/(N-1))$. We do so to facilitate the approximation of $E[L_{t+1}(\gamma)]$, which is required in the size experiment.

Given the portfolio strategies we consider, it is not generally possible to find a parameterization that implies $E[L_{t+1}(\gamma)] = 0$ for all $\gamma \in \Gamma$. In the size experiment we therefore test the null hypothesis $E[L_{t+1}(\gamma)] = \zeta_m(\gamma)$ instead of zero, where $\zeta_m(\gamma) \neq 0$ is the population value of $E[L_{t+1}(\gamma)]$, which we estimate using 100,000 simulations. The power experiment tests the null hypothesis $E[L_{t+1}(\gamma)] = 0$ for all $\gamma \in \Gamma$.

4.2 Forecast comparison via tail quantile forecasts of portfolio returns

We next study a scenario comparing quantile forecasts of portfolio returns implied by multivariate forecasting models. We simulate returns data using a GARCH-DCC model (Engle, 2002) with normal errors, parameterized to resemble the properties of daily asset returns.

We compare two widely-used models: (i) a GARCH-DCC model with normal errors, and (ii) a multivariate normal distribution with the RiskMetrics covariance estimator (Riskmetrics, 1996). We let the $N \times 1$ return vector Y_{t+1} follow a GARCH-DCC process

$$\begin{aligned}
Y_{t+1} &= \mu_{t+1} + H_{t+1}^{1/2} C_{t+1}^{1/2} \nu_{t+1}, \\
\nu_{t+1} &= (\nu_{t+1,1}, \ldots, \nu_{t+1,N})' \sim iid\, N(0, I), \\
H_{t+1} &= \operatorname{diag}\left(h_{t+1,1}, \ldots, h_{t+1,N}\right), \\
h_{t+1,i} &= \omega_0 + \omega_1 h_{t,i} + \omega_2 h_{t,i} \nu_{t,i}^2, \\
C_{t+1} &= \operatorname{diag}\left(\tilde{C}_{t+1}\right)^{-1/2} \tilde{C}_{t+1} \operatorname{diag}\left(\tilde{C}_{t+1}\right)^{-1/2}, \\
\tilde{C}_{t+1} &= (1 - \xi_1 - \xi_2)\bar{C} + \xi_1 \tilde{C}_t + \xi_2 \nu_t \nu_t', \\
\bar{C} &= [\bar{C}]_{ij}, \text{ where } [\bar{C}]_{ij} = 1 - \frac{|i-j|}{N}.
\end{aligned} \tag{31}$$

We choose GARCH parameters $\omega_0 = 0.05$, $\omega_1 = 0.10$, $\omega_2 = 0.85$ and DCC parameters $\xi_1 = 0.025$, $\xi_2 = 0.95$, to match time-varying volatility and correlation patterns commonly found in daily equity returns. We set $\mu_{t+1} = 0$ for simplicity. To fix the value of $\bar{C}$ we use the covariance matrix generated by a Bartlett kernel with bandwith set to N. This specification generates a diverse set of correlations, and ensures positive definiteness of $\bar{C}$.

We are interested in one-period-ahead α-quantile forecasts for portfolio returns $\tilde{Y}_{t+1}(\gamma) = \gamma' Y_{t+1}$, with $\alpha = 5\%$, for a range of portfolio weights $\gamma \in \Gamma$. The first forecast we consider is the optimal forecast, based on the GARCH-DCC model above. This forecast is given by

$$Q^{(1)}_{t+1,\alpha}(\gamma) = \Phi^{-1}(\alpha) \cdot \gamma' H^{1/2}_{t+1} C_{t+1} H^{1/2}_{t+1} \gamma, \tag{32}$$

where $\Phi^{-1}(\alpha)$ denotes the α-quantile of the standard normal distribution.

The RiskMetrics forecast is

$$Q^{(2)}_{t+1,\alpha}(\gamma) = \Phi^{-1}(\alpha) \cdot \gamma' \hat{\Sigma}_{t+1} \gamma, \tag{33}$$

$$\text{where } \hat{\Sigma}_{t+1} = c_{\lambda,m} \sum_{j=0}^{m} \lambda^j \left(Y_{t-j+1} - \hat{\mu}_t\right) \left(Y_{t-j+1} - \hat{\mu}_t\right)' \tag{34}$$

where $\hat{\mu}_t = \frac{1}{m} \sum_{j=1}^{m} Y_{t-j+1}$ and $c_{\lambda,m}$ is a constant that normalizes the summed weights $\sum_{j=0}^{m} \lambda^j$ to one. As is standard for the RiskMetrics approach using daily returns, we set $\lambda = 0.94$.

We obtain the scores (losses) $S^{(i)}_{t+1}(\gamma)$ using the tick loss function, which is a consistent loss function for the quantile, and is defined as

$$S^{(i)}_{t+1}(\gamma, \alpha) = \left(\mathbb{1}\left\{\tilde{Y}^p_{t+1}(\gamma) < Q^{(i)}_{t+1,\alpha}(\gamma)\right\} - \alpha\right)\left(Q^{(i)}_{t+1,\alpha}(\gamma) - \tilde{Y}^p_{t+1}(\gamma)\right), \tag{35}$$

and obtain loss differences $L_{t+1}(\gamma) = S^{(2)}_{t+1}(\gamma) - S^{(1)}_{t+1}(\gamma)$.

We consider all portfolio return vectors with positive weights summing to one, i.e. $\Gamma = \{\gamma : \gamma_i \geq 0, i = 1, \ldots, N, \sum_{i=1}^{N} \gamma_i = 1\}$. Γ is thus the $(N-1)$-simplex, and drawing uniformly from Γ is particularly easy using the Dirichlet distribution of order N with concentration parameters set to one. See Kotz et al. (2000, Ch. 49) for an elaborate treatment of the Dirichlet distribution.

As in the previous example, to study the finite-sample size of this test we consider the null hypothesis that $E[L_{t+1}(\gamma)] = \zeta_m(\gamma)$ instead of zero, where $\zeta_m(\gamma) \neq 0$ is the population value of $E[L_{t+1}(\gamma)]$, which we estimate using 100,000 simulations. The power experiment tests the null hypothesis $E[L_{t+1}(\gamma)] = 0$ for all $\gamma \in \Gamma$.

4.3 Forecast comparison via Murphy diagrams of quantile forecasts

Under mild regularity conditions (see, e.g., Gneiting, 2011a) the α-quantile of a random variable Y_{t+1} is elicitable using the "generalized piecewise linear" (GPL) class of scoring rules:

$$S(Y_{t+1}, x; \alpha, g) = \left(\mathbb{1}(Y_{t+1} < x) - \alpha\right)\left(g(x) - g(Y_{t+1})\right), \tag{36}$$

where $g(\cdot)$ is a non-decreasing function. A commonly used member of this family is the tick loss function, which sets $g(z) = z$, and which was used in the previous section. In this example we test for differences in predictive ability of competing α-quantile forecasts across the set of *all* consistent scoring rules for α-quantiles, denoted $\mathcal{S}^{\alpha}_{GPL}$, using the mixture representation for this class of loss functions presented in Ehm et al. (2016):

$$S(Y_{t+1}, x; \alpha, g) = \int_{-\infty}^{\infty} \tilde{S}(Y_{t+1}, x; \alpha, \gamma) dH(\gamma; g) \tag{37}$$

$$\text{where } \tilde{S}(Y_{t+1}, x; \alpha, \gamma) = \big(\mathbb{1}(Y_{t+1} < x) - \alpha\big)\big(\mathbb{1}(\gamma < x) - \mathbb{1}(\gamma < Y_{t+1})\big), \tag{38}$$

where $H(\cdot; g)$ some non-negative measure.

Consider two forecasts $Q^{(1)}_{t+1,\alpha}$ and $Q^{(2)}_{t+1,\alpha}$. We say that $Q^{(2)}_{t+1,\alpha}$ dominates $Q^{(1)}_{t+1,\alpha}$ if

$$\begin{aligned} E_t[L(Y_{t+1}, Q^{(1)}_{t+1,\alpha}, Q^{(2)}_{t+1,\alpha}, \alpha, g)] &\equiv E_t[S(Y_{t+1}, Q^{(2)}_{t+1,\alpha}, \alpha, g) - S(Y_{t+1}, Q^{(1)}_{t+1,\alpha}, \alpha, g)] \\ &\leq 0 \text{ for all } S \in \mathcal{S}^{\alpha}_{GPL} \end{aligned} \tag{39}$$

Corollary 1 in Ehm et al. (2016) establishes that $Q^{(2)}_{t+1,\alpha}$ dominating $Q^{(1)}_{t+1,\alpha}$ is implied by

$$\begin{aligned} E_t[\tilde{L}(Y_{t+1}, Q^{(1)}_{t+1,\alpha}, Q^{(2)}_{t+1,\alpha}, \alpha, \gamma)] &\equiv E_t[\tilde{S}(Y_{t+1}, Q^{(2)}_{t+1,\alpha}, \alpha, \gamma) - \tilde{S}(Y_{t+1}, Q^{(1)}_{t+1,\alpha}, \alpha, \gamma)] \\ &\leq 0 \text{ for all } \gamma \in \mathbb{R} \end{aligned} \tag{40}$$

We can therefore test for equal predictive ability by testing the above condition. It should be noted that our theory requires the range of γ, denoted Γ, to be bounded, and so we cannot test over all $\gamma \in \mathbb{R}$. However, in practice we can make Γ large enough to cover all relevant parameter values, since $\tilde{S}(Y_{t+1}, Q^{(2)}_{t+1,\alpha}; \alpha, \gamma) - \tilde{S}(Y_{t+1}, Q^{(1)}_{t+1,\alpha}; \alpha, \gamma)$ is known to be identically zero for $\gamma \notin \big[\min\big(Y_{t+1}, Q^{(1)}_{t+1,\alpha}, Q^{(2)}_{t+1,\alpha}\big), \max\big(Y_{t+1}, Q^{(1)}_{t+1,\alpha}, Q^{(2)}_{t+1,\alpha}\big)\big]$.

