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

Sander Barendse* Andrew Patton

Oxford University Duke University

February 14, 2019

Preliminary Draft, Please Do Not Circulate

Abstract

We develop joint out-of-sample tests for multiple testing problems that arise when comparing predictive accuracy using loss or utility functions that contain *shape parameters*. Our tests cover forecast comparison scenarios in which the shape parameter (vector) takes values in some subset of Euclidean space. We apply our tests to three such forecast evaluation problems. First, we consider hypotheses of equal (superior) expected utility between two portfolio strategies, defined over an interval of risk aversion parameter values. Second, we consider hypotheses of equal (superior) predictive ability between two conditional quantile forecast models using *Murphy diagrams*. Finally, we consider hypotheses of equal (superior) predictive ability of univariate quantile forecasts of portfolio returns–as generated by multivariate models of the portfolio assets–by examining all portfolios with positive weights summing to one. In empirical applications we show that the new tests reject at least as often as benchmarks tests, such as the standard Wald test or Bonferonni multiple correction, and are better behaved than the benchmarks in practice, in that p-values remain stable as we test at more elements of the multiple hypothesis. Monte-Carlo experiments verify that our tests have good size and power properties in small sample.

Keywords: Forecasting; Model comparison; Model selection; Out-of-sample testing.

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

1 Introduction

Many forecast comparison problems in economics are really joint testing problems over a continuum of *shape parameters*, either (i) because the loss function or utility function with which we evaluate differences in forecasts depends on some arbitrary shape parameter (vector) in Euclidean space, or (ii) because the maintained hypothesis of equal or superior predictive ability has a continuum of testable implications. Instead of testing the single (arbitrarily chosen or conventional) hypothesis, testing the multiple hypothesis can improve robustness of the results by being potentially less sensitivity to small sample variation, and can increase power when the test of the single hypothesis is less sensitive to deviations from the null than a test of other elements of the multiple hypothesis.

Examples of hypotheses that depend on some arbitrary parameter include hypotheses 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 a discussion) we really should consider a multiple testing problem of hypotheses of equal expected utility over a range of risk aversion values that researchers find convincing, 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. In finance there is a large literature that evaluates models in this way. Fleming et al. (2001, 2003); Marquering and Verbeek (2004) study the benefits of volatility timing for a mean-variance investor, with several values of risk aversion. Similarly, Engle and Colacito (2006) propose tests to compare bivariate dynamic correlation models using quadratic utility, for several values of risk aversion. In DeMiguel et al. (2007) differences in certainty equivalent returns of strategies are evaluated for mean-variance investors, with testing performed individually for several values of the risk aversion parameter.

An example of an hypothesis of equal or superior predictive ability that has a continuum of testable implications is the evaluation of *Murphy diagrams*. Ehm et al. (2016) shows that superior forecasts of statistical quantities, such as quantiles and expectiles, should have lower expected loss in terms of a family of so-called *elementary score functions*, or *elementary loss functions* in econometric parlance, indexed by some real parameter (the plot of the elementary loss functions is called a Murphy diagram). Superior predictive ability therefore implies that a continuum of elementary loss function differentials has mean zero. Ehm et al. (2016) notice

that joint testing of the Murphy diagram has not been fully developed yet for time series data. Recently joint tests for Murphy diagrams based on controlling the *family-wise error rate* have been introduced, such as in Ziegel et al. (2017), but such p-value corrections using bounds can be unsuitable to large-scale multiple testing problems, see Hand (1998, p. 115) and White (2000). Moreover, these tests consider only a finite subset of the testable implications instead of a continuum.

Another example of an hypothesis of equal or superior predictive ability that has a continuum of testable implications arises when we evaluate quantile forecasts of multivariate models, since an equal (superior) model should generate better forecasts for any linear combination of the random variables. The hypothesis of equal (superior) predictive ability therefore implies that the mean loss differential for all linear combinations should be zero (smaller than zero). In practice, comparisons of tail quantile (*or* Value-at-Risk) forecasts of portfolio returns, as generated from multivariate models, often only consider the equally-weighted portfolio or some other fixed linear combination of portfolio constituents, see e.g. McAleer and Da Veiga (2008); Santos et al. (2012); Kole et al. (2017). Considering only a single *well-diversified* weight vector can put more advanced multivariate models at a disadvantage by *diversifying away* features in the data that can only be captured by more sophisticaled models.

Our contribution in this paper is that we develop out-of-sample tests that delivers asymptotically appropriate p-values for multiple testing problems over a continuum of shape parameters – which includes the above three examples – and which do not rely on bounds, such as the Bonferonni multiple correction, meanwhile taking into account the time-series nature of the data that is inherent to most forecasting settings. To our knowledge such out-of-sample tests, which do not rely on bounds and are applicable to time-series data, have not yet been developed in these cases. Moreover, our theory can cover many other relevant out-of-sample testing scenarios. We also contribute to the literature by considering simulation studies and empirical applications related to the three above-mentioned examples. In the simulation studies we consider DGPs that closely resemble the data in those scenarios.

We derive our tests using the supremum or average of Diebold-Mariano tests related to each of the shape parameter values in the multiple hypothesis, and obtain critical values using the moving blocks bootstrap of Bühlmann (1995). This bootstrap procedure is suitable to the multiple testing problem at hand, because it is applicable to weakly dependent empirical processes indexed by classes of functions that are general enough to cover our cases of interest:

utility functions and loss functions parameterized by some vector that can take values in a bounded subset of Euclidian space.

In simulation studies and empirical analyses we show that our tests are generally more powerful than benchmarks, or reject similarly or more frequently than in practice, with benchmarks including the multivariate Wald test and a test using the Bonferonni multiple correction. (i) In a comparison of expected utility of naive and minimum-variance portfolio strategies (see DeMiguel et al. (2007) for an elaboration) we show that our tests reject hypotheses of equal and superior predictive ability as often or more frequently than the benchmarks for similar data as used in DeMiguel et al. (2007). (ii) For a multiple test of the Murphy diagram for quantile forecasts generated by univariate GARCH and Riskmetrics (1996) models we find that our tests rejects more frequently than the benchmarks. (iii) For a multiple test of quantile forecasts generated by multivariate GARCH-DCC and Riskmetrics models – for the linear portfolios defined by all weight vectors in the unit simplex (i.e., positive weights, with weights summing up to one) – we find that our tests reject more often than the benchmarks. Finally, (iv) Monte-Carlo experiments confirm that the bootstrap has good size and power properties in small sample.

Our tests build on the out-of-sample testing framework of Diebold and Mariano (1995), of which the asymptotic properties are further developed in West (1996). Similar to Giacomini and White (2006) we consider the forecasting methods (which, besides the model, includes the forecasting scheme and choices of in-sample and out-of-sample periods) rather than just forecasting models. Finally, White (2000) and Hansen (2005) develop tests of superior predictive ability for some model against finitely many benchmark. Our tests differ from the latter tests in that we allow for the joint testing over a continuum of shape parameters rather than a finite subset.

In what follows we first discuss three illustrative examples in Section 2. In Section 3 we discuss 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 Illustrative scenarios

In this paper we consider the following running examples of backtesting scenarios that we study in simulation exercises and empirical illustrations. In Example 1 we compare portfolio

strategies in terms of expected utility, i.e. with an economic loss function. In Example 2 we consider the multiple testing problem implied by Murphy diagrams. In Example 3 we consider the multiple testing problem implied by studying differences in tail quantile forecasts obtained from multivariate models.

Example 1. *(Comparison of portfolio strategies) We want to compare two portfolio strategies in terms of expected utility. Consider some vector of returns* Y_{t+1}*, and portfolio returns* $Y'_{t+1}\hat{\omega}^{(1)}_{t,m_1}$ *and* $Y'_{t+1}\hat{\omega}^{(2)}_{t,m_2}$*, where, for* $i = 1, 2$*,* $\hat{\omega}^{(i)}_{t,m_1}$ *is a portfolio weight vector of the* i*th strategy and estimated using observations* $Y_t, \ldots, Y_{t-m_i+1}$*, and* $\max(m_1, m_2) \leq m$*. Consider a utility function* $u^{(i)}_{t+1}(\gamma) = u(Y'_{t+1}\hat{\omega}^{(1)}_{t,m_1}; \gamma)$ *that depends on some parameter vector* $\gamma \in \Gamma$*. For instance, when* $u(\cdot)$ *is the exponential utility function* γ *denotes the (scalar) risk aversion parameter, and* $\Gamma \in [a, b]$*, with* $0 < a < b < \infty$*. We then have loss differences* $L(Y_{t+1}, Y_t, \ldots, Y_{t-m+1}; \gamma) = u^{(1)}_{t+1}(\gamma) - u^{(2)}_{t+1}(\gamma)$*, and the multiple hypothesis of equal (superior) expected utility* $E[L(Y_{t+1}, Y_t, \ldots, Y_{t-m+1}; \gamma)] = 0$ *(*$E[L(Y_{t+1}, Y_t, \ldots, Y_{t-m+1}; \gamma)] \leq 0$*), for all* $\gamma \in \Gamma$*.*

Example 2. *(Forecast comparison of portfolio return characteristics) Let* Y_{t+1} *denote some vector of returns. The portfolio return for some vector* γ *is then given by* $\tilde{Y}_{t+1}(\gamma) = Y'_{t+1}\gamma$*. Also consider* $W_t = (Y_t, X_t)$*, with* X_t *a vector of other relevant variables. We are interested in forecasting some statistic* ψ_t *of* $\tilde{Y}_{t+1}(\gamma)$ *conditional on* $W_t, W_{t-1}, \ldots$*, such as a conditional quantile, for all portfolios implied by* $\gamma \in \Gamma$*. When we consider all long-only portfolios with weights summing to one* Γ *equals the unit simplex, i.e.* $\Gamma = \{\gamma : \sum_{i=1}^{h} \gamma_i = 1, \gamma_i \geq 0,$ *for all* $i = 1, \ldots, r\}$*. Consider the following forecasts of* ψ_t*:* $g^{(1)}_t(\gamma) = g^{(1)}(W_t, \ldots, W_{t-m+1}; \hat{\beta}^{(1)}_{t,m_1}, \gamma)$ *and* $g^{(1)}_t(\gamma) = g^{(2)}(W_t, \ldots, W_{t-m+1}; \hat{\beta}^{(2)}_{t,m_2}, \gamma)$*, where, for* $i = 1, 2$*,* $g^{(i)}(\gamma)$ *is a measurable function and* $\hat{\beta}^{(i)}_{t,m_i}$ *is an estimated parameter vector using observations* $W_t, \ldots, W_{t-m_i+1}$*, and* $\max(m_1, m_2) \leq m$*. Consider a loss function* $S^{(i)}_{t+1}(\gamma) = S(\tilde{Y}_{t+1}, g^{(1)}_t)$*, e.g. the tick-loss function when we consider the conditional quantile. We have loss differences* $L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma) = S^{(1)}_{t+1}(\gamma) - S^{(2)}_{t+1}(\gamma)$*, and the multiple hypothesis of equal (superior) predictive ability:* $E[L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma)] = 0$ *(*$E[L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma)] \leq 0$*), for all* $\gamma \in \Gamma$*.*

