

Comparing Density Forecasts via Weighted Likelihood Ratio Tests. Asymptotic and Bootstrap Methods.

Raffaella Giacomini*

University of California, San Diego

This version: January 2002

Abstract

This paper proposes and analyzes tests that can be used to compare the accuracy of alternative density forecasts of a variable. The tests are also valid in the broader context of model selection based on out-of-sample predictive ability. We restrict attention to the case of density forecasts derived from parametric models that are non-nested or overlapping, with known or estimated parameters. For simplicity, we consider univariate density forecasts, but the results can be easily extended to the multivariate case. We propose asymptotic and bootstrap weighted likelihood ratio tests that focus power on different regions of the unconditional distribution of the variable, as a way to incorporate loss functions into the evaluation procedure. The loss functions proposed are defined over the distance between the density forecast and the true density, which represents a departure from the treatment of loss functions in the point forecasting literature. We show how the likelihood ratio test for non-nested hypotheses proposed by Vuong (1989) can be obtained in our framework. A simulation exercise analyzes size and power properties of this last test in the context of density forecasting. In an application using S&P500 daily returns, all the tests indicate that density forecasts from a GARCH-type model with Student's t disturbances outperform density forecasts from a GARCH model with Generalized Error Distribution (GED) disturbances. The same conclusion holds when the first density forecast is from a GARCH with skewed- t disturbances. The weighted likelihood ratio tests suggest that t -GARCH forecasts outperform GED -GARCH forecasts in 'normal' days, while skewed t -GARCH forecasts outperform GED -GARCH forecasts in both 'normal' days and in days when the returns are relatively high.

*I am deeply indebted to Clive W. J. Granger for introducing me to this line of research and for many interesting discussions. I would also like to thank Carlos Capistran, Graham Elliott, Andrew Patton, Kevin Sheppard and Allan Timmermann for valuable comments. Address: Department of Economics 0508, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093. E-mail: rgiacomini@ucsd.edu.

1 Introduction

This paper proposes and analyzes tests for comparing two alternative sequences of density forecasts of the same variable.

A density forecast is an estimate of the future probability distribution of a random variable, conditional on the information available at the time the forecast is made. It thus represents a complete characterization of the uncertainty associated with the forecast, as opposed to a point forecast, which provides no information about the uncertainty of the prediction. Density forecasting is receiving increasing attention in both macroeconomics and finance (see Tay and Wallis, 2000 for a survey). A famous example of density forecasting in macroeconomics is the ‘fan-chart’ of inflation and GDP published by the Bank of England and by the Sveriges Riksbank in Sweden in their quarterly Inflation Reports (for other examples of density forecasting in macroeconomics, see also Diebold, Tay and Wallis, 1999 and Clements and Smith, 2000). In finance, where the wide availability of data and the increasing computational power make it possible to produce more accurate estimates of densities, the examples are numerous. Leading cases are in risk management, where forecasts of portfolio distributions are issued with the purpose of tracking measures of portfolio risk such as the Value-at-Risk (see, e.g., Duffie and Pan, 1996) or the Expected Shortfall (see, e.g., Artzner *et al.*, 1997). Another example is the literature that focuses on extracting density forecasts from option price data (see, e.g. Soderlind and Svensson, 1997). The vast literature on forecasting volatility with GARCH-type models (see Bollerslev, Engle and Nelson, 1994) and its extensions to forecasting higher moments of the conditional distribution (see Hansen, 1994) can also be seen as precursors to density forecasting. The use of sophisticated distributions for the standardized residuals of a GARCH model and the modeling of time dependence in higher moments is in many cases just an attempt to capture relevant features of the data to better approximate the true distribution of the variable. With density forecasting becoming more and more widespread in applied econometrics, it is necessary to develop reliable techniques to evaluate the forecasts’ performance. A popular method for evaluating a sequence of conditional density forecasts was proposed by Diebold, Gunther and Tay (1998). These authors suggested evaluating a sequence of density forecasts by considering the probability integral transforms (z_t) of the realized data with respect to the forecast densities. If the density forecasts coincide with the true conditional densities, the sequence $\{z_t\}$ is i.i.d. $U(0,1)$ ¹. While Diebold *et*

¹The basic result showing the $U(0,1)$ distribution of the probability integral transform when the density is correctly specified is attributed to Fisher (1932). The same approach adopted by Diebold *et al.* (1998) was also proposed by Dawid (1984) and Kling and Bessler (1989), among others.

al. (1998) adopted mainly qualitative tools for testing the i.i.d. $U(0,1)$ behavior of the transformed data, formal tests of the same hypothesis have been recently suggested by Berkowitz (2000), Hong (2000) and Hong and White (2000).