In small samples it can occur that $\tilde{S}(Y_{t+1}, Q^{(2)}_{t+1,\alpha}; \alpha, \gamma) - \tilde{S}(Y_{t+1}, Q^{(1)}_{t+1,\alpha}; \alpha, \gamma) = 0$ for all observations in a given sample, for some values of γ. As a result, $\hat{\sigma}^2_n(\gamma) = 0$ for these values of γ. To circumvent this we fix $\sigma^2_n(\gamma) = 1$ for all $\gamma \in \Gamma$, i.e. we consider sup- and ave-tests based on $t_n(\gamma) \equiv \sqrt{n}\bar{L}_n(\gamma)$ instead of $t_n(\gamma) \equiv \sqrt{n}\frac{\bar{L}_n(\gamma)}{\hat{\sigma}_n(\gamma)}$. The p-values remain valid under the bootstrap. The HAC covariance estimators used in the calculation of the multivariate Wald test and the Bonferroni-corrected test suffer from the same singularity. However, inference is no longer valid for these tests, because the limit law of these test statistics is no longer standard without studentization.

We use the same quantile forecast models as in the simulation design in the previous section, but set $N = 1$, so that the quantile forecasts defined in equations (32) and (33) are obtained from a GARCH model instead of a GARCH-DCC model. For additional comparison, we also consider a rolling window sample quantile estimated over the previous 250 days.

As in the previous examples, to study the finite-sample size of this test we consider the null hypothesis that $E[L_{t+1}(\gamma)] = \zeta_m(\gamma)$ instead of zero, where $\zeta_m(\gamma) \neq 0$ is the population value of $E[L_{t+1}(\gamma)]$, which we estimate using 100,000 simulations. The power experiment tests the null hypothesis $E[L_{t+1}(\gamma)] = 0$ for all $\gamma \in \Gamma$.

4.4 Simulation results

Table 1 presents rejection rates for the size and power experiments introduced in Section 4.1, based on 1,000 Monte Carlo simulations. We consider out-of-sample period lengths $n = 120$ and 600 observations, corresponding to 10 and 50 years of monthly returns data. We consider increasingly large grids of Γ, with the number of grid points set to $K_n = 1, 10, 50, 100$, and 250. We obtain critical values using $B = 1{,}000$ bootstrap samples.

From Panels A and B we observe that the (two-sided) sup-t^2 and ave-t^2 tests are somewhat oversized for $n = 120$ out-of-sample observations, and appropriately sized for $n = 600$. The one-sided sup-t test is also oversized for $n = 120$ but approximately correctly sized for $n = 600$. Reassuringly, we observe that the sup-t^2, ave-t^2 and sup-t tests are stable in terms of rejection rates once K_n is moderately large, indicating robustness to this tuning parameter.

The benchmark tests (based on a Wald test or a Bonferroni correction) perform poorly as K_n increases, with the Wald test having large size distortions, and the Bonferroni-corrected test becoming conservative. This sensitivity can be avoided by implementing these tests on just a single value of γ, though of course in that case the conclusion of the test is sensitive to the particular choice of γ.

Panels C and D shows rejection rates in the power experiment. We observe the sup- and ave-test rejection probabilities being stable once the value of K_n is moderately large. In contrast, but as anticipated, the Bonferroni-adjusted tests have power declining with K_n. The power of the Wald test in this application also declines with K_n, and we note that this test cannot be implemented when $n = 120$ and $K_n = 250$ as in that case the covariance matrix is singular.

[Table 1 about here.]

Table 2 presents small sample rejection rates of the size and power experiments introduced in Section 4.2, for portfolios composed of 30 assets. Results are presented for sample sizes of n = 500 and 2,000 observations, corresponding approximately to two and eight years of daily asset returns, and six sets of weight vectors, which are drawn as follows. We first consider a set of 31 deterministic weight vectors: the equal weighted portfolio weights and the 30 basis vectors. Subsequently we randomly draw $S_n - 31$ weight vectors from J, for $S_n = 50, 100, 250, 500$ and 1,000. We use B = 1,000 bootstrap samples to obtain critical values.

Panels A and B provide rejection rates in the size experiment. We observe that the sup-t^2, ave-t^2 and sup-t tests have reasonable size properties, although the tests are slightly conservative for n =2,000 (rejection rates between 1-2%). The same holds for the Bonferroni-corrected tests at all S_n. The benchmark Wald test is oversized for all S_n at n = 500, and for all S_n larger than 31 at n = 2,000, illustrating the problems with this test as S_n grows. As in the previous example, the rejection rates of the sup-t^2, ave-t^2 and sup-t tests are stable across S_n.

Panels C and D show rejection rates in the power experiment. As in the previous section, we observe the sup- and ave-test rejection probabilities being stable across values of S_n. We again observe the Bonferroni-adjusted tests having decreasing power as S_n increases. The power of the Wald test in this application increases, driven by its failure to control size as S_n increases.

[Table 2 about here.]

Tables 3 and 4 provide small sample rejection rates of the tests in size and power experiments introduced in Section 4.3. Table 3 compares GARCH forecasts with RiskMetrics forecasts, and Table 4 compares GARCH forecasts with rolling window forecasts. The sample sizes considered are n = 500 and n = 2,000, corresponding approximately to two and eight years of daily returns data. We consider grids of Γ with $K_n = 50, 100$, and 250 grid points equally spaced over the interval $[-20, 0]$. We select this interval because outside of it the elementary score differences are equal to zero in almost all realizations. As we cannot always (across values of γ in the elementary scores) compute the asymptotic covariance required to obtain the benchmark Wald and Bonferroni-corrected tests, we do not implement those tests here. Instead, we present the results of a standard Diebold-Mariano test based on the tick loss function as benchmark. These are reported in the rows labeled "1".

Panel A of Tables 3 and 4 provide rejection rates for the size experiments. In Table 3 we compare the GARCH and RiskMetrics forecasts and we observe that the sup-t^2 and ave-t^2 tests are correctly sized at both out-of-sample period lengths considered ($n = 500$ and 2,000). The one-sided sup-t test is slightly conservative at $n = 500$, but is close to nominal rates for $n =$ 2,000. The benchmark Diebold-Mariano test using the tick loss function, is also approximately correctly sized. The rejection rates of the sup- and ave-tests are stable across values K_n, indicating robustness to this tuning parameter. Panel A of Table 4 provides rejection rates for the size experiment where we compare the GARCH and the rolling window quantile forecasts. We find that the two-sided Diebold-Mariano test and the ave-t^2-test are oversized, whereas the sup-t^2 and sup-t tests have reasonable rejection rates.

Panel B of Table 3 provides rejection rates in the power experiment in which we compare the GARCH and RiskMetrics forecasts. We observe that this particular alternative is not very distant from the null, and the ave- and sup-tests have no power at $n = 500$. At $n = 2{,}000$ the proposed tests have power against the alternative, but the benchmark test is more powerful. Panel B of Table 4 provides rejection rates in the power experiment comparing the GARCH and rolling window quantile forecasts. The sup- and ave-tests reject more often, and have power properties similar to the benchmark Diebold-Mariano test.

[Table 3 about here.]

[Table 4 about here.]

5 Applications to equity return forecast environments

5.1 Comparing the utility of two portfolio strategies

In this section we compare equal weighted and minimum-variance portfolio strategies in terms of average utility, across a range of levels of risk aversion. We use monthly returns data on 30 U.S. industry portfolios, from September 1926 to December 2017, a total of 1,098 observations.[1] The minimum-variance portfolio weights are estimated over a rolling window of $m = 120$ months. We use the exponential utility function $u(y;\gamma) = -\exp\{-\gamma y\}/\gamma$, with $\gamma \in [1, 10]$, which covers all risk aversion parameter values considered in DeMiguel et al. (2007).

Table 5 provides p-values for the proposed new tests, as well as the benchmark Wald and Bonferroni-adjusted tests. The left panel presents results for tests of the null of equal predictive ability for these two portfolio strategies, and the tests for both one-sided null hypotheses. The

[1] The returns can be obtained from the data library at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

row labeled $K_n = 1$ tests the hypothesis of equal predictive ability for one particular value of risk aversion $\gamma = 2.5$, while the remaining rows use $K_n = 10, 50, 100, 250$ equally-spaced points in $[1, 10]$. We set B = 1,000.

In Panel A of Table 5 we consider $\Gamma = [1, 10]$. When $K_n = 1$, the null of equal predictive ability is rejected using all tests. We also find that the one-sided test rejects the null of weakly greater utility from the minimum variance portfolio compared with the equal-weighted portfolio and fails to reject the opposite null, indicating that the equal-weighted strategy dominates the minimum variance portfolio for this particular risk aversion. When we increase K_n, and test the hypothesis of equal average utility over the entirety of Γ, we find that the ave-t^2 test fails to reject the null, while the sup-t^2 test continues to reject the null. The one-sided tests reveal that the equal-weighted portfolio dominates the minimum variance portfolio across all values for the risk aversion parameter.

Looking at the benchmark tests, the Bonferroni-corrected test fails to reject the null hypothesis for larger values of K_n. This finding is consistent with the conservativeness of this test as K_n increases documented in the simuation study in the previous section. The Wald test rejects the null hypothesis for all values of K_n except $K_n = 250$, but since the simulation exercise shows the test has severe size distortions, we should not rely on conclusions drawn form the Wald test.