Example 3. *(Murphy diagrams) Let* Y_{t+1} *denote some scalar return, and let* ξ_t *denote some statistic of* Y_{t+1} *in which we are interested, conditional on* $W_t, W_{t-1}, \ldots$*, with* $W_t = (Y_t, X_t)$*, and* X_t *a vector of other relevant variables. Under some distributional assumptions many statistics, such as the (conditional) quantile, expectile, and (special case) mean admit families* $\mathcal{S}$ *of strictly consistent scoring functions, such that for all loss functions functions in this family* $S \in \mathcal{S}$*, it*

holds that $E_t[S(Y_t, \xi_t)] < E_t[S(Y_t, x)]$*, for all* x *in the domain of* ξ_t*, other than the value of* ξ_t *itself, and where* $E_t[\cdot]$ *denotes the expectation conditional on* $W_t, W_{t-1}, \ldots$*. Consider the following forecasts of* ξ_t*:* $g_t^{(1)} = g^{(1)}(W_t, \ldots, W_{t-m+1}; \hat{\beta}_{t,m_1}^{(1)})$ *and* $g_t^{(1)} = g^{(2)}(W_t, \ldots, W_{t-m+1}; \hat{\beta}_{t,m_2}^{(2)})$*, where, for* $i = 1, 2$*,* $g^{(i)}$ *is a measurable function and* $\hat{\beta}_{t,m_i}^{(i)}$ *is an estimated parameter vector using observations* $W_t, \ldots, W_{t-m_i+1}$*, and* $\max(m_1, m_2) \leq m$*. If we want to compare* $g_t^{(1)}$ *to* $g_t^{(2)}$ *we usually select one or several of these loss functions for testing. This can be problematic when tests using different members of the family* $\mathcal{S}$ *generate different conclusions. Ehm et al. (2016) show that all members of the strictly consistent loss function families for quantiles, expectiles, and the mean, can be written as mixtures of the elementary loss functions* $\tilde{S}(\cdot, \cdot; \gamma)$ *, i.e. for all* $S \in \mathcal{S}$ *we have* $S(Y_t, x) = \int_{-\infty}^{\infty} \tilde{S}(Y_t, x; \gamma) dH(\gamma)$*, for some non-negative measure* H*, and* $\gamma \in \Gamma \subset \mathbb{R}$*. We can consider all elementary loss differences* $L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma) = \tilde{S}(Y_{t+1}, g_t^{(1)}; \gamma) - \tilde{S}(Y_{t+1}, g_t^{(1)}; \gamma)$ *simultaneously, by testing the multiple hypothesis of equal (superior) predictive ability* $E[L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma)] = 0$ *(*$E[L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma)] \leq 0$*), for all* $\gamma \in \Gamma$*. Ehm et al. (2016) establish that this implies equal (superior) predictive ability for all members of* $\mathcal{S}$*. The graph plotting the sample mean of all elementary loss differences is usually referred to as the Murphy diagram.*

3 Theory

Consider the stochastic process $W = \left\{W_t : \Omega \to \mathbb{R}^{N+s}, N \in \mathbb{N}_+, s \in \mathbb{N}, t = 1, 2, \ldots\right\}$ 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 $\mathcal{F}_t = \sigma(W_1', \ldots, W_t')'$.

In our sample we denote the total sample size by T and estimate n out-of-sample forecasts. 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 measurable loss (difference) $L_{m,t+1}(\gamma) = L(W_{t+1}, W_t, \ldots, W_{t-m+1}; \gamma)$ 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. By setting $m < \infty$ we impose 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. This limited memory condition is also imposed in Giacomini and White (2006).

Since the forecasting model parameters do not converge in probability to the true parameters due to finite in-sample period length m, the choice of m, and the choice between fixed or moving window forecast schemes are important determinants of forecasting performance. By evaluating performance differences between forecasts from models with finite length in-sample periods we are therefore comparing forecasting *methods*, rather than just forecasting *models*, where the forecasting method specification includes the forecasting model, the forecasting scheme, and the choice of in-sample and out-of-sample period lengths. The test framework of Giacomini and White (2006) therefore differs from West (1996) who does require the in-sample period length to diverge, and therefore evaluate performance differences between forecasting *models*. Two benefits of using the Giacomini and White (2006) framework are that our tests remain valid for nested models (see, e.g., Clark and McCracken (2001) for valid tests of nested models when in-sample period length does diverge), and our tests do not have to be corrected for estimation error (see, e.g., West (1996), McCracken (2000), and Escanciano and Olmo (2010) for robust tests when in-sample period length diverges) when the in-sample period length is of the same order as the out-of-sample period length.

We are interested in testing the two-sided hypothesis

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

and the one-sided hypothesis

$$H_0': \ E\left[L_{m,t+1}(\gamma)\right] \leq 0 \ \forall \, \gamma \in \Gamma, \tag{2}$$

Testing these unconditional hypotheses provides information on differences in forecasting performance or utility on average. We will first derive tests for H_0, and subsequently extend our theory to tests of H_0'.

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

3.1 Equal predictive ability tests

To develop a test of H_0 we use as ingredients the following Diebold-Mariano test statistics (Diebold and Mariano, 1995) at each $\gamma \in \Gamma$:

$$t_{m,n}(\gamma) \equiv \sqrt{n}\frac{\bar{L}_{m,n}(\gamma)}{\hat{\sigma}_{m,n}(\gamma)},$$

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

It should be noted that when autocorrelation is present in the $L_{m,t+1}(\gamma)$ the $t_{m,n}(\gamma)$ do not converge in distribution to a standard normal limit, because $\hat{\sigma}^2_{m,n}(\gamma)$ is not a heteroskedasticity and autocorrelation corrected (HAC) estimator of the asymptotic covariance matrix of $\sqrt{n}\bar{L}_{m,n}(\gamma)$ (see, e.g. Newey and West (1987)). We divide by $\hat{\sigma}_{m,n}(\gamma)$ because we are not aware of uniform convergence in probability results for HAC estimators, which we do require in our theory for $\hat{\sigma}_{m,n}(\gamma)$. Moreover, inference is not affected because the bootstrap accounts for time-series features in the data to obtain critical values of our tests. Finally, in some scenarios it might be better not to studentize, and calibrate $\hat{\sigma}^2_{m,n}(\gamma) = \sigma^2_m(\gamma) = 1$ instead. Such scenarios include those for which $\hat{\sigma}^2_{m,n}(\gamma)$ is close to zero in small samples. That $\hat{\sigma}_{m,n}(\gamma)$ does not have to be a consistent of the variance of $\sqrt{n}\bar{L}_{m,n}(\gamma)$, is also noted by Hansen (2005, under Cor. 3).

To test H_0 we employ the test statistics

$$\sup t^2_{m,n} \equiv \sup_{\gamma\in\Gamma} t^2_{m,n}(\gamma),$$

and

$$\text{ave}\, t^2_{m,n} \equiv \int_\Gamma t^2_{m,n}(\gamma) dJ(\gamma),$$

where J is some weight function over Γ. For instance, in Example 1 we let J be uniform on $\Gamma = [a, b]$. Each of these tests can be written as functions $v(t_{m,n})$, where v maps functionals on Γ to $\mathbb{R}$ and we write $t_{m,n} = \{t_{m,n}(\gamma) : \gamma \in \Gamma\}$ 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 $v(Z) \to \infty$.

Under the following assumptions we derive the asymptotic distribution of the tests.

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

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

Assumption 4. *$\hat{\sigma}^2_{m,n}(\gamma)\xrightarrow{a.s.}\sigma^2_m(\gamma)$ uniformly over $\gamma\in\Gamma$. Moreover, $\inf_{\gamma\in\Gamma}\sigma^2_m(\gamma)>0$ and $\Sigma_{m,n}(\cdot,\cdot)\equiv Cov\big(\sqrt{n}\bar{L}_{m,n}(\cdot),\sqrt{n}\bar{L}_{m,n}(\cdot)\big)>0$ for n sufficiently large.*

Theorem 1. *Let Assumptions 1 to 4 be satisfied. It follows that, for some $m<\infty$, under H_0, $\sqrt{n}\bar{L}_{m,n}(\cdot)\Rightarrow Z_{m,n}(\cdot)$, for some Gaussian process $Z_{m,n}(\cdot)$ with covariance kernel $\Sigma_{m,n}(\cdot,\cdot)$. Moreover, it follows that $v(t_{m,n})\xrightarrow{d}v(\tilde{t}_{m,n})$, with $\tilde{t}_{m,n}(\cdot)\equiv Z_{m,n}(\cdot)/\sigma_m(\cdot)$.*

The β-mixing condition in Assumption 1 – which is stronger than α, but weaker than ϕ-mixing – is usually assumed when deriving functional CLTs for time series data with unbounded moments. Moreover, the decay rate is quite strong, but required for the bootstrap of Bühlmann (1995). Bühlmann (1995) notices that this 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. Bradley et al. (2005, Thm. 3.7) shows that geometric ergodicity implies β-mixing with at least exponential rate, which satisfies Assumption 1. Assumption 2 is only slightly stronger than the existence of the fourth moments of the loss (differences) $L_{m,t+1,i}$, and Assumption 3 requires a Lipschitz condition to hold for these moments. Assumption 4 requires that $\hat{\sigma}_{m,n}(\cdot)$ satisfies a strong Uniform Law of Large Numbers (see, e.g., Andrews (1992)), and also imposes uniform non-singularity of $\sigma^2_m(\cdot)$ and $\Sigma_{m,n}(\cdot,\cdot)$.

Consider the alternative hypothesis

$$H_a:\ \Big|E\big[L_{m,t+1}(\gamma^{\dagger})\big]\Big|\geq\Delta>0,\quad\text{for some }\gamma^{\dagger}\in\Gamma. \tag{3}$$

The following result establishes inference under the null and alternative hypothesis.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied, and let $\Sigma_{m,n}(\cdot,\cdot)$ be nondegenerate. Under H_0 it follows that $P\left(v(t_{m,n})>c_m(1-\alpha)\right)\to\alpha=P\left(v(\tilde{t}_{m,n})>c_m(1-\alpha)\right)$, where $c_m(1-\alpha)$ is chosen such that $P\left(v(\tilde{t}_{m,n})>c_m(1-\alpha)\right)=\alpha$. Moreover, let $\Gamma^{\dagger}=\{\gamma:|\gamma-\gamma^{\dagger}|^{\lambda}<\Delta/B\}$ have positive J-measure. Under H_a it follows that $v(t_{m,n})\xrightarrow{d}\infty$, and $P\big(v(t_{m,n})>c_m(1-\alpha)\big)\to1$.*

We establish the consistency of the blockwise bootstrap for general empirical processes of

Bühlmann (1995) for $v(t_{m,n})$. The blockwise bootstrap was first studied by Künsch (1989) for general stationary observations.