Even with rigorous testing procedures of the i.i.d. $U(0,1)$ hypothesis available, this approach has some drawbacks. First, the key result that the sequence $\{z_t\}$ is i.i.d. $U(0,1)$ relies on the assumption of no parameter estimation uncertainty in the densities, and it is not clear that the validity of the procedure would hold when estimation uncertainty is taken into account. Second, it is important to emphasize that the method is only valid in absolute terms, that is, to evaluate the ‘goodness’ of a particular sequence of density forecasts, relative to the data-generating process. In practice, it is likely that any econometric model used to produce the sequence of density forecasts is misspecified. In this situation a more relevant question is how to decide which one of two (or more) given alternative density forecasts is preferable. The issue of comparative evaluation of density forecasts has not yet been explored in the theoretical literature, but, in spite of this, there are a few examples of empirical research that attempt to compare density forecasts. The Diebold *et al.* (1998) technique, in particular, has inspired a number of applications where alternative density forecasts are evaluated by constructing their corresponding sequences of probability integral transforms and comparing their relative ‘closeness’ to the uniform distribution, see, e.g., Clements and Smith (2000), Weigend and Shi (2000) and Bauwens, Giot, Grammig and Veredas (2000). This closeness is in most situations assessed only through visual inspection, and no measure of distance or formal testing is utilized. This paper attempts to fill the gap in the literature and propose formal tests that can be utilized to rank alternative density forecasts. The necessity of reliable techniques for comparing parametric density forecasts from alternative model specifications is also clear in the context of copula modeling, that is gaining interest in financial econometrics (see Patton, 2001a, b). A copula is a complete characterization of the dependence between variables and the use of copulas in econometrics represents a flexible way to model multivariate distributions, by specifying distinct parametric models for the marginal distributions and for the copula. The out-of-sample evaluation of copula models is complicated by the fact that the multivariate distribution arising from the use of different copula specifications are often non-nested, and the available techniques do not easily lend themselves to this situation. The multivariate version of all the tests proposed in this paper can be naturally utilized for the evaluation and selection of copula models.

In the paper, we restrict attention to a specific forecasting situation. The relevant environment is one in which two alternative parametric conditional models are used to generate density forecasts

for the variable of interest. The models are non-nested and the parameters of the densities are either known or estimated. We emphasize that even though the paper focuses on a density forecasting environment, the techniques proposed can be used in the more general context of model selection. In this case, two competing models will be analyzed in terms of their *ex-post* predictive performance, in an out-of-sample evaluation exercise that compares forecasts of the entire density, rather than simple point forecasts implied by the two models. As such, our tests can be used in conjunction with Diebold-Mariano (1995) type of tests, that compare models according to their (point) forecasting accuracy or in terms of the relative loss implied by some economically meaningful criterion.

One of the contributions of the paper is a first step towards thinking about loss functions and forecast evaluation in a density forecasting framework. When the object of interest is a density forecast, rather than a point forecast, the standard framework of loss functions defined over the forecast errors (see, e.g., Christoffersen and Diebold, 1997) is no longer valid. We consider instead loss functions defined over the distance between the density forecast and the true density, and show how this set-up leads to the development of ‘weighted likelihood ratio tests’ for comparing density forecasts. We point out that the standard likelihood ratio test for comparison of non-nested models proposed by Vuong (1989) can be justified in our framework by a particular functional form of the loss function. Introducing a general class of loss functions allows us to formulate tests that compare the performance of density forecasts in different regions of the unconditional distributions of the variable, allowing for example to distinguish predictive ability in ‘normal’ days from that in ‘extreme’ days..

The paper is organized as follows. Section 2 introduces the notation and the assumptions utilized in the paper. In Section 3, we examine a way of introducing loss functions into the evaluation of density forecasts. The class of loss functions proposed leads to the development of asymptotic and bootstrap weighted likelihood ratio tests, discussed in Section 4. Section 5 shows how the standard likelihood ratio test for non-nested hypotheses proposed by Vuong (1989) is a special case of the asymptotic test examined in the previous section, with a simpler expression for the asymptotic variance when the parameter estimates are QMLEs. A bootstrap and a bootstrap-*t* likelihood ratio tests are also considered, and the small sample properties of the tests are analyzed in a Monte Carlo simulation in Section 6. All the tests proposed in the paper are used in Section 7 to compare density forecasts for the S&P 500 daily returns obtained from GARCH models with different distributional assumptions. Finally, Section 8 concludes.