Figure 1 shows the sample mean of expected utility differences and the pointwise 95% confidence bounds for each $\gamma \in [1, 10]$. We observe that for γ less than about 3, the confidence interval does not include zero, indicating that for lower levels of risk aversion the equal-weighted portfolio significantly outperforms the minimum variance portfolio. For higher levels of risk aversion, the pointwise confidence intervals either contain zero, or lie below zero. Since the sup-t^2 test considers the largest deviation from the zero line, while the ave-t^2 considers (weighted) averages of all deviations, it is intuitive that the ave-t^2 test statistic indicates weaker evidence against the null due to the large subinterval of Γ for which the mean utility differential is insignificantly different from zero.

Panels B and C of Table 5 show results for $\Gamma = [1, 5]$ and $\Gamma = [5, 10]$, to examine the sensitivity of the conclusions of these tests to the range of values of risk aversion considered. For less risk averse investors, with $\Gamma = [1, 5]$, we see that the both the ave-t^2 and the sup-t^2 tests reject the null-hypothesis. If we consider more risk averse investors, with $\Gamma = [5, 10]$, we see that the null hypothesis cannot be rejected by either the ave- or the sup-tests. Conclusions drawn from the Bonferroni-corrected tests do not change much, with this test failing to reject the null hypothesis for larger values of K_n. From the one-sided test results we observe that the equal-weighted strategy dominates the minimum variance strategy over $\Gamma = [1, 5]$, while we cannot distinguish between the strategies over $\Gamma = [5, 10]$.

[Figure 1 about here.]

[Table 5 about here.]

5.2 Quantile forecasts of portfolio returns from multivariate models

In this analysis we compare two multivariate volatility models, a GARCH-DCC model and the RiskMetrics model, by the quality of their forecasts for the 5% Value-at-Risk (i.e., the 5% quantile) of the returns on portfolios of the underlying assets. We use daily returns on the same 30 U.S. industry portfolios as in the previous section, over the period January 1998 to December 2017, a total of 5,032 observations. The GARCH-DCC model is estimated over a rolling window of 1,000 observations. We set $B = 1{,}000$.

We consider the following sets of portfolios. For $S_n = 1$ we consider only the equal weighted portfolio. For $S_n = 31$ we consider the equal weighted portfolio and all 30 single-asset portfolios. For $S_n > 31$, we additionally consider $S_n - 31$ random weight vectors drawn from the 30-dimensional simplex.

Table 6 presents p-values of the tests of equal or superior predictive ability. We show results for the full sample, as well as the first and second halves of the sample period. The results for the full sample show that the models have approximately equal performance, since the null hypotheses of equal predictive ability cannot be rejected. Similarly, the one-sided superior predictive tests do not reject either null hypothesis, irrespective of which model is selected as benchmark. We note that that the sup- and ave-tests generate p-values that are stable for larger values of S_n, whereas the Bonferroni p-values approach one as S_n increases, consistent with the conservativeness of this test found in the simulation study. The Wald tests are unreliable for larger values S_n, as shown in the simulation exercises, and are not reported for $S_n > 31$.

In Panel B of Table 6 we find evidence against the null of equal predictive accuracy using the sup-t test, with p-values around 0.04 to 0.06. The one-sided test taking the GARCH-DCC model as benchmarks generates p-values of 0.02 to 0.04, indicating evidence that the RiskMetrics model significantly outperforms the GARCH-DCC model for part of the shape parameter space (i.e., for some portfolio weight vectors). On the other hand, the one-sided test taking the RiskMetrics model as benchmark does not reject the null, indicating that the RiskMetrics model dominates the GARCH-DCC model in this sub-sample. In the second sub-sample, similar to the full sample, we cannot reject the hypotheses of equal or superior predictive ability using any test.

Figure 2 plots the sample mean of the tick loss differences and the 95% confidence bounds for the first 100 portfolio weights that we draw, over the full sample and subsamples, and sorted on mean tick loss difference. In the first half of the sample we indeed find that the RiskMetrics forecasts perform better, although for only few portfolio vectors we observe (pointwise) confidence intervals that exclude zero. The sup-tests detect the outperformance of the RiskMetrics forecasts better than the ave-test as observed in Table 6. In the second half of the sample the average tick loss differences are generally negative, which is in agreement with the null of weakly better GARCH-DCC forecasts.

[Figure 2 about here.]

[Table 6 about here.]

5.3 Quantile forecast comparison via Murphy diagrams

Our final empirical analysis compares two forecast models for the 5% quantile (i.e., the 5% Value-at-Risk) of a single asset. We compare three models: a GARCH(1,1) model (Bollerslev, 1986), the RiskMetrics model, and a simple rolling window sample quantile calculated over the previous 250 days. We use the same data as the previous sub-section: daily returns on 30 U.S. industry portfolios, over the period January 1998 to December 2017, a total of 5,032 observations. The GARCH model is estimated over a rolling window of 1,000 observations. The parameter space for the elemetary scoring rule shape parameter (see Section 4.3 for details) is $\Gamma = [-20, 0]$, and we consider an increasingly fine grid of equally-spaced points in this space when implementing the new tests. We set $B = 5{,}000$.

We implement the tests for each of the 30 industry portfolio returns separately. In Table 7 we present detailed results for a single representative portfolio (the "Transportation" industry portfolio), and in Table 8 we present a summary of the results across all 30 industry portfolios.

Panel A of Table 7 compares RiskMetrics and GARCH forecasts. We find that the sup-t^2 test rejects the null of equal predictive accuracy, with p-values between 0.01 and 0.02. The ave-t^2 test has p-values around 0.075, indicating weaker evidence against the null of equal accuracy. The benchmark Diebold and Mariano (1995) test using the tick loss function fails to reject the null, with a p-value of 0.77. The one-sided sup-t tests reject the null of weakly superior GARCH forecasts with p-values around 0.03, but do not reject in the opposite direction, indicating that the RiskMetrics forecasts dominate the GARCH forecasts for the Transportation industry portfolio.

The upper panel of Figure 3 shows the "Murphy diagram" for this comparison, applied to the Transportation portfolio, and reveals that for most values of the elementary scoring rule parameter the GARCH and RiskMetrics forecasts have similar average losses. For values of the parameter around -1 the GARCH forecast significantly outperforms the RiskMetrics forecast, with the pointwise confidence intervals being far from zero, whereas the RiskMetrics forecast shows some (pointwise) significant outperformance for values of the parameter around -2. The results of the tests in Table 7, however, indicate that the RiskMetrics forecasts dominate the GARCH forecasts overall.

Panel B of Table 7 compares the GARCH forecast with the rolling window sample quantile forecast. We find p-values for the tests of equal accuracy are zero to two decimal places, indicating strong evidence against this null. The one-sided tests finds no evidence that the rolling window sample quantile outperforms the GARCH forecast for any value of elementary scoring rule parameter, whereas we do find evidence in the opposite direction, indicating that the GARCH forecast dominates the rolling window forecasts for the Transportation industry portfolio. The lower panel of Figure 3 shows that the difference in average loss is negative almost everywhere, consistent with the tests in Table 7, and the pointwise confidence intervals exclude zero for a large part of the parameter space.

Panel A of Table 8 compares Riskmetrics and GARCH forecasts across 30 industry portfolios and reports the proportion of portfolios for which we can reject the null at the 5% level. We see that the two-sided Diebold and Mariano (1995) test using the tick loss rejects the null of equal predictive accuracy for none of the 30 portfolios, while the ave-t^2 and sup-t^2 tests reject for around 23% and 10% of portfolios respectively. The right columns report forecast dominance outcomes based on the one-sided sup-t tests discussed in Section 3.3. Focusing on the K_n =1,000 row, we see that for 17% of the 30 portfolios either the RiskMetrics model dominates (10%) or the GARCH model dominates (7%), while for the remaining portfolios we are unable to statistically distinguish the performance of these two models. Consistent with this, results for the one-sided Diebold and Mariano (1995) tests using tick loss (not reported in the interests of space) do not reject for any portfolio in any direction.

Panel B of Table 8 reveals a clear ordering of the rolling window and GARCH forecasts. The two-sided tests reject the null of equal accuracy for all 30 portfolios. The one-sided tests reveal that the GARCH model provides uniformly superior forecasts to the rolling window model, for all 30 portfolios. One-sided Diebold and Mariano (1995) tests using the tick loss function lead to the same conclusion.

[Figure 3 about here.]

[Table 7 about here.]

[Table 8 about here.]

6 Concluding remarks

In many empirical applications, researchers are faced with the problem of comparing forecasts using a loss function that contains a shape parameter; examples include comparisons using average utility across a range of values for the level of risk aversion, and comparisons using characteristics of a portfolio return across a range of values for the portfolio weight vector. We propose new forecast comparison tests, in the spirit of Diebold and Mariano (1995) and Giacomini and White (2006), that may be applied in such applications. We consider tests of the null of equal forecast accuracy across all values of the shape parameter, against the alternative of unequal forecast accuracy for some value of the shape parameter. We also consider one-sided tests for superior forecast accuracy. The asymptotic properties of the test statistics are derived using bootstrap theory for empirical processes, see Bühlmann (1995).

We show via an extensive simulation study that the tests have satisfactory finite-sample properties, unlike the leading existing alternatives which break down when a large number of values of the shape parameter is considered. We illustrate the new tests in three empirical applications: comparing portfolio strategies using average utility across a range of levels of risk aversion; comparing multivariate volatility models via their Value-at-Risk forecasts for portfolios of the underlying assets across a range of values for the portfolio weight vector; and comparisons using recently-proposed "Murphy diagrams" (Ehm et al., 2016) for classes of consistent scoring rules for quantile forecasting.