The resampling scheme of the blockwise bootstrap uses overlapping blocks of consecutive observations. Let the block length be noted by $l = l(n)$, and let $k = n/l$. We then consider a bootstrap sample $\{L^*_{m,m+1}, \ldots, L^*_{m,T}\}$, with the first block of l observations $\{L^*_{m,m+1}, \ldots, L^*_{m,m+l}\}$ drawn as $\{L_{m,V_1+1}, \ldots, L^*_{m,V_1+l}\}$, the second block of l observations $\{L^*_{m,m+(2-1)l+1,1}, \ldots, L^*_{m,m+2l}\}$ drawn as $\{L_{m,V_2+1}, \ldots, L^*_{m,V_2+l}\}$, and so on up to the kth block of l observations $\{L^*_{m,m+(k-1)l+1,1}$, $\ldots, L^*_{m,m+kl}\}$ drawn as $\{L_{m,V_k+1}, \ldots, L^*_{m,V_k+l}\}$, with the V_i iid uniform distributed variables on $\{m, m+1, \ldots, m+n-l\}$. When $kl > n$ occurs in practice, we remove the superfluous observations at the end of the kth block of bootstrap observations.

The blockwise bootstrap counterpart of $\sqrt{n}\bar{L}_{m,n}(\gamma)$ is then given by

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

with $\mu^*_n(\gamma) \equiv \frac{1}{n-l+1}\sum_{i=1}^{n-l+1}\frac{1}{l}\sum_{t=m+i}^{i+l-1} L_{m,t}(\gamma)$ denoting the expectation of $\frac{1}{n}\sum_{t=m}^{T-1} L^*_{m,t+1}(\gamma)$ conditional on the original sample. Similarly, let $t^*_{m,n}(\gamma) \equiv \sqrt{n}\bar{L}^*_{m,n}(\gamma)/\hat{\sigma}_{m,n}(\gamma)$, and let $c^*_{m,n}(1-\alpha)$ denote the α-quantile of $t^*_{m,n}(\gamma)$, conditional on the data.

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

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

The following result establishes consistency of the bootstrap.

Theorem 3. *Let the assumptions of Theorem 1 and Assumption 5 hold. Under H_0 it follows that $\sqrt{n}\bar{L}^*_{m,n}(\cdot) \Rightarrow Z_{m,n}(\cdot)$ almost surely. Moreover, $P(v(t_{m,n}) > c^*_{m,n}(1-\alpha)) \to \alpha$.*

Theorem 3 suggests that we can approximate $c^*_{m,n}(1-\alpha)$ through Monte-Carlo simulation. Specifically, let $c^B_{m,n}(1-\alpha)$ denote the $\alpha \cdot 100\%$ percentile of the test statistics $v(t^{*(1)}_{m,n}(\gamma)), \ldots,$ $v(t^{*(B)}_{m,n}(\gamma))$ obtained from $B < \infty$ bootstrap samples. As $B \to \infty$, $c^B_{m,n}(1-\alpha)$ becomes arbitrarily close to $c^*_{m,n}(1-\alpha)$.

3.2 Superior predictive ability tests

We now consider the one-sided hypothesis H'_0 and its alternative

$$H'_a: \; E\big[L_{m,t+1}(\gamma^\dagger)\big] \geq \Delta > 0, \quad \text{for some } \gamma^\dagger \in \Gamma. \tag{4}$$

Notice that H_0 is the element of H_0' least favorable to the alternative. We can therefore construct a valid test of H_0' from

$$\sup t_{m,n} \equiv \sup_{\gamma \in \Gamma} t_{m,n}(\gamma).$$

More specifically, the results under H_0 in Theorems 1, 2, and 3 still hold for H_0' for a demeaned version of $\sup t_{m,n}$, denoted $\sup \tau_{m,n} \equiv \sqrt{n}\frac{\bar{L}_{m,n}(\gamma) - E[\bar{L}_{m,n}(\gamma)]}{\hat{\sigma}_{m,n}(\gamma)}$. We must use $\sup \tau_{m,n}$ because H_0' allows for $E[\bar{L}_{m,n}(\gamma)] \leq 0$. Notice that we can still obtain valid critical values under H_0' from Theorem 3. On the other hand, the result in Theorem 2 under H_a still holds under H_a' for $\sup t_{m,n}$ (not just for $\sup \tau_{m,n}$), which implies that $\sup t_{m,n}$ still has asymptotic power against fixed alternatives.

3.3 Practical implementation

When Γ contains infinitely many elements we cannot evaluate the test statistics over all elements of Γ in practice. We therefore provide two numerical approximations $v(t_{m,n})$, for which Theorems 1 and 2 remain valid.

We first look at discretizations of Γ that becomes increasingly dense. Consider a d-dimensional 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 approximation to $v(t_{m,n})$:

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

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

$$\widehat{\sup t_{m,n}} \equiv \max_{i=1,\ldots,K_n} t_{m,n}(\gamma_{n,i}).$$

If we let Γ be hyperrectangular 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, $v_n \to \infty$, 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 easy to implement. Other choices of J might 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 can grow quickly with large d. As a result the calculation of $\widehat{\text{ave}\, t_{m,n}^2}$, $\widehat{\sup t_{m,n}^2}$, and $\widehat{\sup t_{m,n}}$ becomes problematic as n grows. In those scenarios we can instead use Monte Carlo simulation from J to obtain an approximation of

$v(t_{m,n})$. Consider S_n independent draws $\gamma^{(i)}$ from J, $i = 1, \ldots, S_n$, and approximations

$$\widetilde{\text{ave}\, t^2_{m,n}} \equiv \frac{1}{S_n} \sum_{i=1}^{S_n} t^2_{m,n}(\gamma^{(i)}),$$

$$\widetilde{\sup t^2_{m,n}} \equiv \max_{i=1,\ldots,S_n} t^2_{m,n}(\gamma^{(i)}), \text{and}$$

$$\widetilde{\sup t_{m,n}} \equiv \max_{i=1,\ldots,S_n} t_{m,n}(\gamma^{(i)}).$$

Lemma 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 $n \to \infty$, $\widetilde{v(t_{m,n})} \xrightarrow{p} v(t_{m,n})$. Moreover, for $S_n \to \infty$, it follows that $\widetilde{v(t_{m,n})} \xrightarrow{p} v(t_{m,n})$, with probability approaching one under the J-measure.*

3.4 Benchmarks

We consider standard joint Wald tests and the Bonferonni multiple-comparison correction as benchmarks to the tests proposed above. It should be noted that these tests can only test at $M < \infty$ number of points in Γ, and therefore generally cannot uniformly test over Γ in our setup.

Consider some discrete parameter set Γ_M. We define the standard Wald test statistic as

$$\hat{Q}^h_{m,n} \equiv n \tilde{L}_{m,n}(\Gamma_M)' \hat{\Omega}^{-1}_{M,m,n} \tilde{L}_{m,n}(\Gamma_M), \tag{5}$$

where $\tilde{L}_{m,n}(\Gamma_M) \equiv \left(\bar{L}_{m,n}(\gamma_1)', \ldots, \bar{L}_{m,n}(\gamma_M)'\right)'$, and $\hat{\Omega}_{M,m,n}$ is some HAC estimator of the asymptotic covariance matrix of $\sqrt{n}\tilde{L}_{m,n}(\Gamma_M)$. We use the HAC estimator of Newey and West (1987) since it is commonly used in practice.

A two-sided α-level test rejects the null hypothesis when $\hat{Q}^h_{m,n} > \chi^2_{qM,1-\alpha}$, where $\chi^2_{qM,1-\alpha}$ denotes the $(1-\alpha)$-quantile of a χ^2 distribution with qM degrees of freedom. As M increases $\hat{\Omega}_{M,m,n}$ becomes near-singular, and singular if $M > n$. This inflates the test statistic in small samples.

A two-sided α-level test using the Bonferonni correction rejects the null hypothesis if, for at least one $\gamma \in \Gamma_M$, we find $n\bar{L}^2_{m,n}(\gamma)/\tilde{\sigma}^2_{m,n} > \chi^2_{q,1-\alpha/M}$, with $\tilde{\sigma}_{m,n}$ a HAC asymptotic covariance estimator of $\sqrt{n}\bar{L}_{m,n}(\gamma)$. Contrary to the standard multivariate Wald test, the test using the Bonferonni correction becomes conservative as M increases, or when the test statistics are positively correlated.

A one-sided α-level test using the Bonferonni correction rejects the null hypothesis if, for at

least one $\gamma \in \Gamma_M$, we find $n\bar{L}_{m,n}(\gamma)/\tilde{\sigma}_{m,n} > z^{-1}_{1-\alpha/M}$, with $z^{-1}_{1-\alpha}$ denoting the $(1-\alpha)$-quantile of the standard normal distribution.

4 Simulation studies

4.1 Simulation study I: Expected utility of portfolio strategies

In an application of Example 1 we evaluate the difference in expected utility of two commonly used portfolio management strategies, being the naive (*or* equally-weighted) portfolio and the minimum variance (mv) portfolio. For a N-length vector of monthly excess returns Y_{t+1} these strategies can be defined in terms of portfolio weight vectors:

$$\begin{aligned} w^{\text{naive}}_{t+1} &= \frac{1}{N}\iota, \\ w^{\text{mv}}_{t+1} &= \frac{1}{\iota'\Sigma^{-1}_{t+1}\iota}\Sigma^{-1}_{t+1}\iota, \end{aligned} \tag{6}$$

with $\Sigma_{t+1} = \text{Cov}(Y_{t+1})$, and ι a N-length vector of ones. Denote the portfolio returns as $Y^{\text{naive}}_{t+1} = w^{\text{naive}\prime}_{t+1} Y_{t+1}$, and $Y^{\text{mv}}_{t+1} = w^{\text{mv}\prime}_{t+1} Y_{t+1}$.

The feasible counterpart of w^{mv}_{t+1}, depends on an estimate of Σ_{t+1}. We follow convention and use the unbiased moving average sample covariance matrix based on the prior $m$observations.

DeMiguel et al. (2007) compare these strategies, and several others, in terms of Sharpe ratio, certainty equivalent return (CEQ), and turnover, which are commonly used evaluation criteria in portfolio management. Here we will test whether the naive and minimum-variance portfolio returns have equivalent (superior) expected utility, which is similar to testing for equivalent (superior) CEQ.

We model utility with the exponential utility function $u(y;\gamma) = -e^{-\gamma y}/\gamma$. Exponential utility investors with normally distributed returns are mean-variance investors (see, e.g. Sargent (1979, p. 150-151)), such that we are close to the mean-variance investors considered in DeMiguel et al. (2007).

There is debate about reasonable values of the risk aversion parameter γ. Bliss and Panigirtzoglou (2004) provide an overview of estimates of γ in the literature, which range from close to zero to as high as 60. Bliss and Panigirtzoglou (2004) themselves find full sample point estimates between 2.98 to 10.56 in their Table VI. Meanwhile, DeMiguel et al. (2007) perform their comparisons for investors with with risk aversion ranging between 1 and 10. We will therefore test

for equal expected utility over $\Gamma = [1, 10]$, i.e. $E[L_{t+1}(\gamma)] = E[u(1+Y_{t+1}^{\text{naive}};\gamma) - u(1+Y_{t+1}^{\text{mv}};\gamma)] = 0$, for all $\gamma \in \Gamma$. We draw uniformly over Γ for the ave-test.