2 Description of environment

2.1 Notation

For simplicity we will restrict attention to the univariate case. The extension to multivariate is relatively straightforward.

The density forecasts are based on the two alternative conditional models $F_\theta \equiv \{f(x_{t+1}|\Omega_t; \theta); \theta \in \Theta\}$ and $G_\gamma \equiv \{g(x_{t+1}|\Omega_t; \gamma); \gamma \in \Gamma\}$. The parameter spaces Θ and Γ are respectively k_1 and k_2 dimensional. The forecasts are conditional on the information set $\Omega_t = \{x_{t-j}, z_{t+1-j}; j \geq 0\}$, containing the past history of the variable of interest X_t and possibly the history of exogenous variables denoted jointly as Z_{t+1} . The available sample of size T is divided in two parts, with the first R data used for estimation and the last n for out-of-sample evaluation. The first forecasts are formed using data from 1 to R , the second using data 1 to $R + 1$ and so forth. The last forecasts are produced by estimating the models on data from 1 to $R + n - 1 \equiv T - 1$. Let $\hat{\theta}_t$ and $\hat{\gamma}_t$ denote the estimators based on data from 1 to t . The procedure will generate two sequences of n density forecasts

$$\{f(x_{t+1}|\Omega_t; \hat{\theta}_t)\}_{t=R}^{T-1} \text{ and } \{g(x_{t+1}|\Omega_t; \hat{\gamma}_t)\}_{t=R}^{T-1} \quad (1)$$

for the variables X_{R+1}, \dots, X_T . We let $h(x_{t+1}|\Omega_t)$ indicate the true conditional density of the variable X_{t+1} and let θ^* and γ^* denote the probability limits respectively of $\hat{\theta}_t$ and $\hat{\gamma}_t$. Finally, we define a weight function $w : \mathbb{R} \rightarrow [0, 1]$, whose meaning and use will be made more precise in Section 3.

2.2 Assumptions

The first set of assumptions is related to the regularity conditions utilized by West (1996). The assumptions only stated in terms of f and θ implicitly hold for g and γ . The symbol ∇_θ^k will denote the k -th derivative operator with respect to θ .

Assumption 1. Let N be an open neighborhood of θ^* : (a) $f(x_{t+1}|\Omega_t, \cdot)$ is continuously differentiable of order 2 on N and the weight function $w(\cdot)$ is twice continuously differentiable on \mathbb{R} .

(b) There exists a constant $D < \infty$ such that for all t , $\sup_{\theta \in N} |\nabla_\theta^2 \log f(X_{t+1}|\Omega_t, \theta)| < m_t$ for a measurable m_t that satisfies $Em_t < D$.

(c) The estimate $\hat{\theta}_t$ satisfies $\hat{\theta}_t - \theta^* = B^f(t)A^f(t)$, where $B^f(t)$ is $k_1 \times q$ and $A^f(t)$ is $q \times 1$, with $B^f(t) \xrightarrow{a.s.} B^f$, B^f matrix of rank k_1 and $A^f(t) = t^{-1} \sum_{s=1}^t a_s^f(\theta^*)$ for a $q \times 1$ orthogonality condition $a_s^f(\theta^*)$ such that $Ea_s^f(\theta^*) = 0$.

(d) For some $d, d', d'' > 1$, $\sup_t E|\log f(X_{t+1}|\Omega_t, \theta^*)|^{4d}$, $\sup_t E|\nabla_{\theta_i} \log f(X_{t+1}|\Omega_t, \theta^*)|^{4d'}$, $\sup_t E|a_t^f|^{4d''} < \infty$, for all i , where ∇_{θ_i} is the i -th component of the gradient.

(e) $\{X_t\}$ is strong mixing, with mixing coefficients of size $-3d/(d-1)$.

(f) $[w(X_{t+1}) \log f(X_{t+1}|\Omega_t, \theta^*), (w(X_{t+1}) \nabla_{\theta} \log f(X_{t+1}|\Omega_t, \theta^*))', a_t^f]'$ is covariance stationary.

Assumption 1 imposes conditions on the density models, the weight function, the estimation procedure and the data-generating process of the random variable X_t . The restrictions are fairly standard, and allow for application of the results to a wide range of situations that arise in practice. In particular, we demand the use of smooth density functions for the forecast models, but this requirement could be relaxed along the lines of McCracken (2000). We also require existence of at least four moments of the log-likelihoods and the scores. This requirement, in general, depends on both the density models and the true density, and its plausibility should thus be established case by case. The parameters of the models can be estimated by a variety of linear and nonlinear techniques, including Maximum Likelihood, OLS and GMM. The restrictions on memory and heterogeneity of the data-generating process still allow for conditional heterogeneity and serial dependence.²

Assumption 2. (a) The conditional models F_{θ} and G_{γ} are non-nested: $F_{\theta} \not\subseteq G_{\gamma}$ and $G_{\gamma} \not\subseteq F_{\theta}$.
(b) $f(\cdot|\cdot; \theta^*) \neq g(\cdot|\cdot; \gamma^*)$.