References

Andrews, D. W. (1992). Generic Uniform Convergence. *Econometric Theory*, 8(2):241–257.

Andrews, D. W. (1994). Empirical Process Methods in Econometrics. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics, Volume 4*, pages 2247–2294. Elsevier.

Bliss, R. R. and Panigirtzoglou, N. (2004). Option-Implied Risk Aversion Estimates. *Journal of Finance*, 59(1):407–446.

Boussama, F., Fuchs, F., and Stelzer, R. (2011). Stationarity and Geometric Ergodicity of BEKK Multivariate GARCH Models. *Stochastic Processes and their Applications*, 121(10):2331–2360.

Bradley, R. C. et al. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, 2:107–144.

Bühlmann, P. (1995). The Blockwise Bootstrap for General Empirical Processes of Stationary Sequences. *Stochastic Processes and Their Applications*, 58(2):247–265.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric Evaluation of Linear Macroeconomic Models. *Review of Economic Studies*, 53:671–690.

Davydov, Y., Lifshits, M. A., and Smorodina, N. (1998). *Local Properties of Distributions of Stochastic Functionals*. American Mathematical Society.

DeMiguel, V., Garlappi, L., and Uppal, R. (2007). Optimal Versus equal weighted Diversification: How Inefficient is the 1/N Portfolio Strategy? *Review of Financial Studies*, 22(5):1915–1953.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

Doukhan, P., Massart, P., and Rio, E. (1994). The Functional Central Limit Theorem for Strongly Mixing Processes. *Annales de l'IHP Probabilités et statistiques*, 30(1):63–82.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance Principles for Absolutely Regular Empirical Processes. *Annales de l'I.H.P. Probabilités et Statistiques*, 31(2):393–427.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations and Forecast Rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):505–562.

Engle, R. (2002). Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. *Journal of Business and Economic Statistics*, 20(3):339–350.

Engle, R. and Colacito, R. (2006). Testing and Valuing Dynamic Correlations for Asset Allocation. *Journal of Business and Economic Statistics*, 24(2):238–253.

Fissler, T. and Ziegel, J. F. (2016). Higher Order Elicitability and Osband's Principle. *Annals of Statistics*, 44(4):1680–1707.

Fleming, J., Kirby, C., and Ostdiek, B. (2001). The Economic Value of Volatility Timing. *Journal of Finance*, 56(1):329–352.

Giacomini, R. and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica*, 74(6):1545–1578.

Gneiting, T. (2011a). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gneiting, T. (2011b). Quantiles as Optimal Point Forecasts. *International Journal of Forecasting*, 27(2):197–207.

Hand, D. J. (1998). Data Mining: Statistics and More? *American Statistician*, 52(2):112–118.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics*, 23(4):365–380.

Jin, S., Corradi, V., and Swanson, N. R. (2017). Robust Forecast Comparison. *Econometric Theory*, 33(6):1306–1351.

Kole, E., Markwat, T., Opschoor, A., and Van Dijk, D. (2017). Forecasting Value-at-Risk under Temporal and Portfolio Aggregation. *Journal of Financial Econometrics*, 15(4):649–677.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Kotz, S., Johnson, N. L., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions*. Wiley, New York.

Künsch, H. R. (1989). The Jackknife and the Bootstrap for General Stationary Observations. *Annals of Statistics*, 17(3):1217–1241.

Marquering, W. and Verbeek, M. (2004). The Economic Value of Predicting Stock Index Returns and Volatility. *Journal of Financial and Quantitative Analysis*, 39(2):407–429.

McAleer, M. and Da Veiga, B. (2008). Single-Index and Portfolio Models for Forecasting Value-at-Risk Thresholds. *Journal of Forecasting*, 27(3):217–235.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703.

Patton, A. J. (2018). Comparing Possibly Misspecified Forecasts. *Journal of Business and Economic Statistics*. Forthcoming.

Post, T., Potì, V., and Karabati, S. (2018). Nonparametric Tests for Superior Predictive Ability. *Available at SSRN 3251944*.

Riskmetrics (1996). JP Morgan Technical Document.

Santos, A. A., Nogales, F. J., and Ruiz, E. (2012). Comparing Univariate and Multivariate Models to Forecast Portfolio Value-at-Risk. *Journal of Financial Econometrics*, 11(2):400–441.

West, K. D. (1996). Asymptotic Inference About Predictive Ability. *Econometrica*, 64(5):1067–1084.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126.

White, H. (2001). *Asymptotic Theory for Econometricians*. Academic Press, Cambridge, MA.

Wolak, F. A. (1989). Testing Inequality Constraints in Linear Econometric Models. *Journal of Econometrics*, 41(2):205–235.

Ziegel, J. F., Krüger, F., Jordan, A., and Fasciati, F. (2017). Murphy Diagrams: Forecast Evaluation of Expected Shortfall. *arXiv preprint arXiv:1705.04537*.

A Mathematical appendix

A.1 Proof of Theorem 1

Finite dimensional convergence of $\sqrt{n}\bar{L}_n(\cdot)$ follows from a CLT for (centered) stationary mixing sequences (e.g. Theorem 4 in Doukhan et al. (1994)), and the Crámer-Wold device (Proposition 5.1 in White (2001)), under Assumptions 1, 2, and 4. The mixing condition of Theorem 4 in Doukhan et al. (1994)) is satisfied if $\lim_{T\to\infty}\sum_{t=1}^{T} t^{1/(r-1)}\alpha(t) < \infty$. It is easy to see that this holds for $\alpha(t) = O(t^{-A})$, with $A > r/(r-1)$. Notice that β-mixing implies α-mixing, with relation $\alpha(t) \leq \frac{1}{2}\beta(t)$ between β-mixing and α-mixing coefficients (see, e.g., Doukhan et al. (1995, p. 397)). But under Assumption 1 $\beta(t)$ diminishes at a faster, geometric rate, such that the mixing condition is satisfied.

We apply Theorem 1 of Doukhan et al. (1995) to establish stochastic equicontinuity of $\sqrt{n}\bar{L}_n(\cdot)$. First, notice from Application 1 in Doukhan et al. (1995) that the mixing condition is satisfied if $\lim_{T\to\infty}\sum_{t=1}^{T} t^{1/(r-1)}\beta(t) < \infty$, which was established in the preceding. Second, notice that under Assumption 2, the $L_{t+1}(\cdot)$ belong to $\mathcal{L}_{2r}$, where $\mathcal{L}_{2r}$ denotes the class of functions satisfying $\|f\|_{2r} < \infty$. From Application 1 in Doukhan et al. (1995) we then find that the entropy condition is satisfied if $\int_0^1 \sqrt{H_{[]}(\delta,\Gamma,\|\cdot\|_{2r})}du < \infty$, where $H_{[]}(\delta,\Gamma,\|\cdot\|_{2r})$ is defined as the natural logarithm of the $\mathcal{L}_{2r}$ bracketing numbers $N_{[]}(\delta,\Gamma,\|\cdot\|_{2r})$.

We can always choose N points in Γ, denoted γ_k, for $k = 1,\ldots,N$, and collected in Γ_N, such that for each $\gamma \in \Gamma$, $\min_k|\gamma-\gamma_k| < GN^{-1/d}$, because Γ is a bounded subset of $\mathbb{R}^d$.

Assumption 3 implies that $\|L_{t+1}(\gamma) - L_{t+1}(\gamma')\|_{2r} \leq \|L_{t+1}(\gamma) - L_{t+1}(\gamma')\|_{4r} \leq \bar{B}|\gamma-\gamma'|^{\lambda}$, for all $\gamma,\gamma' \in \Gamma$.

Setting $N(\delta) = \delta^{-d/\gamma}G^dB^{-d/\lambda}$, we therefore find that for all $\gamma \in \Gamma$ there exists a $\gamma_k \in \Gamma_N$ such that $\|L_{t+1}(\gamma) - L_{t+1}(\gamma_k)\|_{2r} \leq B|\gamma-\gamma_k|^{\lambda} \leq BG^{\lambda}N^{-\lambda/d} = \delta$. Hence, $N(\delta) = \delta^{-d/\lambda}G^d\bar{B}^{-d/\lambda}$ satisfies the definition of the $\mathcal{L}_{2r}$-bracketing numbers. Moreover, the entropy condition $\int_0^1 H_{[]}(\delta,\Gamma,\|\cdot\|_{2r})du = \int_0^1 \log(B^{d/\lambda}G^d\delta^{-d/\lambda})d\delta = d\log(B^{1/\lambda}G) + \int_0^1 \delta^{-d/\lambda}d\delta = d\log(B^{1/\lambda}G) + \frac{1}{2}\sqrt{\pi d/\lambda} < \infty$ holds.

It follows from Theorem 1 in Doukhan et al. (1995) that $\sqrt{n}\bar{L}_n(\cdot)$ is stochastically equicontinuous. Together with finite dimensional convergence this implies $\sqrt{n}\bar{L}_n(\cdot) \Rightarrow Z(\cdot)$, with $Z(\cdot)$ a Gaussian process with covariance kernel $\Sigma(\cdot,\cdot)$.

Note that $\sigma_n^2(\cdot) \xrightarrow{a.s} \sigma^2(\cdot)$ uniformly over Γ under Assumption 4. That $v(t_n) \xrightarrow{d} v(t_n)$ follows by application of the Continuous Mapping Theorem. □

A.2 Proof of Theorem 2

The result under H_0 follows from Theorem 1 and the distribution function of $v(\tilde{t})$ being absolutely continuous on $(0, \infty)$. The absolute continuity of the distribution function of $v(\tilde{t})$ follows from $Z(\cdot)$ having a nondegenerate covariance kernel, and thus $\tilde{t}(\cdot)$ having nondegenerate covariance kernel under Assumption 4, and the particular functional forms of $v(\cdot)$ under consideration (see Theorem 11.1 of Davydov et al. (1998)).