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

$$\begin{aligned} Y_{i,t+1} &= \alpha_i + \beta_i Y_{t+1}^f + \eta_{i,t+1}, \\ \nu_{i,t+1} &\sim \text{iid}\,\mathcal{N}(0, \sigma_{\eta,i}^2), \\ Y_{t+1}^f &\sim \text{iid}\,\mathcal{N}(\mu_f, \sigma_f^2). \end{aligned} \tag{7}$$

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 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 generally cannot obtain a parameterization m that implies $E[L_{t+1}(\gamma)] = 0$ for all $\gamma \in \Gamma$. In the size experiment we therefore test the null hypotheses at $\zeta_m(\gamma) = E[L_{t+1}(\gamma)]$ instead of zero, and we estimate $\zeta_m(\gamma)$ using a repeated simulation of $\bar{L}_{m,n}(\gamma)$.

4.2 Simulation study II: Tail quantile forecasts of portfolio returns

We study a scenario similar to Example 2 by formally comparing tail quantile forecasts of portfolio returns as generated by multivariate models. We model the financial assets using a GARCH-DCC model (Engle, 2002) with normal errors, parameterized to resemble the properties of daily asset returns.

We compare two models: (i) a fully parameterized GARCH-DCC model with normal errors, and (ii) a multivariate normal distribution with the Riskmetrics covariance estimator (Riskmetrics, 1996), i.e. an exponentially weighted moving average covariance estimator. Both models are frequently used in practice. We are therefore relatively confident that the simulated loss

differences will resemble those observed in practice.

We let the N-dimensional return vector Y_{t+1} follow a GARCH-DCC process

$$\begin{aligned}
Y_{t+1} &= \mu_t + H_{t+1}^{1/2} C_{t+1}^{1/2} \nu_{t+1}, \\
\nu_{t+1} &= (\nu_{t+1,1}, \ldots, \nu_{t+1,N})' \sim iid\, N(0, I), \\
H_{t+1} &= \operatorname{diag}\left(h_{t+1,1}, \ldots, h_{t+1,N}\right), \\
h_{t+1,i} &= \omega_0 + \omega_1 h_{t,i} + \omega_2 h_{t,i} \nu_{t,i}^2, \\
C_{t+1} &= \operatorname{diag}\left(\tilde{C}_{t+1}\right)^{-1/2} \tilde{C}_{t+1} \operatorname{diag}\left(\tilde{C}_{t+1}\right)^{-1/2}, \\
\tilde{C}_{t+1} &= (1 - \xi_1 - \xi_2) C + \xi_1 \tilde{C}_t + \xi_2 \nu_t \nu_t', \\
C &= [C]_{ij}, \text{ with } [C]_{ii} = 1 - \frac{|i-j|}{N}.
\end{aligned} \tag{8}$$

We choose GARCH parameters $\omega_0 = 0.05$, $\omega_1 = 0.10$, $\omega_2 = 0.85$, to closely match conditional volatility patterns in daily equity returns, and we choose DCC parameters $\xi_1 = 0.025$, $\xi_2 = 0.95$, to closely match time-varying correlation patterns commonly found in daily equity returns. We set $\mu_t = 0$, since average daily returns are close to zero.

In C we recognize 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 C.

We are interested in one-period-ahead τ-quantile forecasts for portfolio returns $Y_{t+1}^p(\gamma) = \gamma' Y_{t+1}$, with $\tau = 5\%$, for all $\gamma \in \Gamma$. The forecasts are given as follows.

The GARCH-DCC forecast is the perfect forecast, and is generated by the GARCH-DCC model without estimation error. We do not estimate the parameters, because doing so would be excessively costly in terms of computation time. The forecast is given by

$$Q_{t+1,\tau}^{(1)}(\gamma) = z_\tau^{-1} \cdot \gamma' H_t^{1/2} C_t H_t^{1/2} \gamma, \tag{9}$$

with z_τ^{-1} denoting the τ-quantile of the standard normal distribution.

The Riskmetrics forecasting method provides the following forecast

$$Q_{t+1,\tau}^{(2)}(\gamma) = z_\tau^{-1} \cdot \gamma' \hat{\Sigma}_{m,t+1} \gamma, \tag{10}$$

where $\hat{\Sigma}_{m,t+1}$ is given by

$$\hat{\Sigma}_{m,t+1} = c_l(1-\lambda)\sum_{j=0}^{m}\lambda^j\left(Y_{t-j+1}-\hat{\mu}_t\right)\left(Y_{t-j+1}-\hat{\mu}_t\right)$$

with $\hat{\mu}_t = (\hat{\mu}_{1t},\ldots,\hat{\mu}_{Nt})' = \frac{1}{m}\sum_{j=1}^{m} Y_{t-j+1}$, and c_l a constant that normalizes the summed weights $(1-\lambda)\sum_{j=0}^{m}\lambda^j$ to one. We choose $\lambda = 0.94$ to follow convention.

The covariance matrix estimator $\hat{\Sigma}_{m,t+1}$ is the moving window analogue of the exponentially weighted moving average Riskmetrics (1996) estimator obtained from the recursion $\Sigma_{t+1} = (1-\lambda)\left(Y_t-\mu_t\right)\left(Y_t-\mu_t\right) + \lambda\Sigma_t$.

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

$$\tilde{L}_{t+1}^{(i)}(\gamma) = \left(\tau - \mathbb{1}\left\{Y_{t+1}^{p}(\gamma) < Q_{t+1,\tau}^{(i)}(\gamma)\right\}\right)\left(Y_{t+1}^{p}(\gamma) < Q_{t+1,\tau}^{(i)}(\gamma)\right), \tag{11}$$

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

We subtract the losses of the GARCH-DCC model from the Riskmetrics model, because the Riskmetrics models has larger mean losses by definition, due to the GARCH-DCC model having no estimation error. As a result the one-sided tests should have power against H_0' for this ordering.

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 also known as 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.

Finally, in our size experiment the mean loss differences must equal zero. We do not however search for parameterizations that agree with the unconditional null hypothesis, because this is a computationally infeasible. Rather, we test the null hypotheses at $\zeta_m(\gamma) = E[\bar{L}_{m,n}(\gamma)]$ instead of the zero vector. For each γ under consideration we obtain an estimate of $\zeta_m(\gamma)$ through repeated simulation of $\bar{L}_{m,n}(\gamma)$.

4.3 Simulation study III: Murphy diagrams of tail quantile forecasts

We study a scenario similar to Example 3. We test for differences in predictive ability of τ-quantile forecasts of an individual return using a multiple hypothesis based on *elementary loss*

functions.

Let Y_{t+1} be some single asset return (i.e. $N = 1$). Under mild regularity conditions (see, e.g., Gneiting (2011)) the (conditional) τ-quantile of Y_{t+1} admits a class of consistent loss functions given by:

$$S(Y_{t+1}, x) = \big(\mathbb{1}(Y_{T+1} < x) - \tau\big)\big(g(Y_{t+1}) - g(x)\big), \tag{12}$$

for all non-decreasing functions $g(\cdot)$; i.e. $E_t[S(Y_{t+1}, Q_t)] \leq E_t[S(Y_{t+1}, x)]$, for Q_t the true τ-quantile, and all x in the domain of Q_t. A commonly used member of this family is the tick-loss function (see Equation (11)) which sets $g(z) = z$, and which we use in Simulation study II.

Ehm et al. (2016) show that each $S(Y_{t+1}, x)$ has a mixture representation

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

where

$$\tilde{S}(Y_{t+1}, x; \gamma) = \big(\mathbb{1}(Y_{T+1} < x) - \tau\big)\big(\mathbb{1}(\gamma < x) - \mathbb{1}(\gamma < Y_{t+1})\big),$$

and with H some non-negative measure, which is unique given some $g(\cdot)$. Moreover, $S_\tau(Y_{t+1}, x)$ is strictly consistent if H assigns positive mass to some finite interval. It follows that the elementary loss functions $\tilde{S}(Y_{t+1}, x; \gamma)$ are strictly consistent loss functions by selecting the H assigning point mass to γ.

Consider two forecasts $Q^{(1)}_{t+1,\tau}$ and $Q^{(2)}_{t+1,\tau}(\gamma)$. We say that $Q^{(2)}_{t+1,\tau}$ dominates (is superior to) $Q^{(1)}_{t+1,\tau}(\gamma)$ if $E_t[S(Y_{t+1}, Q^{(2)}_{t+1,\tau}) - S(Y_{t+1}, Q^{(1)}_{t+1,\tau})] \leq 0$ for all members $S(\cdot)$ of the family of consistent loss functions for the τ-quantile. Corollary 1 in Ehm et al. (2016) establishes that $Q^{(2)}_{t+1,\tau}(\gamma)$ dominating $Q^{(1)}_{t+1,\tau}(\gamma)$ is implied by $E_t[S(Y_{t+1}, Q^{(2)}_{t+1,\tau}; \gamma) - \tilde{S}(Y_{t+1}, Q^{(2)}_{t+1,\tau}; \gamma)] \leq 0$, for all $\gamma \in \mathbb{R}$. We can therefore test for equal (superior) predictive ability using the joint hypothesis $E_t[S(Y_{t+1}, Q^{(2)}_{t+1,\tau}; \gamma) - \tilde{S}(Y_{t+1}, Q^{(1)}_{t+1,\tau}; \gamma)] = 0$ ($E_t[S(Y_{t+1}, Q^{(2)}_{t+1,\tau}; \gamma) - \tilde{S}(Y_{t+1}, Q^{(1)}_{t+1,\tau}; \gamma)] \leq 0$), for all $\gamma \in \Gamma$.

It should be noted that our theory requires Γ to be bounded, such that 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 $\tilde{S}(Y_{t+1}, Q^{(2)}_{t+1,\tau}; \gamma) - \tilde{S}(Y_{t+1}, Q^{(1)}_{t+1,\tau}; \gamma) = 0$ for $\gamma > \max\big(Y_{t+1}, Q^{(1)}_{t+1,\tau}, Q^{(2)}_{t+1,\tau}\big)$ and $\gamma < \min\big(Y_{t+1}, Q^{(1)}_{t+1,\tau}, Q^{(2)}_{t+1,\tau}\big)$.