Part (a) indicates that the models can be either strictly non-nested ($F_{\theta} \cap G_{\gamma} = \emptyset$) or overlapping ($F_{\theta} \cap G_{\gamma} \neq \emptyset$ but $F_{\theta} \not\subseteq G_{\gamma}$ and $G_{\gamma} \not\subseteq F_{\theta}$). Strictly non-nested densities are, e.g., the normal and the lognormal, the Student's t and the Generalized Error Distribution (*GED*) with finite degrees of freedom parameters. Alternatively, non-nestedness can be achieved when both f and g belong to the same family of distributions, but the models specify non-nested expressions for, say, the conditional mean or variance. An example is the case of two different non-linear specifications that cannot be obtained from each other, or of models that use different exogenous variables. Overlapping arises when the two models are not nested but still possess some common elements, as in the case of conditional moment equations that depend on some common explanatory variable or of two families of distributions that both nest the normal. For a more complete discussion, see Vuong (1989).

Part (b) is relevant in the case of overlapping models. It requires the density forecasts to be distinct only when evaluated at the respective probability limits of the parameters. In practice, unless the probability limits of the parameter estimates are known *a priori*, one will have to pre-test for condition (b). In the example where f and g belong to the same family of distributions and specify conditional moment equations that depend on common and non-common variables, one should verify

²The assumption of covariance stationarity (Assumption 1-(f)) is mainly imposed for convenience in estimating the asymptotic variance matrix. This assumption could be relaxed, at the price of increased complexity (see Rivers and Vuong, 1999).

that at least one of the coefficients on the non-common variables is significantly different from zero³. In this case, it is guaranteed that the two density forecasts evaluated at the probability limits of the parameters are distinct.

Assumption 3. As $T \rightarrow \infty$, $R, n \rightarrow \infty$ and $\lim_{T \rightarrow \infty}(n/R) = \pi$, $0 \leq \pi < \infty$.

Assumption 3 allows the in-sample and the out-of-sample sizes to diverge at the same rate, or the in-sample size to grow faster than the out-of-sample. This assumption concerns the way the asymptotic distribution is achieved. In particular, letting the in-sample and the out-of-sample diverge at the same rate is a way to state that the asymptotic distribution of the test statistics will take into account the uncertainty due to estimated parameters. Imposing $\pi = 0$, on the other hand, is an artificial way to ensure that estimation uncertainty will not affect the asymptotic distribution.

3 Loss functions and density forecasting

Traditionally, the vast majority of the forecast evaluation literature has focused on assessing the statistical accuracy of a forecast. Recently, however, a few studies (e.g., Granger and Pesaran, 2000, Pesaran and Skouras, 2000) have advocated a closer link between the forecast evaluation and the underlying decision problem, proposing economically meaningful evaluation methods.

There is now a quite large literature on loss functions and evaluation of point forecasts (see, e.g., Christoffersen and Diebold, 1996 and 1997). In this section, we explore the possibility of incorporating loss functions into the evaluation of density forecasts, and argue that the standard framework of loss functions for forecast evaluation is not appropriate when the forecast is a density forecast. Our perspective differs from the treatment in Granger and Pesaran (2000) and Pesaran and Skouras (2000), who consider the decision problem of a user basing his decisions on a density forecast and propose evaluating the forecast according to its economic value for the user. The incorporation of loss functions into the forecasting problem has until now focused on the definition of classes of loss functions of the form $L(\hat{x}_{t,\tau}, x_{t+\tau})$, where $\hat{x}_{t,\tau}$ is a τ -step-ahead point forecast of $X_{t+\tau}$ and $x_{t+\tau}$ is the realization of the variable. In the vast majority of cases, the loss function is assumed to only depend on the forecast error, $e_{t,\tau} = \hat{x}_{t,\tau} - x_{t+\tau}$, as for quadratic loss or general asymmetric loss (see, e.g., Christoffersen and Diebold, 1996, 1997, Weiss, 1996). Weiss (1996) shows that, in this framework, the optimal predictor is some summary measure of the true conditional density of the variable $X_{t+\tau}$ (the mean for quadratic loss, the median for absolute error loss, etc.). This means that a user with, say, a

³In this case, the following test will not have exact size α . Instead, α will represent an upper bound on the actual size of the test.

quadratic loss function will only care about the accuracy of the mean prediction and will be indifferent among density forecasts that yield the same forecast for the conditional mean. In other words, if the user has a loss function of the form $L(\hat{x}_{t,\tau}, x_{t+\tau})$, it becomes unnecessary to issue a density forecast in the first place, and the forecaster should only concentrate on accurately forecasting the relevant summary measure of the true density. The discussion of loss functions relevant for density forecasting must thus involve a shift of focus.