The result under H_1 is established as follows. Under the assumptions of Theorem 1 it follows that $\bar{L}_n(\gamma) \xrightarrow{a.s} E[L_{t+1}(\gamma)] \equiv \Delta(\gamma)$, uniformly over Γ. Now notice that, for any $\gamma \in \Gamma$, $\left|E[L_{t+1}(\gamma^{\dagger})]\right| - \left|E[L_{t+1}(\gamma^{\dagger}) - L_{t+1}(\gamma)]\right| \leq \left|E[L_{t+1}(\gamma)]\right|$ by the Triangle Inequality. Furthermore, from Jensen's Inequality, Hölder's Inequality and under Assumption 3, it follows that

$$
\begin{aligned}
\left|E[L_{t+1}(\gamma^{\dagger}) - L_{t+1}(\gamma)]\right| &\leq E\left[\left|L_{t+1}(\gamma^{\dagger}) - L_{t+1}(\gamma)\right|\right] \\
&\leq \left\|L_{t+1}(\gamma^{\dagger}) - L_{t+1}(\gamma)\right\|_{4r} \leq B|\gamma - \gamma^{\dagger}|^{\lambda}.
\end{aligned}
$$

Hence, if $\Gamma^{\dagger} = \{\gamma : |\gamma - \gamma'|^{\lambda} < \Delta/B\}$ has positive J-measure, there exists a $\Delta' > 0$ such that $|E[L_{t+1}(\gamma)]| = |\Delta(\gamma)| \geq \Delta'$, for all $\gamma \in \Gamma^{\dagger}$. It follows that $|\bar{L}_n(\gamma)| > \Delta'$, a.s., uniformly over $\Gamma^{\dagger}$.

Additionally, under Assumption 4 it follows that $\hat{\sigma}^2_n(\gamma) \xrightarrow{a.s.} \sigma^2_m(\gamma)$ uniformly over $\gamma \in \Gamma$, and $\inf_{\gamma \in \Gamma} \sigma^2(\gamma) > 0$, such that there exists a $\Delta'' > 0$ so that $n^{-1/2}|t_n(\gamma)| \xrightarrow{a.s.} n^{-1/2}\frac{|\bar{L}_n(\gamma)|}{\sigma(\gamma)} > \Delta''$, a.s., uniformly over $\Gamma^{\dagger}$. By application of the Continuous Mapping Theorem it follows that, a.s., $n^{-1}v(t_n^2) > 0$. Hence, $P[v(t_n^2) > c] \to 1$, for any constant $c \in \mathbb{R}$. □

A.3 Proof of Theorem 3

That $\sqrt{n}\bar{L}^*_n(\cdot) \Rightarrow Z(\cdot)$ almost surely follows from Theorem 1 in Bühlmann (1995). Assumption A1, A2, and A3 in Bühlmann (1995) are satisfied under Assumptions 1, 2, and 5 respectively. Finally, Assumption A4 in that paper is established in the proof of Theorem 1, since $N(\delta)$ satisfies the definition of the $\mathcal{L}_{4r}$ bracketing numbers, and $N(\delta) = \delta^{-d/\lambda} G^d \bar{B}^{-d/\lambda}$, for all $\delta > 0$.

Note that $\sigma^2_n(\cdot) \xrightarrow{a.s} \sigma^2(\cdot)$ uniformly over Γ under Assumption 4, such that $t^*_n(\cdot) \Rightarrow \tilde{t}(\cdot)$ almost surely under the Continuous Mapping Theorem.

That $v(t^*_n) \xrightarrow{d} v(t_n)$ in probability follows by application of a Continuous Mapping Theorem for bootstrapped processes (see Theorem 10.8 in Kosorok (2008)), given that the bootstrap is consistent in probability, which is implied by $\sqrt{n}\bar{L}^*_n(\cdot) \Rightarrow Z(\cdot)$ almost surely. The result follows. □

A.4 Proof of Proposition 1

We show the result for ave t_n^2. The result for the other tests follows from similar steps. *Part 1:* The weak convergence of t_n^2 as established in Theorem 1 and the Continuous Mapping Theorem, implies stochastic equicontinuity (see, e.g., Proposition 1 in Andrews (1994)), i.e., for all $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} P\left(\sup_{|\gamma-\gamma'|<\delta} \left| t_n^2(\gamma) - t_n^2(\gamma') \right| > \varepsilon \right) = 0,$$

where we again use the Euclidean metric to metrize Γ.

From absolute continuity of J it follows that that $\int_\Gamma dJ(\gamma) = \sum_{i=1}^{K_n} \int_{\Gamma_n^i} dJ(\gamma)$. Hence,

$$\begin{aligned}
\left| \int_\Gamma t_n^2(\gamma) dJ(\gamma) - \sum_{i=1}^{K_n} t_n^2(\gamma_{n,i}) \int_{\Gamma_n^i} dJ(\gamma) \right| &\leq \sum_{i=1}^{K_n} \int_{\Gamma_n^i} \left| t_n^2(\gamma) - t_n^2(\gamma_{n,i}) \right| dJ(\gamma) \\
&\leq \sum_{i=1}^{K_n} \sup_{\gamma\in\Gamma_n^i} \left| t_n^2(\gamma) - t_n^2(\gamma_{n,i}) \right| \int_{\Gamma_n^i} dJ(\gamma) \\
&\leq \sup_{|\gamma-\gamma'|<\delta_n} \left| t_n^2(\gamma) - t_n^2(\gamma_{n,i}) \right| \sum_{i=1}^{K_n} \int_{\Gamma_n^i} dJ(\gamma) \\
&= \sup_{|\gamma-\gamma'|<\delta_n} \left| t_n^2(\gamma) - t_n^2(\gamma_{n,i}) \right|,
\end{aligned}$$

For any $\varepsilon > 0$ there exists a $\delta > 0$ (with $\delta_n < \delta$ eventually), such that

$$\begin{aligned}
\limsup_{n\to\infty} P\left(\left| \int_\Gamma t_n^2(\gamma) dJ(\gamma) - \sum_{i=1}^{K_n} t_n^2(\gamma_{n,i}) \int_{\Gamma_n^i} dJ(\gamma) \right| > \varepsilon' \right) &\leq \limsup_{n\to\infty} P\left(\sup_{|\gamma-\gamma'|<\delta_n} \left| t_n^2(\gamma) - t_n^2(\gamma_{n,i}) \right| > \varepsilon \right) \\
&\leq \limsup_{n\to\infty} P\left(\sup_{|\gamma-\gamma'|<\delta} \left| t_n^2(\gamma) - t_n^2(\gamma_{n,i}) \right| > \varepsilon \right) \\
&< \varepsilon,
\end{aligned}$$

where the last display follows from the stochastic equicontinuity of $t_n^2(\gamma)$. Because δ is arbitrary, the result follows.

Part 2: We cover Γ with some hyperrectangle $\bar{\Gamma}$, which we can do because Γ is a bounded subset of Euclidian space. Consider the d-dimensional hyperrectangular grid of $\bar{\Gamma}$ with $\bar{K}_n$ elements $\{\bar{\Gamma}_n^i\}_{i=1}^{\bar{K}_n}$, such that $\sup_{\gamma,\gamma'\in\bar{\Gamma}_n^i} |\gamma - \gamma'| < \delta_n$, for all $i = 1, \ldots, \bar{K}_n$.

Now let $\{\Gamma_i^n\}_{i=1}^{K_n}$ be the K_n elements of $\{\bar{\Gamma}_n^i\}_{i=1}^{\bar{K}_n}$, such that $\Gamma_i^n \cap \Gamma$ is nonempty, and choose the $\gamma_{n,i}$ such that $\gamma_{n,i} \in \Gamma$.

We can expand

$$
\begin{aligned}
\widetilde{\operatorname{ave} t_n^2} - \widehat{\operatorname{ave} t_n^2} &= \frac{1}{S_n}\sum_{j=1}^{S_n} t_n^2(\gamma^{(j)}) - \sum_{i=1}^{K_n} t_n^2(\gamma_{n,i}) \int_{\Gamma_n^i} dJ(\gamma) \\
&= \sum_{i=1}^{K_n} t_n^2(\gamma_{n,i}) \left\{ \frac{1}{S_n}\sum_{j=1}^{S_n} \mathbb{1}\left(\gamma^{(j)} \in \Gamma_i^n\right) - \int_{\Gamma_n^i} dJ(\gamma) \right\} \\
&\quad + \sum_{i=1}^{K_n} \frac{1}{S_n}\sum_{j=1}^{S_n} \left(t_n^2(\gamma^{(j)}) - t_n^2(\gamma_{n,i}) \right) \mathbb{1}\left(\gamma^{(j)} \in \Gamma_i^n\right) \\
&= A_n + B_n.
\end{aligned}
$$

Notice that

$$
\begin{aligned}
|A_n| &\leq \sup_{\gamma\in\Gamma} t_n^2(\gamma) \cdot \sum_{i=1}^{K_n} \left| \frac{1}{S_n}\sum_{j=1}^{S_n} \mathbb{1}\left(\gamma^{(j)} \in \Gamma_i^n\right) - \int_{\Gamma_n^i} dJ(\gamma) \right| \\
&\leq K_n \sup_{\gamma\in\Gamma} t_n^2(\gamma) \sup_{\Gamma'\subset\Gamma} \left| \frac{1}{S_n}\sum_{j=1}^{S_n} \mathbb{1}\left(\gamma^{(j)} \in \Gamma'\right) - \int_{\Gamma'} dJ(\gamma) \right| \\
&= K_n O_p(1) C_n,
\end{aligned}
$$

where the last line follows from Theorem 1.