Finally, in small samples it can occur that $\tilde{S}(Y_{t+1}, Q^{(2)}_{t+1,\tau}; \gamma) - \tilde{S}(Y_{t+1}, Q^{(1)}_{t+1,\tau}; \gamma) = 0$ for all $t = T+1, \ldots, m$, for some elements of Γ. As a result, $\hat{\sigma}^2_{m,n}(\gamma) = 0$ for these values of γ. To circumvent this we calibrate $\sigma^2_{m,n}(\gamma) = 1$ for all $\gamma \in \Gamma$, i.e. we consider sup- and ave-tests based on $t_{m,n}(\gamma) \equiv \sqrt{n}\bar{L}_{m,n}(\gamma)$ instead of $t_{m,n}(\gamma) \equiv \sqrt{n}\frac{\bar{L}_{m,n}(\gamma)}{\hat{\sigma}_{m,n}(\gamma)}$. The p-values remain valid under the bootstrap. The HAC covariance estimators used in the calculation of the multivariate Wald test and the Bonferonni-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. Instead of using the multivariate Wald test and the Bonferonni test we therefore use the Diebold-Mariano test based on the tick-loss function as benchmark.

We use the same quantile forecast models as in Simulation study II but set $N = 1$, such that the quantile forecasts defined in Equations (9) and (10) are effectively obtained from a GARCH model instead of a GARCH-DCC model and a univariate Riskmetrics model. We also consider a moving sample quantile estimated over the previous 250 days.

In the size experiment the mean loss differences must equal zero. Again, we do not search for parameterizations that agree with the unconditional null hypothesis, because of computational infeasibility. Rather, we test the null hypotheses at $\zeta_m(\gamma) = E[\bar{L}_{m,n}(\gamma)]$ instead of the zero vector. For each γ under consideration we obtain an estimate of $\zeta_m(\gamma)$ through repeated simulation of $\bar{L}_{m,n}(\gamma)$.

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. The minimum-variance portfolio weights are estimated using data from the previous $m = 120$ months, amounting to ten years of data. We consider out-of-sample period lengths $n = 120$ and 600 months. Sixty years of data ($m+n = 720$) is about the maximum sample size encountered in portfolio applications. We consider increasingly large grids of Γ, with the number of grid points set to $K_n = 1, 10, 50, 100$, and 250. Finally, we obtain critical values using $B = 1{,}000$ bootstrap samples.

From Panel A we observe that the sup-t^2 and ave-t^2 tests are slightly oversized for $n = 120$ out-of-sample observations, and appropriately sized for $n = 600$. We see that the benchmark tests also suffer from larger than nominal size at $K_n = 1$, which suggests that the size distortion is caused by one or more individual Diebold-Mariano test statistics–which the sup- and ave-tests are functions of–rather than the specific functional forms of our test statistics.

Importantly, we observe that the sup-t^2 and ave-t^2 tests are stable in terms of rejection rates as K_n increases, which suggests that the researcher can choose a relatively large value for K_n and be confident that test conclusions are stable for increases beyond this value.

We also observe the expected behavior from the benchmark tests as K_n increases; with the Wald test having large size distortions, and the Bonferonni corrected test becoming conservative. In practice we often do not know what the appropriate value of K_n is, such that the standard Wald tests is generally unreliable for some arbitrarily chosen K_n. The increasingly conservative rejection rates of the Bonferonni-corrected test suggest that K_n can also not be set to some arbitrary large number.

The one-sided tests show largely similar behavior, with the sup-t test being correctly sized for $n = 600$, and the Bonferonni test becoming conservative as K_n increases.

Panel B shows rejection rates in the power experiment. We observe similar behavior of the tests, with the sup- and ave-tests being stable for large values of K_n, and the Bonferonni-corrected test rejection rates decreasing with K_n. As concerns the standard Wald test we find that the rejection rates in the power experiment reduces with K_n.

Given that the choice of K_n can influence test conclusions, we now look at power properties at $K_n = 250$. We do so, because *ex-ante* we generally do not know the appropriate value of K_n. This leaves the test vulnerable to the critique of p-hacking through K_n. At $n = 120$ the sup-tests have largest power in the experiment with approximately 95% of alternatives rejected, which is quite a bit larger than the Bonferonni-corrected tests, which are the secondmost powerful with approximately 0.54 at large K_n. At $n = 600$ all tests, except for the standard Wald test, reject all cases.

[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 $N = 30$ assets. The rejection rates are based on 1,000 Monte-Carlo simulations. Results are presented for out-of-sample periods of $n = 500$ and 2,000 days, and six sets of weight vectors, which are drawn as follows. We first consider a set of $N + 1$ deterministic weight vectors, being the equally weighted portfolio vector and the h 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. The Riskmetrics models is estimated over $m = 75$ lags to follow convention.

Panel A provides rejection rates in the size experiment. We observe that the sup- and ave-tests have good size properties, although the tests are slightly conservative for n =2,000 (rejection rates between 1-2%). The same holds for the Bonferonni-corrected tests at all S_n. The standard Wald test is oversized for all S_n at n = 500, and for all S_n larger than 31 at n =2,000, which illustrates the unreliability of the standard Wald test, even in moderately large samples. The rejection rates of the sup- and ave-tests are stable as S_n increases.

Panel B shows rejection rates in the power experiment. We observe that sup-t^2 test is most powerful amongst the two-sided tests at S_n =1,000, which is the largest value of S_n under consideration. The sup-t test is most powerful amongst one-sided tests at S_n =1,000. The Bonferonni-corrected test power decreases with S_n. The standard Wald test is unreliable, as shown in Panel A.

One interesting observation is that the ave-t^2 tests at S_n = 31 is most powerful for all correctly sized tests and S_n under consideration, and considerably so. At S_n = 31, we test the multiple hypothesis at the equally weighted portfolio return and the individual asset returns. Because these returns are specifically selected, instead of the product of randomly drawn portfolio vectors or some arbitrary point on a grid, we could opt for using this particular test instead of setting S_n large. In the empirical study we find some evidence for this ranking as well, although the sup-t^2 test rejects similarly (and for all S_n large) for the empirical data set under consideration.

[Table 2 about here.]

Table 3 provides small sample rejection rates of the tests in size and power experiments introduced in Section 4.3, based on 1,000 Monte-Carlo simulations. Moreover, we consider grids of Γ with $K_n = 10, 50, 100$, and 250 grid points equally spaced over the interval $[-20, 0]$. We select this interval because outside of it the elementary score differences are generally equal to zero.

Panel A provides rejection rates for the size experiment in which we compare the GARCH and Riskmetrics forecasts. 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, a Diebold-Mariano test using the tick-loss function, is also correctly sized. The rejection rates of the sup- and ave-tests are all stable for larger values of K_n.

Panel B provides rejection rates for the size experiment where we compare the GARCH and

the moving sample quantile forecasts. We find that the two-sided Diebold-Mariano test and the ave-test are oversized, whereas the sup-tests are less so for larger values of K_n.

Panel C provides rejection rates in the power experiment in which we compare the GARCH and Riskmetrics forecasts. We observe that the particular alternative DGP is not very distant from the null, in that the ave- and sup-tests have no power at $n = 500$. The benchmark test has larger power. At $n = 2,000$ our tests have power against the althernative, but again the benchmark test is more powerful. The ranking in terms of power of the tests contradicts with what we find in the empirical study in Section 5.3, which shows the sup- and ave-tests reject much more often than the benchmark test.

Panel D provides rejection rate in the power experiment between the GARCH and moving sample quantile forecasts. The sup- and ave-tests reject more often, and have similar power properties to the benchmark.

[Table 3 about here.]

5 Empirical results

We apply the tests of equal (superior) predictive ability and expected utility to empirical data. Specifically, we will use daily and monthly returns data on 30 U.S. Industry portfolios collected by Ken French, which are freely available on his website.[1]

We use a sample of monthly returns running from September 1926, the first available monthly return, to December 2017 to test for equal utility of the naive and minimum-variance portfolios. The full sample consists of 1,098 monthly return observations per portfolio.

We use twenty years of daily returns, running from January 2, 1998, to December 29, 2017, to test for equal predictive ability of tail quantile forecasts. The full sample consists of 5,032 daily return observations per portfolio. We also consider subsamples consisting of the first and second halves of the full sample.

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

5.1 Expected utility of minimum-variance and equally-weighted portfolio strategies

In this empirical illustration we test the expected utility hypotheses obtained from naive and minimum-variance portfolio strategies, as applied to monthly industry portfolio returns. The minimum-variance portfolio weights are estimated over moving windows of $m = 120$ months. Table 4 provides p-values for the sup- and ave-tests as well as the benchmark tests. Results are presented for several choices of the risk aversion parameter interval Γ, being $\Gamma = [1, 10]$, $[1, 5]$, and $[5, 10]$. We consider these intervals to see how the tests are affected by different choices of Γ.

We first discuss results for $\Gamma = [1, 10]$, which covers all risk aversion parameter values considered in DeMiguel et al. (2007). At $K_n = 1$ we test the hypothesis of equal (superior) expected utility for a exponential utility investor with risk aversion $\gamma = 2.5$. We find that the hypothesis is rejected using all tests. When we increase K_n, and test the multiple hypothesis of equal (superior) expected utility over the entirety of Γ, we find that the sup-t^2 and sup-t tests still reject the equal and superior expected utility hypotheses. The ave-t^2 test fails to reject the null hypothesis, but, like the sup-tests, has stable p-values for larger values of K_n.

On the other hand, the Bonferonni-corrected test cannot reject the null hypothesis for larger K_n. This finding agrees with the increasing conservativeness with K_n that we find in the simulation study. The standard 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 individual 95% confidence bounds for each $\gamma \in [1, 10]$. We observe that for smaller values of γ, i.e. $1 < \gamma < 3$, the confidence interval of the utility differential lies above the zero line, whereas for larger values of γ, the individual confidence intervals either contain zero, or lie below zero. This suggests that investors with greater risk appetite (lower γ) prefer the naive portfolio over the minimum-variance portfolio, whereas investors with lesser risk appetite prefer the minimum-variance portfolio. Since the sup-t^2 test consider the largest deviation from the zero line, and the ave-t^2 considers the averages over all deviations[2], it is intuitive that the ave-t^2 test statistics is pushed towards zero more strongly due to the large subinterval of Γ for which the mean utility

[2]Formally, the tests consider the deviation divided by its standard error, but here the standard errors are quite stable over Γ.

differentials is insignificantly different from zero. Moreover, by including this subinterval Γ, we inflate the p-value of the Bonferonni-corrected test, by including more elements close to zero in the multiple hypothesis.

Panels B and C of Table 4 show results for $\Gamma = [1, 5]$, and $\Gamma = [5, 10]$, respectively. Given that we only care for investors with larger risk appetite ($\Gamma = [1, 5]$), we see that the ave-t^2 also rejects the null-hypothesis. On the other hand, if we consider only risk averse investors ($\Gamma = [5, 10]$) we see that the null hypothesis cannot be rejected by the ave- and sup-tests. Conclusions drawn from the Bonferonni-corrected tests do not change much, with both tests failing to reject the null hypothesis for larger values of K_n.

[Figure 1 about here.]

[Table 4 about here.]

5.2 Tail quantile forecasts of equity portfolio returns from GARCH-DCC and Riskmetrics models

Table 5 presents p-values of the joint tests of equal (superior) predictive ability applied to quantile forecasts of GARCH-DCC models and Riskmetrics models for the daily industry portfolio returns. The GARCH-DCC model is now estimated over a moving window of the previous 1,000 observations. We differ from the simulation study by considering the GARCH-DCC models as the benchmark, such that a rejection of the one-sided test suggests that the GARCH-DCC model does not have equal or lower losses than the Riskmetrics model for all portfolios under consideration.