Starting from the assumption that it is the whole density that one wants to predict, we consider loss functions that are defined over the distance between the density forecast⁴ $f(x_{t+1}|\Omega_t; \theta^*)$ and the true density $h(x_{t+1}|\Omega_t)$:

$$L(f, h) \equiv L(\rho(f, h)), \quad (2)$$

where $\rho \in \mathbb{R}^+ \cup \{0\}$ denotes any measure of divergence⁵ between functions. The loss function $L: \mathbb{R}^+ \cup \{0\} \rightarrow \mathbb{R}$ satisfies the following requirements:

1. $L(\cdot)$ is increasing in its argument
2. $L(0) = 0$.

The second requirement states that if the density forecast coincides with the true density (and thus $\rho(f, h) = 0$), any loss should be zero. This condition formalizes the conclusion reached by Diebold et al. (1998) and Granger and Pesaran (2000) that the true density is the only density forecast preferred by all users, regardless of loss functions.

A general discussion of plausible classes of loss functions and their properties is beyond the scope of the present paper. Instead, we focus our attention on a particular class of loss functions that has intuitive appeal and that will lead to the development of ‘weighted likelihood ratio’ tests. The starting point is the idea that a user might be especially interested in a density forecast that is accurate in predicting events that lay in a particular region of the unconditional distribution of the variable of interest. An example could be a user who only cares about predicting (loosely defined) tail events, as in the case when different investment strategies or policy implications would arise if the future realizations of the variable fall into the tails of the distribution. If the user is presented with two alternative density forecasts, he might then want to place greater emphasis on the performance

⁴Notice that, in the definition of the loss function, the density forecast f is evaluated at the probability limit of the parameter estimate θ^* .

⁵We use the concept of divergence, instead of distance, because a measure of divergence need not be symmetric, while a distance must be.

of the competing models in the tails of the distribution, and less emphasis on what happens in the center. Another situation that might be of interest is a focus on predicting events that fall near the unconditional mean of the variable, as a way to ignore the influence of possible outliers on predictive performance. Finally, one might want to separate the predictive performance of the models in the right and in the left tail of the distribution, as in the case, e.g., of forecasting models for risk management, where only the left tail is relevant.

In this paper, we propose a class of loss functions that can be used to address these kinds of problems. If $f(x_{t+1}|\Omega_t; \theta^*)$ is the density forecast and $h(x_{t+1}|\Omega_t)$ the true density, we consider a loss function defined as

$$L(f, h) \equiv E[w(X_{t+1}) \log h(X_{t+1}|\Omega_t) - w(X_{t+1}) \log f(X_{t+1}|\Omega_t; \theta^*)], \quad (3)$$

where $w(\cdot)$ is the weight function and the expectation is taken with respect to the true density of $(X_1, Z_1, \dots, X_{t+1}, Z_{t+1})$. Intuitively, the above loss function can be seen as a generalization of the Kullback-Leibler information criterion (KLIC), defined as $I(f, h) \equiv E[\log h - \log f]$. The KLIC is a special case of (3), obtained for a weight function identically equal to one. While the KLIC measures the divergence between density f and the true density h by their expected log difference, the loss function (3) is a measure of expected ‘weighted’ log difference, with a higher weight placed on the region of the unconditional distribution of the variable X_t that is of interest. For example, when the data have unconditional mean 0 and variance 1, one could consider the following weight functions.

- Center of distribution: $w_1(x) = \phi(x)$, ϕ standard normal pdf
- Tails of distribution: $w_2(x) = 1 - \phi(x)/\phi(0)$, ϕ standard normal pdf
- Right tail: $w_3(x) = \Phi(x)$, Φ standard normal cdf
- Left tail: $w_4(x) = 1 - \Phi(x)$, Φ standard normal cdf

Plots of the weight functions $w_1 - w_4$ are shown in Figure 1.