Furthermore, we can show that $S_n^{1/2-\eta} C_n = o_p(1)$, for any $\eta \in (0, 1/2)$, where the probability statement now holds under the J-measure, by a CLT for iid empirical processes. Notice that due to the hyperrectangular shape of the Γ_i^n, we have for each $\Gamma_i^n \subset \bar{\Gamma}$

$$
\mathbb{1}\left(\gamma \in \Gamma_i^n\right) = \prod_{i=1}^{d} \mathbb{1}\left(\gamma_i \leq \overline{\gamma_i^n}\right) \prod_{i=1}^{d} \left(1 - \mathbb{1}\left(\gamma_i \leq \underline{\gamma_i^n}\right)\right), \tag{41}
$$

with $\bar{\gamma}_i^n$ denotes the maximum of the ith coordinate of all points in Γ_i^n, and with $\underline{\gamma}_i^n$ denoting the mininum.

Indicator functions such as the factors in (41) are type I(b) functions in the definition of Andrews (1994), and by Theorem 3 in Andrews (1994) so is the product (41). A functional CLT follows from Theorem 1 and 2 in Andrews (1994)), and by application of the Continuous Mapping Theorem we find $\sup_{\Gamma'\subset\bar{\Gamma}} \left| \frac{1}{S_n} \sum_{j=1}^{S_n} \mathbb{1}\left(\gamma^{(j)} \in \Gamma'\right) - \int_{\Gamma'} dJ(\gamma) \right| = O_p(S_n^{-1/2})$. Hence, $S_n^{1/2-\eta} C_n = O_p(1)$.

Furthermore, notice that

$$\begin{aligned}
|B_n| &\leq \frac{1}{S_n}\sum_{j=1}^{S_n}\sum_{i=1}^{K_n}\left|t_n^2(\gamma^{(j)}) - t_n^2(\gamma_{n,i})\right| \mathbb{1}\left(\gamma^{(j)} \in \Gamma_i^n\right) \\
&\leq 2^d \frac{1}{S_n}\sum_{j=1}^{S_n} \sup_{|\gamma^{(j)}-\gamma'|<\delta_n} \left|t_n^2(\gamma^{(j)}) - t_n^2(\gamma')\right| \\
&\leq 2^d \sup_{|\gamma-\gamma'|<\delta_n} \left|t_n^2(\gamma) - t_n^2(\gamma')\right| = o_p(1),
\end{aligned}$$

by the stochastic equicontinuity of $t_n^2(\gamma)$ and where 2^d equals the maximum number of vertices shared amongst hyperrectangles in a hyperrectangular grid.

If we can choose $K_n = o(S_n^{-1/2+\eta})$ then $|A_n| = O_p(1)K_nC_n = O_p(1)o_p(1) = o_p(1)$. Hence, $|\widetilde{\text{ave}\, t_n^2} - \widehat{\text{ave}\, t_n^2}| = o_p(1)$. But we are free to choose the rate at which $K_n \to \infty$ as $n \to \infty$, so the result follows.

$\square$

Table 1: Small sample rejection rates of equal expected utility tests on equal weighted and minimum-variance portfolio strategies

	Two-sided tests				One-sided tests	
K_n	Wald	Bonferroni	ave-t^2	sup-t^2	Bonferroni	sup-t
Panel A: Size properties, n=120						
				n = 120		
1	0.14	0.14	0.13	0.13	0.17	0.19
10	0.58	0.06	0.09	0.12	0.07	0.15
50	0.50	0.04	0.08	0.14	0.04	0.16
100	0.50	0.04	0.09	0.14	0.05	0.18
250	-	0.03	0.09	0.13	0.04	0.17
Panel B: Size properties, n=600						
1	0.06	0.06	0.06	0.06	0.07	0.08
10	0.48	0.02	0.04	0.06	0.02	0.07
50	0.50	0.01	0.06	0.07	0.01	0.10
100	0.50	0.01	0.06	0.08	0.01	0.10
250	0.51	0.00	0.05	0.07	0.01	0.08
Panel C: Power properties, n=120						
1	0.12	0.12	0.12	0.12	0.01	0.01
10	0.98	0.81	0.47	0.94	0.86	0.96
50	0.53	0.65	0.41	0.92	0.70	0.96
100	0.29	0.61	0.40	0.93	0.67	0.96
250	-	0.51	0.40	0.94	0.57	0.95
Panel D: Power properties, n=600						
1	0.76	0.76	0.77	0.77	0.00	0.00
10	0.99	1.00	1.00	1.00	1.00	1.00
50	0.80	1.00	1.00	1.00	1.00	1.00
100	0.67	1.00	1.00	1.00	1.00	1.00
250	0.51	1.00	1.00	1.00	1.00	1.00

Note: This table presents the rejection rates of the proposed two-sided tests (ave-t^2 and sup-t^2) and one-sided test (sup-t), as well as the benchmark tests (Wald and Bonferroni). The data is generated according to equation (30), and the equal-weighed and minimum-variance portfolio strategies are given in Equation (29). The minimum-variance portfolio weights are estimated using a rolling window of $m = 120$ observations. The out-of-sample period consists of $n = 120$, and 600 observations. We consider discrete grids of $\Gamma = [1, 10]$ formed using $K_n = 1, 10, 50, 100$, and 250 equally spaced grid points. Results for the multivariate Wald test for $(K_n, n) = (250, 120)$ are not shown due to the singularity of the covariance matrix.

Table 2: Small sample rejection rates of quantile forecast tests, for differences between multivariate GARCH-DCC and RiskMetrics models

	Two-sided tests				One-sided tests	
S_n	Wald	Bonferroni	ave-t^2	sup-t^2	Bonferroni	sup-t
Panel A: Size properties, n=500						
31	0.36	0.03	0.03	0.05	0.01	0.01
50	0.93	0.03	0.03	0.04	0.01	0.01
100	1.00	0.02	0.03	0.04	0.00	0.01
250	1.00	0.01	0.03	0.04	0.00	0.01
500	-	0.01	0.03	0.04	0.00	0.01
1000	-	0.01	0.03	0.05	0.00	0.01
Panel B: Size properties, n=2,000						
31	0.03	0.01	0.01	0.02	0.01	0.01
50	0.13	0.01	0.02	0.01	0.00	0.01
100	0.81	0.00	0.01	0.01	0.00	0.01
250	1.00	0.00	0.02	0.01	0.00	0.01
500	1.00	0.00	0.02	0.01	0.00	0.01
1000	1.00	0.00	0.02	0.01	0.00	0.01
Panel C: Power properties, n=500						
31	0.42	0.10	0.20	0.11	0.16	0.15
50	0.91	0.07	0.16	0.10	0.11	0.14
100	1.00	0.05	0.13	0.10	0.07	0.14
250	1.00	0.03	0.11	0.10	0.04	0.14
500	-	0.02	0.12	0.10	0.03	0.14
1000	-	0.01	0.11	0.10	0.02	0.13
Panel D: Power properties, n=2,000						
31	0.33	0.50	0.81	0.52	0.62	0.65
50	0.38	0.42	0.66	0.52	0.55	0.64
100	0.82	0.30	0.52	0.50	0.43	0.63
250	1.00	0.20	0.46	0.50	0.28	0.64
500	1.00	0.14	0.46	0.51	0.21	0.64
1000	1.00	0.09	0.43	0.49	0.15	0.64

Note: This table presents the rejection rates of the proposed two-sided tests (ave-t^2 and sup-t^2) and one-sided test (sup-t), as well as the benchmark tests (Wald and Bonferroni). The quantile forecasts for the portfolio returns from the GARCH-DCC and multivariate RiskMetrics models are defined in equations (32) and (33). The data is generated as in equation (31) with $N = 30$. We test at 31 fixed portfolio weight vector being the equal weighted portfolio vector and the 30 basis vectors, as well as $S_n - 31$ weight vectors drawn uniformly from the unit simplex.

Table 3: Small sample rejection rates of Murphy diagram tests, for quantile forecast differences between GARCH and RiskMetrics models

	Two-sided tests			One-sided tests	
K_n	tick loss	ave-t^2	sup-t^2	tick loss	sup-t
Panel A: Size properties, n = 500					
1	0.07	-	-	0.04	-
50	-	0.04	0.04	-	0.03
100	-	0.03	0.03	-	0.02
250	-	0.02	0.02	-	0.01
Panel B: Size properties, n = 2,000					
1	0.06	-	-	0.04	-
50	-	0.04	0.05	-	0.03
100	-	0.05	0.05	-	0.04
250	-	0.04	0.04	-	0.04
Panel C: Power properties, n = 500					
1	0.11	-	-	0.19	-
50	-	0.04	0.04	-	0.06
100	-	0.04	0.04	-	0.06
250	-	0.03	0.03	-	0.04
Panel D: Power properties, n = 2,000					
1	0.37	-	-	0.49	-
50	-	0.10	0.08	-	0.14
100	-	0.10	0.08	-	0.13
250	-	0.12	0.10	-	0.15

Note: This table presents the rejection rates of the proposed two-sided tests (ave-t^2 and sup-t^2) and one-sided test (sup-t), as well as the benchmark tests (Diebold-Mariano using the tick loss function). The quantile forecasts from the GARCH and RiskMetrics models are given in equations (32) and (33), with $N = 1$. We consider out-of-sample period lengths $n = 500$, and 2,000, and discrete grids of $\Gamma = [-20, 0]$ with $K_n = 50, 100$, and 250 equally spaced grid points. The tick loss based tests use only a single loss function, and are reported in the rows labeled $K_n = 1$.