We show results for the full sample, and the sub-samples consisting of the first and second half of observations. The results for the full sample show that the models are competitive, since the null hypotheses of equal and superior predictive ability cannot be rejected. This suggests that the GARCH-DCC model has superior or equal predictive ability in comparison to the Riskmetrics model. Moreover, we see that sup- and ave-tests generate p-values that are stable for larger values of S_n, whereas the Bonferonni p-values become conservative with S_n. The Wald tests are unreliable, as shown in the simulation exercises, and are not reported.

The results for the first sub-sample show that the hypotheses of equal and superior predictive ability are rejected by the sup-tests. The superior predictive ability test rejects that the

GARCH-DCC model is superior or equal to the Riskmetrics model for all portfolio vectors in Γ. In the second sub-sample we cannot reject the hypotheses of equal or superior predictive ability for any test. Again the sup- and ave-tests, have stable p-values for larger S_n, whereas the Bonferonni p-values increase with S_n.

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 first subsample, and sorted on mean tick loss difference. We see that the GARCH-DCC forecasts perform better, although for only few portfolio vectors we observe have confidence intervals that exclude zero. The sup-tests detect the outperformance of the GARCH-DCC forecasts better than the ave-test as observed in Table 5.

[Figure 2 about here.]

[Table 5 about here.]

5.3 Murphy diagrams of tail quantile forecasts

Table 6 presents the ratio of rejections of the equal (superior) predictive ability hypothesis of quantile forecast models as defined in Section 4.3, over a cross-section of the 30 industry portfolios,. The quantile forecasts are generated by the univariate GARCH model, the Riskmetrics model, and the sample quantile calculated over a moving window of the previous 250 days. The GARCH model is estimated over a moving window of 1,000 observations. The Riskmetrics model is estimated over a moving window of 75 days.

Panel A and B present p-values for the tests applied to the Transportation portfolio, as a representative scenario for the full set of industry portfolios. We provide the cross-sectional ratio of rejections accross all industry portfolios below. Panel A shows that the sup tests reject the hypotheses of equal and superior predictive ability between the GARCH and Riskmetrics models, and that p values are quite stable over time. The ave-test is not significant at the 5% level, which hold for the tick-loss based Diebold-Mariano tests as well. Panel B shows that the 250-day moving sample quantile forecasts underperforms substantially, with all tests rejecting equal or superior predictive ability in comparison with GARCH forecasts.

Panel C and D present the cross-section rejection ratios. In Panel C we observe that the ave- and sup-tests reject the null hypotheses of equal (superior) predictive ability between GARCH

and Riskmetrics forecasts for 13% to 29% of the considered industry portfolios, whereas the benchmark Diebold-Mariano test using the tick-loss function rejects for none of the industry portfolios. Unlike in previous empirical examples, the ave-t^2 tests rejects more often than the sup-t^2 test. Moreover, the ratio of rejections is relative stable for larger values of K_n. The results suggest that sup- and ave-tests of the joint hypothesis based on elemental loss functions are more sensitive to deviations from the null than the commonly used Diebold-Mariano test based on the tick-loss function, and could therefore provide an interesting tool for empirical research.

In Panel D we compare the GARCH model with the sample quantil estimated over a moving window of 250 days. The latter model is generally considered to be much less competitive. Indeed Panel B shows that all tests reject the null hypotheses for all industry portfolios under consideration.

Figures 3 show Murphy diagrams for the two comparisons, applied to the Transportation portfolio, and show that the GARCH and Riskmetrics forecasts are relatively competitive, whereas the GARCH forecast is superior to the moving sample quantile forecasts.

[Figure 3 about here.]

[Table 6 about here.]

6 Concluding remarks

We develop out-of-sample tests for joint testing problems that are defined over a continuum of moment conditions, and which are applicable to time-series data – as required in many forecasting scenarios. We apply our tests to relevant forecasting evaluation problems and show that the new tests reject at least as often as benchmarks tests. Moreover, our tests are better behaved than the benchmarks in that p-values remain stable as we test at more elements of the joint hypothesis. We use simulation experiments to show that our tests have good size and power properties in small sample.

References

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

Andrews, D. W. (1994). Chapter 37 Empirical Process Methods in Econometrics. volume 4 of *Handbook of Econometrics*, pages 2247–2294. Elsevier.

Bliss, R. R. and Panigirtzoglou, N. (2004). Option-Implied Risk Aversion Estimates. *The 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.

Clark, T. E. and McCracken, M. W. (2001). Tests of Equal Forecast Accuracy and Encompassing for Nested Models. *Journal of Econometrics*, 105(1):85–110.

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 Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, 22(5):1915–1953.

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

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

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 & Economic Statistics*, 20(3):339–350.

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

Escanciano, J. C. and Olmo, J. (2010). Backtesting Parametric Value-at-Risk With Estimation Risk. *Journal of Business & Economic Statistics*, 28(1):36–51.

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

Fleming, J., Kirby, C., and Ostdiek, B. (2003). The Economic Value of Volatility Timing Using "Realized" Volatility. *Journal of Financial Economics*, 67(3):473–509.

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

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

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 & Economic Statistics*, 23(4):365–380.

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. *The 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.

McCracken, M. W. (2000). Robust Out-of-Sample Inference. *Journal of Econometrics*, 99(2):195–223.

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

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.

Sargent, T. (1979). *Macroeconomic Theory*. Economic Theory, Econometrics and Mathematical Economics Series. Academic Press.

West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica*, pages 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.

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 $\sqrt{n}\bar{L}_{m,n}(\cdot)$ 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 $\sqrt{n}\bar{L}_{m,n}(\cdot)$.

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

Setting $N(\delta) = \delta^{-d/\gamma} G^d B^{-d/\lambda}$, we therefore find that for all $\gamma \in \Gamma$ there exists a $\gamma_k \in \Gamma_N$ such that $\|L_{m,t+1}(\gamma) - L_{m,t+1}(\gamma_k)\|_{2r} \leq B|\gamma - \gamma_k|^{\lambda} \leq BG^{\lambda}N^{-\lambda/d} = \delta$. 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 $\sqrt{n}\bar{L}_{m,n}(\cdot)$ is stochastically equicontinuous. Together with finite dimensional convergence this implies $\sqrt{n}\bar{L}_{m,n}(\cdot) \Rightarrow Z_{m,n}(\cdot)$, with $Z_{m,n}(\cdot)$ a Gaussian process with covariance kernel $\Sigma_{m,n}(\cdot,\cdot)$.

Note that $\sigma^2_{m,n}(\cdot) \xrightarrow{a.s} \sigma^2_m(\cdot)$ uniformly over Γ under Assumption 4. That $v(t_{m,n}) \xrightarrow{d} v(t_{m,n})$ 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 $v(\tilde{t}_{m,n})$ being absolutely continuous on $(0, \infty)$. The absolute continuity of the distribution function of $v(\tilde{t}_{m,n})$ follows from $Z_{m,n}(\cdot)$ having a nondegenerate covariance kernel, and thus $\tilde{t}_{m,n}$ 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_a is established as follows.

Under the assumptions of Theorem 1 it follows that $\bar{L}_{m,n}(\gamma) \xrightarrow{a.s} E\left[L_{m,t+1}(\gamma)\right] \equiv \Delta(\gamma)$, uniformly over Γ.

Now notice that, for any $\gamma \in \Gamma$, $\left|E[L_{m,t+1}(\gamma^\dagger)]\right| - \left|E[L_{m,t+1}(\gamma^\dagger) - L_{m,t+1}(\gamma)]\right| \leq \left|E[L_{m,t+1}(\gamma)]\right|$ by the Triangle Inequality. Furthermore, from Jensen's Inequality, Hölder's Inequality and under Assumption 3, it follows that

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

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

Additionally, under Assumption 4 it follows that $\hat{\sigma}^2_{m,n}(\gamma) \xrightarrow{a.s.} \sigma^2_m(\gamma)$ uniformly over $\gamma \in \Gamma$, and $\inf_{\gamma\in\Gamma} \sigma^2_m(\gamma) > 0$, such that there exists a $\Delta'' > 0$ so that $n^{-1/2}|t_{m,n}(\gamma)| \xrightarrow{a.s.} n^{-1/2}\frac{|\bar{L}_{m,n}(\gamma)|}{\sigma_m(\gamma)} > \Delta''$, a.s., uniformly over $\Gamma^\dagger$. By application of the Continuous Mapping Theorem it follows that, a.s., $n^{-1}v(t^2_{m,n}) > 0$. Hence, $P[v(t^2_{m,n}) > c] \to 1$, for any constant $c \in \mathbb{R}$. □

A.3 Proof of Theorem 3

That $\sqrt{n}\bar{L}^*_{m,n}(\cdot) \Rightarrow Z_{m,n}(\cdot)$ almost surely follows from Theorem 1 in Bühlmann (1995). Assumption A1, A2, and A3 in Bühlmann (1995) are satisfied under Assumptions 1, 5, and 2, 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 $\sigma^2_{m,n}(\cdot) \xrightarrow{a.s} \sigma^2_m(\cdot)$ uniformly over Γ under Assumption 4, such that $t^*_{m,n}(\cdot) \Rightarrow \tilde{t}_{m,n}(\cdot)$ almost surely under the Continuous Mapping Theorem.

That $v(t^*_{m,n}) \xrightarrow{d} v(t_{m,n})$ 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 $\sqrt{n}\bar{L}^*_{m,n}(\cdot) \Rightarrow Z_{m,n}(\cdot)$ almost surely. The result follows.

□

A.4 Proof of Lemma 1

Part 1: We show the result for $\operatorname{ave} t^2_{m,n}$. The result for the other tests follows from similar steps.

The weak convergence of $t^2_{m,n}$ 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 $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} P\left(\sup_{|\gamma-\gamma'|<\delta} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma') \right| > \varepsilon \right) = 0,$$

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,

$$\begin{aligned}
&\left| \int_\Gamma t^2_{m,n}(\gamma) dJ(\gamma) - \sum_{i=1}^{K_n} t^2_{m,n}(\gamma_{n,i}) \int_{\Gamma_n^i} dJ(\gamma) \right| \\
&\quad \le \sum_{i=1}^{K_n} \int_{\Gamma_n^i} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma_{n,i}) \right| dJ(\gamma) \\
&\quad \le \sum_{i=1}^{K_n} \sup_{\gamma \in \Gamma_n^i} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma_{n,i}) \right| \int_{\Gamma_n^i} dJ(\gamma) \\
&\quad \le \sup_{|\gamma-\gamma'|<\delta_n} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma_{n,i}) \right| \sum_{i=1}^{K_n} \int_{\Gamma_n^i} dJ(\gamma) \\
&\quad = \sup_{|\gamma-\gamma'|<\delta_n} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma_{n,i}) \right|,
\end{aligned}$$

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

$$\begin{aligned}
&\limsup_{n\to\infty} P\left(\left| \int_\Gamma t^2_{m,n}(\gamma) dJ(\gamma) - \sum_{i=1}^{K_n} t^2_{m,n}(\gamma_{n,i}) \int_{\Gamma_n^i} dJ(\gamma) \right| > \varepsilon' \right) \\
&\quad \le \limsup_{n\to\infty} P\left(\sup_{|\gamma-\gamma'|<\delta_n} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma_{n,i}) \right| > \varepsilon \right) \\
&\quad \le \limsup_{n\to\infty} P\left(\sup_{|\gamma-\gamma'|<\delta} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma_{n,i}) \right| > \varepsilon \right) < \varepsilon,
\end{aligned}$$

where the last display follows from the stochastic equicontinuity of $t^2_{m,n}(\gamma)$. Because δ is arbitrary, the result follows.