[FIGURE 1 HERE]

4 Weighted likelihood ratio tests

Although the framework described in the previous section could be utilized to discuss optimality of a particular density forecast, the focus of the paper is on comparison of density forecasts. In this

case, given two alternative density forecasts $f \in F_\theta$ and $g \in G_\gamma$, a user will choose the forecast that yields the lowest loss.

When the loss function is as defined in (3), the relative loss of forecasts f and g with respect to the true density h is given by

$$WLR^* \equiv L(f, h) - L(g, h) = E[w(X_{t+1})(\log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*))]. \quad (4)$$

A test for equal loss of models F_θ and G_γ is thus a test of the null hypothesis

$$H_0 : WLR^* = 0 \text{ against} \quad (5)$$

$$H_g : WLR^* > 0 \text{ or}$$

$$H_f : WLR^* < 0,$$

where the two alternative hypotheses respectively indicate that G_γ is better than F_θ or that F_θ is better than G_γ . We call a test of H_0 a ‘weighted likelihood ratio’ test. The expression for WLR^* depends on the unknown expectation $E[\cdot]$ and probability limits θ^* and γ^* . We propose estimating WLR^* by the out-of-sample analogue

$$WLR_n = n^{-1} \sum_{t=R}^{T-1} wd_{t+1}(\hat{\beta}_t), \quad (6)$$

where $wd_{t+1}(\hat{\beta}_t) \equiv w(x_{t+1})[\log g(x_{t+1}|\Omega_t; \hat{\gamma}_t) - \log f(x_{t+1}|\Omega_t; \hat{\theta}_t)]$, $\hat{\beta}_t \equiv (\hat{\gamma}'_t, \hat{\theta}'_t)'$ and $\{x_{t+1}\}_{t=R+1}^T$ are the realizations of the variable over the out-of-sample period.

4.1 Asymptotic weighted likelihood ratio test

An asymptotic test of hypothesis H_0 can be derived using the framework developed by West (1996). The test will rely on asymptotic normality of the test statistic and, for a general weight function $w(\cdot)$, the asymptotic variance will incorporate terms that reflect parameter estimation uncertainty. We introduce the following notation.

$$q_{t+1}(\beta) \equiv \begin{pmatrix} w(x_{t+1}) \log g(x_{t+1}|\Omega_t; \gamma) \\ w(x_{t+1}) \log f(x_{t+1}|\Omega_t; \theta) \end{pmatrix}, \quad a_t(\beta) \equiv \begin{pmatrix} a_t^g(\gamma) \\ a_t^f(\theta) \end{pmatrix} \quad (7)$$

$$\delta_{qq}(j) = E[(q_t(\beta^*) - Eq_t(\beta^*))(q_{t-j}(\beta^*) - Eq_t(\beta^*))']$$

$$\delta_{qa}(j) = E[(q_t(\beta^*) - Eq_t(\beta^*))a_{t-j}(\beta^*)']$$

$$\delta_{aa}(j) = E[a_t(\beta^*)a_{t-j}(\beta^*)']$$

$$S_{qq} = \sum_{j=-\infty}^{\infty} \delta_{qq}(j), \quad S_{qa} = \sum_{j=-\infty}^{\infty} \delta_{qa}(j), \quad S_{aa} = \sum_{j=-\infty}^{\infty} \delta_{aa}(j)$$

$$F \equiv E \begin{pmatrix} w(X_{t+1})\nabla_{\gamma} \log g(X_{t+1}|\Omega_t; \gamma^*) & 0 \\ 0 & w(X_{t+1})\nabla_{\theta} \log f(X_{t+1}|\Omega_t; \theta^*) \end{pmatrix}, \quad B \equiv \begin{pmatrix} B^g & 0 \\ 0 & B^f \end{pmatrix}$$

$$\Pi \equiv 1 - \pi^{-1} \ln(1 + \pi) \text{ for } 0 < \pi < \infty, \quad \Pi \equiv 0 \text{ for } \pi = 0$$

$$\Omega \equiv S_{qq} + \Pi(FBS'_{qa} + S_{qa}B'F') + 2\Pi FBS_{aa}B'F'.$$

The following result provides the asymptotic weighted likelihood ratio test.

Theorem 1 *Given Assumptions 1, 2, 3, $\sqrt{n}(WLR_n - WLR^*) \xrightarrow{D} N(0, \sigma^2)$, where σ^2 is given by*

$$\sigma^2 = \iota\Omega\iota', \text{ with } \iota = (1, -1).$$

Let $\hat{\sigma}_n^2$ be a consistent estimator⁶ of σ^2 , then

(i) under H_0 : $\sqrt{n}WLR_n/\hat{\sigma}_n \xrightarrow{D} N(0, 1)$

(ii) under H_g : $\sqrt{n}WLR_n/\hat{\sigma}_n \xrightarrow{a.s.} +\infty$

(iii) under H_f : $\sqrt{n}WLR_n/\hat{\sigma}_n \xrightarrow{a.s.} -\infty$.