Table 4: Small sample rejection rates of Murphy diagram tests, for quantile forecast differences between GARCH and rolling window models

	Two-sided tests			One-sided tests	
K_n	tick loss	ave-t^2	sup-t^2	tick loss	sup-t
Panel A: Size properties, n = 500					
1	0.21	-	-	0.05	
50	-	0.14	0.12	-	0.07
100	-	0.14	0.10	-	0.09
250	-	0.15	0.09	-	0.08
Panel B: Size properties, n = 2,000					
1	0.16	-	-	0.06	-
50	-	0.11	0.08	-	0.05
100	-	0.13	0.09	-	0.08
250	-	0.14	0.07	-	0.06
Panel C: Power properties, n = 500					
1	0.32	-	-	0.45	-
50	-	0.26	0.27	-	0.36
100	-	0.33	0.30	-	0.39
250	-	0.30	0.27	-	0.36
Panel D: Power properties, n = 2,000					
1	0.90	-	-	0.95	-
50	-	0.76	0.72	-	0.81
100	-	0.85	0.77	-	0.85
250	-	0.89	0.82	-	0.88

Note: This table presents the rejection rates of the proposed two-sided tests (ave-t^2 and sup-t^2) and one-sided test (sup-t), as well as the benchmark tests (Diebold-Mariano using the tick loss function). The quantile forecasts from the GARCH model in equation (32) with $N = 1$, and a rolling window forecast with window length 250. We consider out-of-sample period lengths $n = 500$, and 2,000, and discrete grids of $\Gamma = [-20, 0]$ with $K_n = 50, 100$, and 250 equally spaced grid points. The tick loss based tests use only a single loss function, and are reported in the rows labeled $K_n = 1$.

Table 5: p-values from tests of equal expected utility from equal-weighted and minimum-variance portfolio strategies

	Two-sided tests				One-sided tests			
					H_1': EW superior		H_1'': MV superior	
K_n	Wald	Bonf.	ave-t^2	sup-t^2	Bonf.	sup-t	Bonf.	sup-t
Panel A: Risk aversion in the range [1,10]								
1	0.03	0.03	0.04	0.04	0.02	0.01	0.97	0.97
10	0.00	0.04	0.13	0.01	0.02	0.00	1.00	0.18
50	0.00	0.22	0.17	0.01	0.11	0.00	1.00	0.20
100	0.00	0.45	0.16	0.01	0.22	0.00	1.00	0.24
250	0.26	1.00	0.16	0.02	0.56	0.01	1.00	0.21
Panel B: Risk aversion in the range [1,5]								
1	0.03	0.03	0.04	0.04	0.02	0.02	0.98	0.98
10	0.00	0.04	0.07	0.01	0.02	0.00	1.00	0.86
50	0.00	0.22	0.07	0.01	0.11	0.00	1.00	0.85
100	0.00	0.45	0.06	0.01	0.22	0.00	1.00	0.83
250	0.00	1.00	0.04	0.01	0.56	0.01	1.00	0.85
Panel C: Risk aversion in the range [5,10]								
1	0.87	0.87	0.87	0.87	0.56	0.57	0.44	0.43
10	0.00	1.00	0.53	0.38	1.00	0.22	1.00	0.19
50	0.00	1.00	0.53	0.36	1.00	0.17	1.00	0.22
100	0.00	1.00	0.62	0.41	1.00	0.22	1.00	0.17
250	0.02	1.00	0.58	0.38	1.00	0.20	1.00	0.20

Note: This table presents p-values from tests for equal (two-sided) and superior (one-sided) predictive accuracy. The equal-weighted and minimum-variance portfolio strategies are given in equation (29). The data consists of monthly returns of 30 industry portfolios and runs from September 1926 to December 2017. The three panels consider three ranges of values for the risk aversion parameter. K_n indicates the number of (equally-spaced) grid points.

Table 6: p-values of quantile forecast tests, for differences between multivariate GARCH-DCC and RiskMetrics models

	Two-sided tests				One-sided tests			
					H_1': RM superior		H_1'': DCC superior	
S_n	Wald	Bonf.	ave-t^2	sup-t^2	Bonf.	sup-t	Bonf.	sup-t
Panel A: Full sample								
1	0.78	0.78	0.79	0.79	0.39	0.39	0.61	0.56
31	0.99	1.00	0.99	0.95	1.00	0.62	1.00	0.93
100	-	1.00	0.94	0.95	1.00	0.64	1.00	0.93
250	-	1.00	0.81	0.93	1.00	0.59	1.00	0.93
500	-	1.00	0.79	0.90	1.00	0.54	1.00	0.92
1000	-	1.00	0.74	0.90	1.00	0.58	1.00	0.93
Panel B: First sub-sample								
1	0.30	0.30	0.30	0.30	0.15	0.15	0.85	0.85
31	0.01	0.05	0.04	0.04	0.02	0.02	1.00	0.99
100	-	0.16	0.22	0.05	0.08	0.03	1.00	1.00
250	-	0.40	0.23	0.05	0.20	0.03	1.00	1.00
500	-	0.80	0.24	0.06	0.40	0.04	1.00	1.00
1000	-	1.00	0.23	0.05	0.80	0.03	1.00	0.99
Panel C: Second sub-sample								
1	0.32	0.32	0.34	0.34	0.84	0.85	0.16	0.11
31	0.33	0.35	0.22	0.17	1.00	0.87	0.17	0.08
100	-	1.00	0.28	0.16	1.00	0.90	0.56	0.09
250	-	1.00	0.28	0.18	1.00	0.91	1.00	0.09
500	-	1.00	0.30	0.18	1.00	0.90	1.00	0.09
1000	-	1.00	0.30	0.17	1.00	0.91	1.00	0.11

This table presents p-values from tests for equal (two-sided) and superior (one-sided) predictive accuracy. The quantile forecasts for the portfolio returns from the GARCH-DCC and multivariate RiskMetrics models are defined in equations (32) and (33). The data consists of twenty years of daily returns of 30 industry portfolios and runs from January 1998 to December 2017. We consider S_n portfolio weight vectors. When $S_n = 1$ we use equal weights. For $S_n \geq 31$ we test at 31 fixed portfolio weight vectors (the equal-weighted portfolio vector and the 30 basis vectors) and $S_n - 31$ weight vectors drawn uniformly from the 30-dimensional unit simplex.

Table 7: Quantile forecast comparison tests for the Transportation industry portfolio

	Two-sided tests			One-sided tests			
				H_1': Comp. superior		H_1'': GARCH superior	
K_n	tick loss	ave-t^2	sup-t^2	tick loss	sup-t	tick loss	sup-t
Panel A: RiskMetrics vs. GARCH							
1	0.77	-	-	0.38	-	0.62	-
50	-	0.00	0.03	-	0.01	-	0.06
100	-	0.01	0.08	-	0.04	-	0.21
250	-	0.02	0.10	-	0.05	-	0.25
500	-	0.01	0.06	-	0.03	-	0.24
1000	-	0.01	0.07	-	0.04	-	0.24
Panel B: Rolling window vs. GARCH							
1	0.00	-	-	1.00	-	0.00	-
50	-	0.00	0.00	-	0.96	-	0.00
100	-	0.00	0.00	-	0.98	-	0.00
250	-	0.00	0.00	-	0.99	-	0.00
500	-	0.00	0.00	-	0.98	-	0.00
1000	-	0.00	0.00	-	0.99	-	0.00

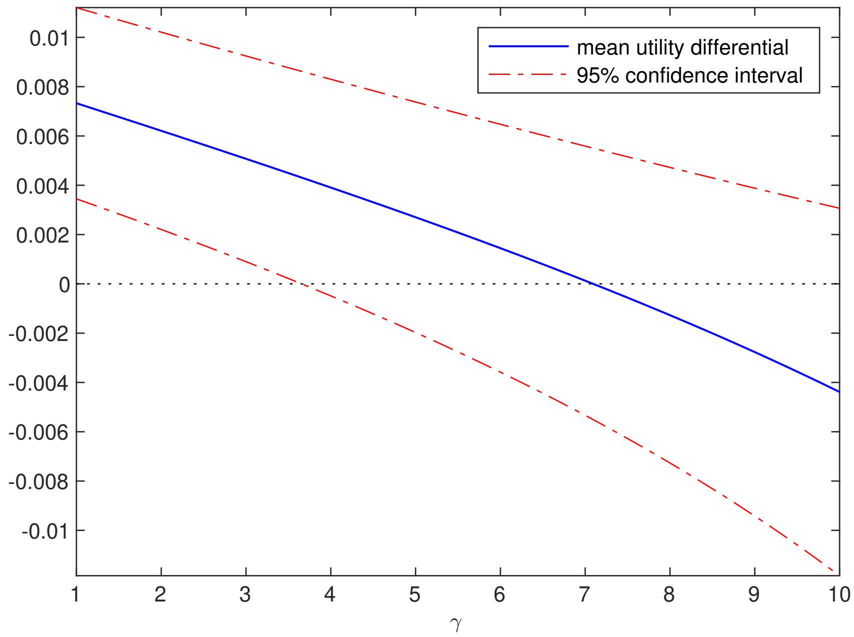
Note: This table presents p-values from tests comparing the predictive accuracy of forecasts of the 5% quantile for daily returns obtained from GARCH, RiskMetrics and rolling window models. We consider one-sided and two-sided tests based on the elementary scoring rules for the quantile, as well as benchmark tests using the tick loss function, applied to daily returns on the Transportation industry portfolio. We consider equally-spaced discrete grids of $\Gamma = [-20, 0]$ with K_n grid points. The tick loss based tests use only a single loss function, and are reported in the rows labeled $K_n = 1$. The model referred to in the panel labeled "Comp. superior" is RiskMetrics in Panel A and rolling window in Panel B.