Part 2: We show the result for $\operatorname{ave} t^2_{m,n}$. The result for $\sup t^2_{m,n}$ follows from similar steps.

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

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

Notice that

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

where the last line follows from Theorem 1.

Furthermore, we can show that $C_n = O(S_n^{-1/2+\eta})$, for some $\eta \in (0, 1/2)$, with probability approaching one 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 $\Gamma_i^n \subset \bar{\Gamma}$

$$\mathbb{1}\left(\gamma \in \Gamma_i^n\right) = \prod_{i=1}^{d} \mathbb{1}\left(\gamma_i \le \overline{\gamma_i^n}\right) \prod_{i=1}^{d} \left(1 - \mathbb{1}\left(\gamma_i \le \underline{\gamma_i^n}\right)\right), \tag{13}$$

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 (13) are type I(b) functions in the definition of Andrews (1994), and by Theorem 3 in Andrews (1994) so is the product (13). 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, for any $\varepsilon > 0$ we have $\lim_{n\to\infty} P\left(S_n^{1/2-\eta} C_n \le \varepsilon\right) = 1$ under the J-measure.

Furthermore, notice that

$$
\begin{aligned}
|B_n| &\leq \frac{1}{S_n} \sum_{j=1}^{S_n} \sum_{i=1}^{K_n} \left| t^2_{m,n}(\gamma^{(j)}) - t^2_{m,n}(\gamma_{n,i}) \right| \mathbb{1}\left(\gamma^{(j)} \in \Gamma^n_i\right) \\
&\leq 2^d \frac{1}{S_n} \sum_{j=1}^{S_n} \sup_{|\gamma^{(j)} - \gamma'| < \delta_n} \left| t^2_{m,n}(\gamma^{(j)}) - t^2_{m,n}(\gamma') \right| \\
&\leq 2^d \sup_{|\gamma - \gamma'| < \delta_n} \left| t^2_{m,n}(\gamma) - t^2_{m,n}(\gamma') \right| = o_p(1),
\end{aligned}
$$

by the stochastic equicontinuity of $t^2_{m,n}(\gamma)$ and where 2^d equals the maximum number of vertices shared amongst hyperrectangles in a hyperrectangular grid.

Hence, $|\widetilde{\text{ave}\, t^2_{m,n}} - \widehat{\text{ave}\, t^2_{m,n}}| = o_p(1)$, with probability approaching one under the J-measure if $K_n = o(S_n^{1/2-\eta})$. The result follows, since we are free to choose the rate at which $K_n \to \infty$ as $n \to \infty$.

□

Table 1: Small sample rejection rates of equal expected utility tests on naive and minimum-variance portfolio strategies ($\Gamma = [1, 10]$).

	2-sided tests				1-sided tests	
K_n	Wald	Bonferonni	ave-t^2	sup-t^2	Bonferonni	sup-t
Panel A: Size properties						
				$n = 120$		
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
				$n = 600$		
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 B: 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
				$n = 600$		
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 p-values of the one-sided and two-sided tests as well as the benchmark tests. The data is generated according to Equation (7), and the naive and portfolio strategies are given in Equation (6). The minimum-variance portfolio weights are estimated using a moving window of $m = 120$ months. The out-of-sample period consists of $n = 120$, and 600 months. We consider discrete grids of $\Gamma = [1, 10]$ with $K_n = 1, 10, 50, 100$, and 250 equally spaced grid points. Results for the multivariate Wald test with $K_n = 250$ are not shown for $n = 120$, due to singularity of the covariance matrix.

Table 2: Small sample rejection rates of quantile forecast tests, for differences between multivariate GARCH-DCC and Riskmetrics models (Γ is unit simplex).

	2-sided tests				1-sided tests	
S_n	Wald	Bonferonni	ave-t^2	sup-t^2	Bonferonni	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
1,000	-	0.01	0.03	0.05	0.00	0.01
				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
1,000	1.00	0.00	0.02	0.01	0.00	0.01
Panel B: 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
1,000	-	0.01	0.11	0.10	0.02	0.13
				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
1,000	1.00	0.09	0.43	0.49	0.15	0.64

Note: This table presents rejection rates of the one-sided and two-sided tests as well as the benchmark tests in our size and power experiments. The quantile forecasts for the portfolio returns from the GARCH-DCC and multivariate Riskmetrics models are defined in Equations (9) and Equation (10), respectively. The data is generated as in Equation (8) with $N = 30$. We let Γ be the set of all portfolio weight vectors with positive portfolio weights summing to one. We test at 31 fixed portfolio weight vector being the equally-weighted portfolio vector and the 30 basis vectors, as well as $S_n - 31$ weight vectors drawn uniformly from Γ, with $S_n = 50, 100, 250, 500$, and 1,000.

Table 3: Small sample rejection rates of Murphy diagram tests, for quantile forecast differences between GARCH and Riskmetrics models ($\Gamma = [-20, 0]$).

	2-sided tests			1-sided tests	
K_n	tick-loss	ave-t^2	sup-t^2	tick-loss	sup-t
Panel A: Size properties, GARCH vs Riskmetrics					
			n = 500		
10	0.07	0.21	0.21	0.04	0.03
50	-	0.04	0.04	-	0.03
100	-	0.03	0.03	-	0.02
250	-	0.02	0.02	-	0.01
			n = 2,000		
10	0.06	0.07	0.07	0.04	0.03
50	-	0.04	0.05	-	0.03
100	-	0.05	0.05	-	0.04
250	-	0.04	0.04	-	0.04
Panel B: Size properties, GARCH vs 250-day moving sample quantile					
			n = 500		
10	0.21	0.48	0.48	0.05	0.04
50	-	0.14	0.12	-	0.07
100	-	0.14	0.10	-	0.09
250	-	0.15	0.09	-	0.08
			n = 2,000		
10	0.16	0.32	0.32	0.06	0.05
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, GARCH vs Riskmetrics					
			n = 500		
10	0.11	0.05	0.05	0.19	0.05
50	-	0.04	0.04	-	0.06
100	-	0.04	0.04	-	0.06
250	-	0.03	0.03	-	0.04
			n = 2,000		
10	0.37	0.07	0.07	0.49	0.09
50	-	0.10	0.08	-	0.14
100	-	0.10	0.08	-	0.13
250	-	0.12	0.10	-	0.15
Panel D: Power properties, GARCH vs 250-day moving sample quantile					
			n = 500		
10	0.32	0.13	0.13	0.45	0.06
50	-	0.26	0.27	-	0.36
100	-	0.33	0.30	-	0.39
250	-	0.30	0.27	-	0.36
			n = 2,000		
10	0.90	0.18	0.18	0.95	0.17
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 rejection rates of the one-sided and two-sided tests as well as the benchmark tests using the tick-loss function in our size and power experiments. The quantile forecasts from the GARCH and Riskmetrics models are given in Equations (9) and Equation (10), respectively, with $N = 1$. We consider out-of-sample period lengths n = 500, and 2,000, and discrete grids of $\Gamma = [-20, 0]$ with $K_n = 10, 50, 100$, and 250 equally spaced grid points. The tick-loss-based test does not depend on the choice of K_n.

Table 4: p-values of equal expected utility tests on naive and minimum-variance portfolio strategies, for several risk aversion parameter intervals Γ.

	2-sided tests				1-sided tests	
K_n	Wald	Bonferonni	ave-t^2	sup-t^2	Bonferonni	sup-t
Panel A: $\Gamma = [1, 10]$						
1	0.03	0.03	0.04	0.04	0.02	0.01
10	0.00	0.04	0.13	0.01	0.02	0.00
50	0.00	0.22	0.17	0.01	0.11	0.00
100	0.00	0.45	0.16	0.01	0.22	0.00
250	0.26	1.00	0.16	0.02	0.56	0.01
Panel B: $\Gamma = [1, 5]$						
1	0.03	0.03	0.04	0.04	0.02	0.02
10	0.00	0.04	0.07	0.01	0.02	0.00
50	0.00	0.22	0.07	0.01	0.11	0.00
100	0.00	0.45	0.06	0.01	0.22	0.00
250	0.00	1.00	0.04	0.01	0.56	0.01
Panel B: $\Gamma = [5, 10]$						
1	0.87	0.87	0.87	0.87	0.56	0.57
10	0.00	1.00	0.53	0.38	1.00	0.22
50	0.00	1.00	0.53	0.36	1.00	0.17
100	0.00	1.00	0.62	0.41	1.00	0.22
250	0.02	1.00	0.58	0.38	1.00	0.20

Note: This table presents rejection rates of the one-sided and two-sided tests as well as the benchmark tests in our size and power experiments. The naive and portfolio strategies are given in Equation (6). The data consists of monthly returns of 30 industry portfolios and runs from September 1926 to December 2017. We consider discrete (sub-)grids of $\Gamma = [1, 10]$ with $K_n = 1, 10, 50, 100,$ and 250 equally spaced grid points.

Table 5: p-values of quantile forecast tests, for differences between multivariate GARCH-DCC and Riskmetrics models (Γ is unit simplex).

	2-sided tests				1-sided tests	
S_n	Wald	Bonferonni	ave-t^2	sup-t^2	Bonferonni	sup-t
Panel A: Full sample						
1	0.78	0.78	0.79	0.79	0.39	0.39
31	0.99	1.00	0.99	0.95	1.00	0.62
100	-	1.00	0.94	0.95	1.00	0.64
250	-	1.00	0.81	0.93	1.00	0.59
500	-	1.00	0.79	0.90	1.00	0.54
1,000	-	1.00	0.74	0.90	1.00	0.58
Panel B: First half of sample						
1	0.30	0.30	0.30	0.30	0.15	0.15
31	0.01	0.05	0.04	0.04	0.02	0.02
100	-	0.16	0.22	0.05	0.08	0.03
250	-	0.40	0.23	0.05	0.20	0.03
500	-	0.80	0.24	0.06	0.40	0.04
1,000	-	1.00	0.23	0.05	0.80	0.03
Panel C: Second half of sample						
1	0.32	0.32	0.34	0.34	0.84	0.85
31	0.33	0.35	0.22	0.17	1.00	0.87
100	-	1.00	0.28	0.16	1.00	0.90
250	-	1.00	0.28	0.18	1.00	0.91
500	-	1.00	0.30	0.18	1.00	0.90
1,000	-	1.00	0.30	0.17	1.00	0.91

This table presents p-values of the one-sided and two-sided tests as well as the benchmark tests. The quantile forecasts for the portfolio returns from the GARCH-DCC and multivariate Riskmetrics models are defined in Equations (9) and Equation (10), respectively. The data consists of twenty years of daily returns of a thirty industry portfolios and runs from January 2, 1998 to December 29, 2017. We let Γ be the set of all portfolio weight vectors with positive portfolio weights summing to one. We test at 31 fixed portfolio weight vector being the equally-weighted portfolio vector and the 30 basis vectors, as well as $S_n - 31$ weight vectors drawn uniformly from Γ, with $S_n = 100, 250, 500,$ and 1,000.