Proof. See Appendix. ■

For a desired level of confidence, one would first choose the corresponding critical value c from the standard normal distribution. If $|\sqrt{n}WLR_n/\hat{\sigma}_n| \leq c$ one would conclude that the two density forecasts give equal loss. If instead $|\sqrt{n}WLR_n/\hat{\sigma}_n| \geq c$, the null would be rejected in favour of H_g (if WLR_n is positive) or H_f (if WLR_n is negative). The test proposed has correct asymptotic size and is consistent, as reflected by the fact that the test statistic has a distribution that does not depend on the parameters under the null hypothesis, and it diverges under the alternative.

4.2 Bootstrap weighted likelihood ratio test

While the computation of the asymptotic test in the previous section can be quite involved, a test that is much easier to implement can be derived utilizing the bootstrap. The test is derived by resampling the test statistic WLR_n (6), in the following way.

⁶ A consistent estimate of the asymptotic variance can be obtained using kernel-based estimators of each component of Ω , such as the Newey-West (1987) estimators. Π can be estimated by $1 - (R/n) \ln(1 + n/R)$. See West (1996) or McCracken (2000) for a more thorough discussion on alternative ways to estimate each component.

A bootstrap artificial sample of size n is obtained by selecting random indexes $\tau(t)$, $t = R, \dots, T-1$ and considering the relative sequence of out-of-sample weighted likelihood ratios $\{wd_{\tau(t)+1}(\hat{\beta}_{\tau(t)}); t = R, \dots, T-1\}$. One can create B such artificial samples and for each calculate the resampled test statistic as

$$WLR_n^b \equiv n^{-1} \sum_{t=R}^{T-1} wd_{\tau(t)+1}(\hat{\beta}_{\tau(t)}), \quad b = 1, \dots, B. \quad (8)$$

There are now several available techniques to do resampling when the data are dependent, as in the time-series case. Popular examples are the moving blocks bootstrap of Künsch (1989) and Liu and Singh (1992) and the stationary bootstrap of Politis and Romano (1994). In the following, we focus attention on the Politis and Romano (1994) stationary bootstrap, but in principle other techniques can be used. The idea behind the stationary bootstrap is to resample blocks of random length, where the length of each block has a geometric distribution. Under some conditions on the growth rate of the average block length, Politis and Romano (1994) show that the stationary bootstrap resampling scheme satisfies desirable consistency and weak convergence properties. See also White (2000, p. 1104) for a description of how to implement the stationary bootstrap.

A bootstrap confidence interval for the weighted likelihood ratio statistic WLR_n can be obtained in many different ways (see Efron and Tibshirani (1993), or Shao and Tu (1995) for a discussion). We consider for simplicity an equal-tailed $(1 - \alpha)100\%$ confidence interval for WLR_n obtained as

$$CI = [WLR_n - q^*(1 - \alpha/2), WLR_n - q^*(\alpha/2)], \quad (9)$$

where $q^*(\alpha/2)$ and $q^*(1 - \alpha/2)$ are respectively the $\alpha/2$ and $1 - \alpha/2$ quantiles of the empirical distribution of $WLR_n^b - WLR_n$, $b = 1, \dots, B$. If $0 \notin CI$ we can reject the null hypothesis (11) of equal expected loss of density forecasts f and g , in favour of f (if $WLR_n < 0$) or in favour of g (if $WLR_n > 0$), at a confidence level α . The validity of this procedure rests on the assumption that the distribution of $\sqrt{n}(WLR_n^b - WLR_n)$, conditional on $\{X_{R+1}, \dots, X_T\}$ converges to the distribution of $\sqrt{n}(WLR_n - WLR^*)$, as n increases. This claim is proven by Politis and Romano (1994)'s Theorem 2 for the case when the resampled statistic depends on known parameters. In the presence of $\hat{\beta}_{\tau(t)}$ in (8), the validity of the bootstrap is obtained at the cost of imposing stronger conditions on the convergence of the parameter estimators to their probability limits and on the relative growth rates of the in-sample and the out-of-sample sizes. This point is argued by White (2000), to whom the reader is referred for a rigorous treatment. For our purposes, it suffices to add the following assumptions to the ones presented in Section 2.