Table 8: Quantile forecast comparison test results across all 30 industry portfolios.

	Two-sided tests			Forecast dominance outcomes			
				(1)	(2)	(3)	(4)
				Comp.	GARCH	No	No
K_n	tick loss	ave-t^2	sup-t^2	dominates	dominates	rejections	ordering
Panel A: RiskMetrics vs. GARCH							
1	0.00	-	-	-	-	-	-
50	-	0.07	0.07	0.17	0.00	0.83	0.00
100	-	0.23	0.17	0.23	0.10	0.67	0.00
250	-	0.20	0.07	0.10	0.03	0.83	0.03
500	-	0.27	0.10	0.10	0.07	0.80	0.03
1000	-	0.23	0.10	0.10	0.07	0.83	0.00
Panel B: Rolling window vs. GARCH							
1	1.00	-	-	-	-	-	-
50	-	1.00	1.00	0.00	1.00	0.00	0.00
100	-	1.00	1.00	0.00	1.00	0.00	0.00
250	-	1.00	1.00	0.00	1.00	0.00	0.00
500	-	1.00	1.00	0.00	1.00	0.00	0.00
1000	-	1.00	1.00	0.00	1.00	0.00	0.00

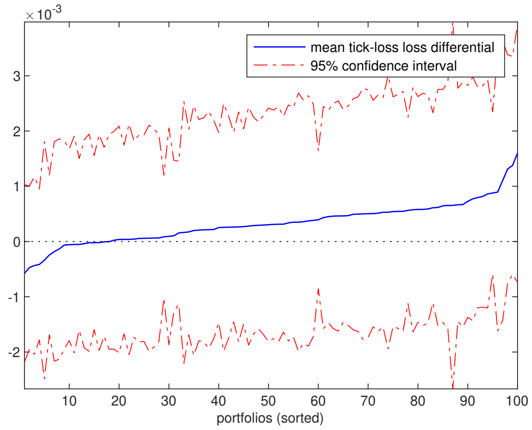
Note: This table presents rejection rates, across 30 industry portfolios, using tests comparing the predictive accuracy of forecasts of the 5% quantile for daily returns obtained from GARCH, RiskMetrics and rolling window models. We consider one-sided and two-sided tests based on the elementary scoring rules for the quantile, as well as benchmark tests using the tick loss function. We consider equally-spaced discrete grids of $\Gamma = [-20, 0]$ with K_n grid points. The tick loss based tests use only a single loss function, and are reported in the rows labeled $K_n = 1$. The forecast dominance results in the final four columns are based on the decision rule described in Section 3.3. The model referred to in the column labeled "Comp. dominates" is RiskMetrics in Panel A and rolling window in Panel B.

Figure 1: Utility differential of equal weighted and minimum-variance portfolio strategies

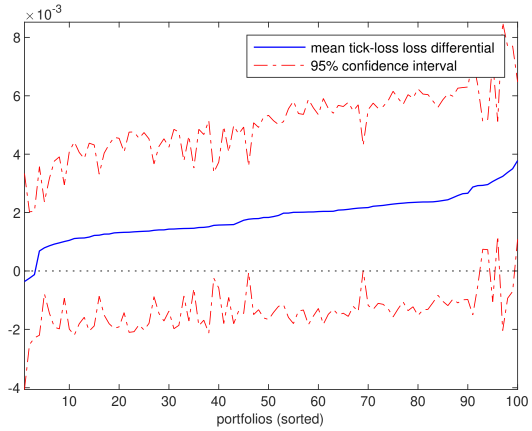


Note: This figure plots the the sample mean of utility from the equal-weighted portfolio strategy *minus* the sample mean of utility derived from the minimum-variance portfolio strategies. The strategies are given equation (29). The data consists of monthly returns of 30 industry portfolios and runs from September 1926 to December 2017. We consider risk aversion, γ, in the range $[1, 10]$.

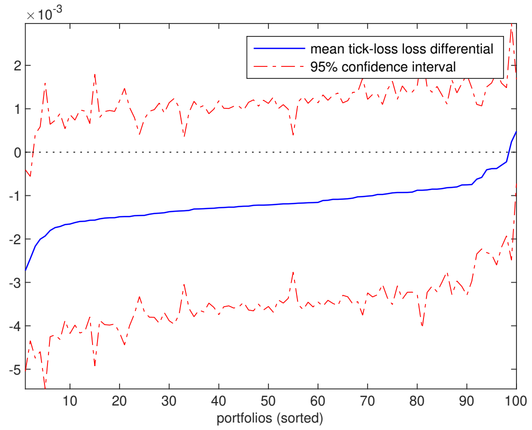
Figure 2: Tick-loss differential for tail quantile forecasts generated by the multivariate GARCH-DCC and RiskMetrics models



(a) Full sample



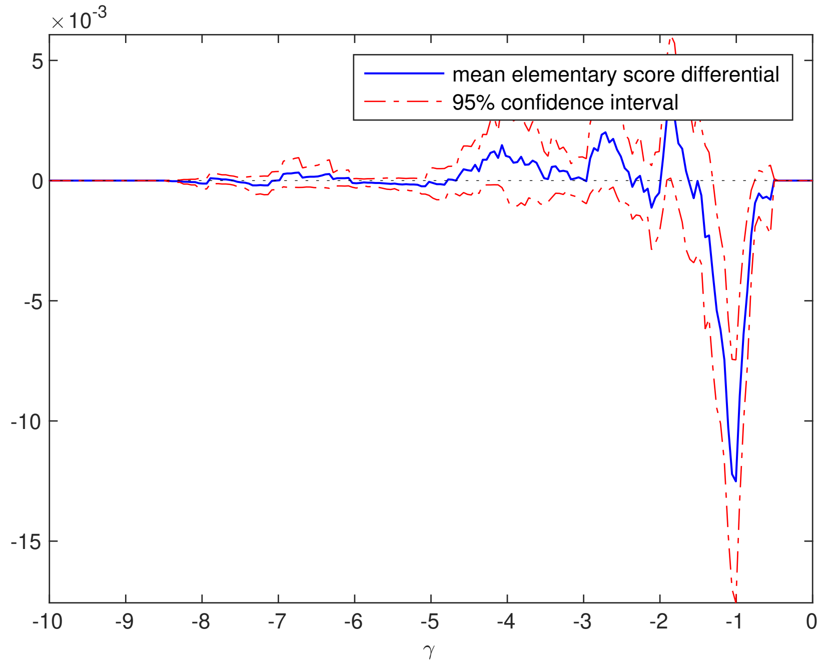
(b) First sub-sample



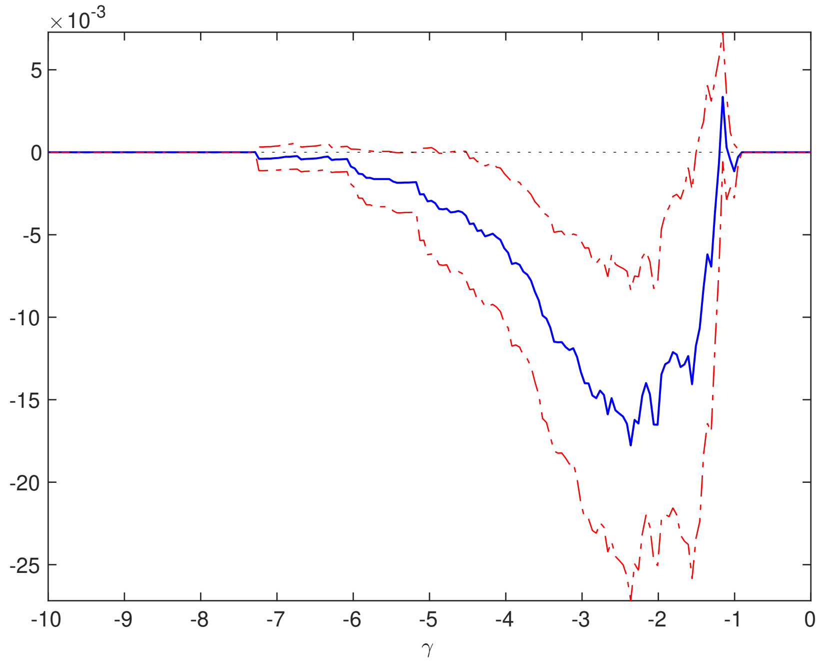
(c) Second sub-sample

Note: This figure plots the sample average of the tick loss from the GARCH-DCC forecasts *minus* the sample mean tick loss of the RiskMetrics forecasts. The forecasts are defined in equations (32) and (33). The mean tick loss differences are shown for $S_n = 100$ portfolio weight vectors, with 31 of these being the equal weighted portfolio vector and the 30 single-asset portfolio vectors, and the rest drawn uniformly from the 30-dimensional simplex. The portfolios are presented sorted on mean tick loss difference. The data consists of daily returns for 30 Industry portfolios and runs from January 1998 to December 2017.

Figure 3: Murphy diagrams for quantile forecasts for the Transportation portfolio returns



(a) GARCH vs. RiskMetrics



(b) GARCH vs. rolling window sample quantile

Note: This figure plots the sample mean of the elementary scoring rule of the GARCH forecasts *minus* the sample mean of the elementary scoring rule of the RiskMetrics forecasts. The forecasts are 5% quantile forecasts for daily returns of the US Transportation industry index from January 1998 to December 2017. The elementary scoring rules are indexed by the scalar parameter $\gamma \in [-10, 0]$.