Table 6: p-values and cross-sectional ratio of 5% rejections of quantile forecast tests, for 30 Industry portfolios ($\Gamma = [-20, 0]$).

	2-sided tests			1-sided tests	
K_n	tick-loss	ave-t^2	sup-t^2	tick-loss	sup-t
Panel A: p-value for Transportation portfolio, Riskmetrics vs. GARCH					
10	0.77	0.38	0.38	0.38	0.15
50	-	0.02	0.00	-	0.01
100	-	0.07	0.01	-	0.04
250	-	0.08	0.02	-	0.04
500	-	0.07	0.01	-	0.03
1000	-	0.08	0.02	-	0.04
1500	-	0.06	0.02	-	0.03
Panel B: p-value for Transportation portfolio, 250-day moving sample quantile vs. GARCH					
10	0.00	0.00	0.00	1.00	0.76
50	-	0.00	0.00	-	0.96
100	-	0.00	0.00	-	0.98
250	-	0.00	0.00	-	0.99
500	-	0.00	0.00	-	0.98
1000	-	0.00	0.00	-	0.99
1500	-	0.00	0.00	-	0.99
Panel C: Rejection ratio over 30 industry portfolios, Riskmetrics vs. GARCH					
10	0.00	0.23	0.23	-	0.27
50	-	0.07	0.07	-	0.13
100	-	0.30	0.17	-	0.23
250	-	0.20	0.07	-	0.13
500	-	0.23	0.10	-	0.20
1000	-	0.23	0.10	-	0.10
1500	-	0.23	0.10	-	0.10
Panel D: Rejection ratio over 30 industry portfolios, 250-day moving sample quantile vs. GARCH					
10	1.00	1.00	1.00	-	0.00
50	-	1.00	1.00	-	0.00
100	-	1.00	1.00	-	0.00
250	-	1.00	1.00	-	0.00
500	-	1.00	1.00	-	0.00
1000	-	1.00	1.00	-	0.00
1500	-	1.00	1.00	-	0.00

Note: This table presents the ratio of rejection at the 5% significance level of the one-sided and two-sided tests as well as the benchmark tests using the tick-loss function, applied to each of the 30 Industry portfolio returns series and the equally weighted portfolio of the Industry portfolios. The quantile forecasts from the GARCH and Riskmetrics models are given in Equations (9) and Equation (10), respectively, with $h = 1$. The moving sample quantile is defined as the 5% percentile over the previous 75 observations. The data consists of twenty years of daily returns of a food industry portfolio and runs from January 2, 1998 to December 29, 2017. We consider discrete grids of $\Gamma = [-20, 0]$ with $K_n = 10, 50, 100, 250, 500$, 1,000, and 1,500 equally spaced grid points. The GARCH model is estimated over a moving window of 1,000 observations. The tick-loss-based test does not depend on choice K_n.

Figure 1: Utility differential of naive and minimum-variance portfolio strategies.

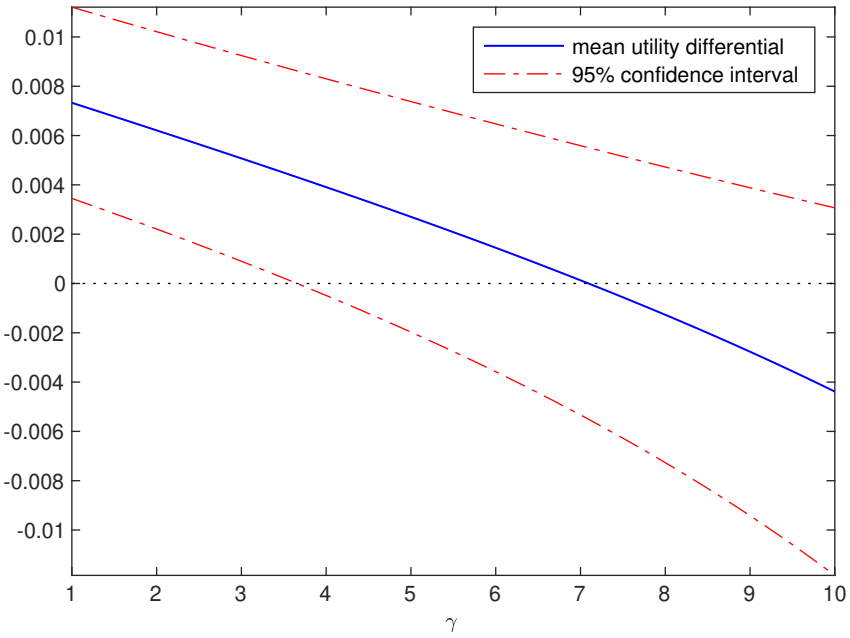


Figure 2: Tick-loss differential for tail quantile forecasts generated by the GARCH-DCC and multivariate RiskMetrics models.

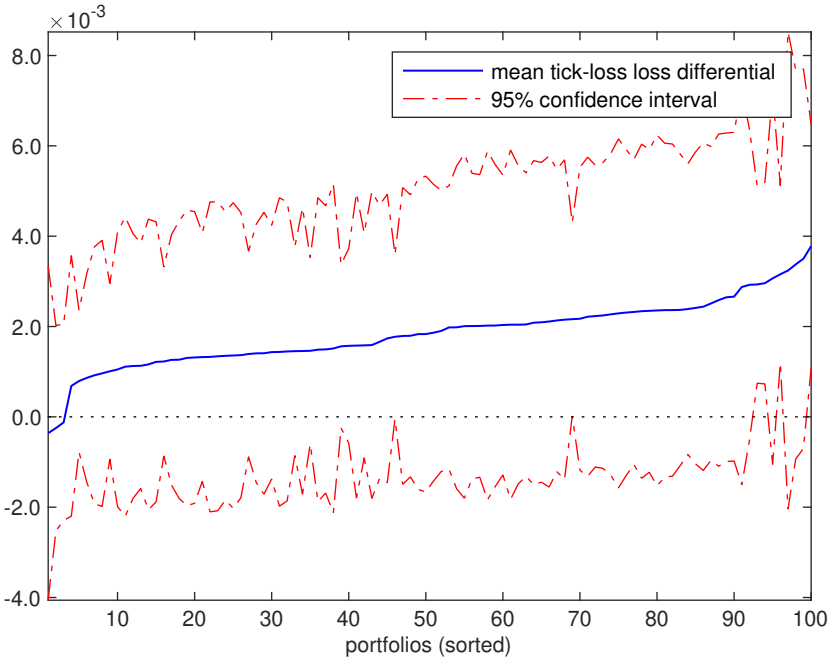
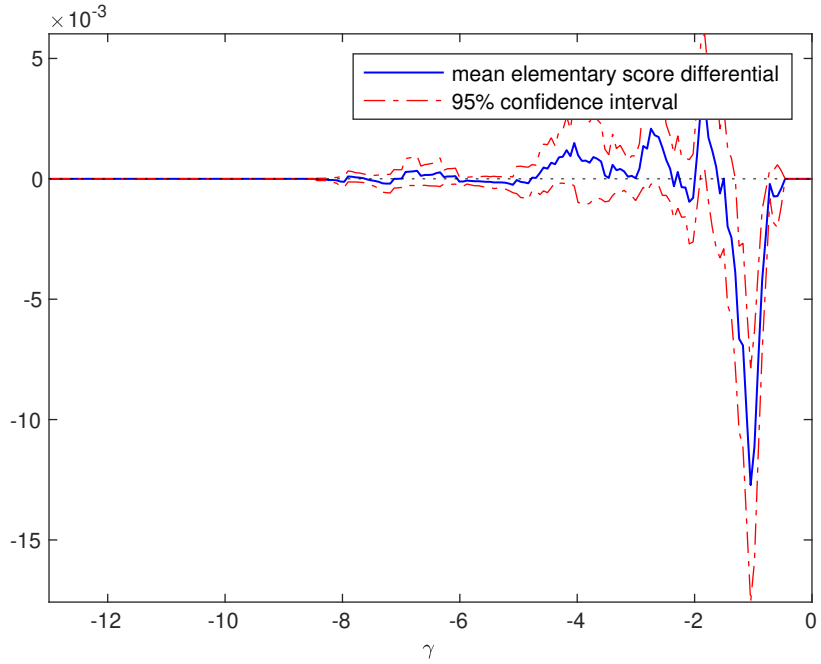
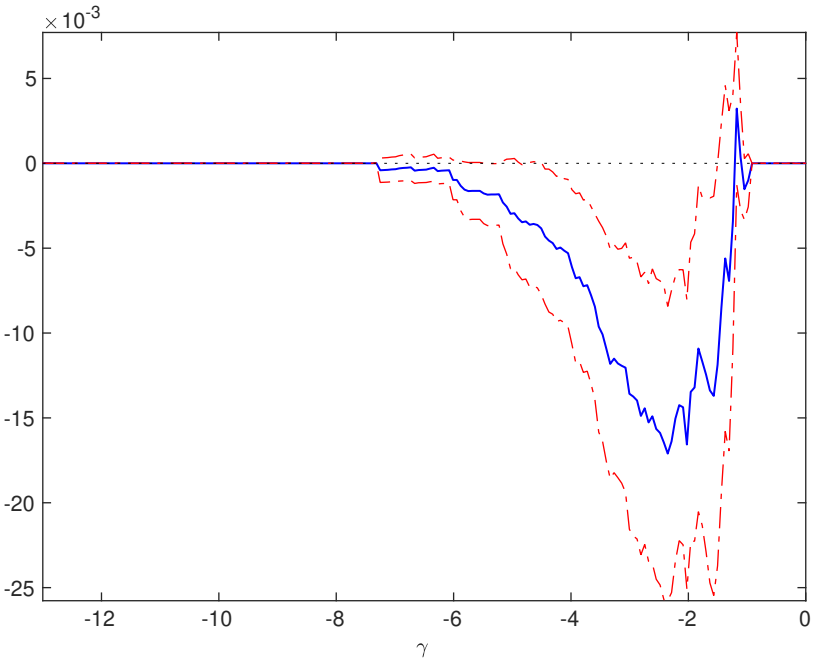


Figure 3: Murphy diagrams; elementary loss differential between quantile forecasts for the Transportation portfolio



(a) GARCH vs. RiskMetrics



(b) GARCH vs. moving sample quantile