Assumption 4: $\hat{\beta}_T$ obeys a law of the iterated logarithm.

Assumption 5: $(n/R) \log \log R \rightarrow 0$ as $T \rightarrow \infty$.

Notice that Assumption 4 is effectively a strengthening of Assumption 3. To guarantee validity of the bootstrap approximation one must thus impose a condition on the relative rate of divergence of R and n . Under Assumptions 1, 2, 4 and 5, Theorem 2.3 of White (2000) guarantees validity of the bootstrap weighted likelihood ratio test.

5 Likelihood ratio tests

In this section we show how a standard likelihood ratio test can be derived as a special case of the tests analyzed in the previous section, when the weight function is identically one. When the parameters of the models are Quasi-Maximum Likelihood Estimators (QMLE), we further show that the asymptotic likelihood ratio test becomes particularly easy to compute. Together with the asymptotic and bootstrap tests already proposed for the general weight function case, we further consider a bootstrap- t test.

The asymptotic test proposed in this section is related to Vuong (1989)'s likelihood ratio test for non-nested hypotheses. In that case, the comparison of alternative models is performed in-sample and under the assumption of independence and identical distribution of the variable of interest. In contrast, our approach focuses on the out-of-sample evaluation of density models, and we allow the variable to be characterized by conditional heterogeneity and serial dependence.

5.1 Asymptotic likelihood ratio test

When the weight function is $w(x) = 1$ the loss function (3) becomes the familiar Kullback-Leibler Information Criterion (KLIC), defined as $L(f, h) = I(h : f) = E[\log h(X_{t+1}|\Omega_t) - \log f(X_{t+1}|\Omega_t; \theta^*)]$. The relative distance of forecasts f and g from the true density h is then given by

$$LR^* \equiv I(h : f) - I(h : g) = E[\log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*)]. \quad (10)$$

A positive value of LR^* would indicate that forecast g will be considered preferable to forecast f by a user whose loss function is the KLIC. A test of equal performance of the two models is a test of

the null hypothesis:

$$\begin{aligned}
H_0 & : LR^* = 0 \text{ against} & (11) \\
H_g & : LR^* > 0 \text{ or} \\
H_f & : LR^* < 0 .
\end{aligned}$$

Similarly to the developments in Section 4, we estimate LR^* by the out-of-sample mean

$$LR_n = n^{-1} \sum_{t=R}^{T-1} [\log g(x_{t+1}|\Omega_t; \hat{\gamma}_t) - \log f(x_{t+1}|\Omega_t; \hat{\theta}_t)], \quad (12)$$

where $\{x_{t+1}\}_{t=R}^{T-1}$ are the realizations of the variable and the parameter estimates $\hat{\gamma}_t$ and $\hat{\theta}_t$ are QMLEs.

Let $d_{t+1}(\beta^*) \equiv \log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*)$, $\beta^* = (\gamma^{*'}, \theta^{*'})'$ and define

$$\begin{aligned}
\sigma^2 & \equiv \sum_{j=-\infty}^{+\infty} \delta_{dd}(j), \text{ where} & (13) \\
\delta_{dd}(j) & = E[(d_t(\beta^*) - Ed_t(\beta^*))(d_{t-j}(\beta^*) - Ed_{t-j}(\beta^*))']
\end{aligned}$$

The following result provides the asymptotic likelihood ratio test.

Theorem 2 *Given Assumptions 1, 2, 3, $\sqrt{n}(LR_n - LR^*) \xrightarrow{D} N(0, \sigma^2)$. Let $\hat{\sigma}_n^2$ be a consistent estimator of σ^2 , then*

$$(i) \text{ under } H_0 : \sqrt{n}LR_n/\hat{\sigma}_n \xrightarrow{D} N(0, 1)$$

$$(ii) \text{ under } H_g : \sqrt{n}LR_n/\hat{\sigma}_n \xrightarrow{a.s.} +\infty$$

$$(iii) \text{ under } H_f : \sqrt{n}LR_n/\hat{\sigma}_n \xrightarrow{a.s.} -\infty.$$

Proof. See Appendix. ■

The expression for the asymptotic variance (13) reveals that estimation uncertainty is asymptotically irrelevant under the assumptions of the theorem. The asymptotic variance is in fact the same that would have been obtained had the parameters been known (as assumed by Diebold and Mariano, 1995), and it coincides (apart from a scale factor) with the spectral density of the de-meanded loglikelihood differences $\{d_t(\beta^*) - Ed_t(\beta^*)\}$ at frequency zero. This is somewhat a special case. If the estimator used is not QMLE, or if the two density forecasts are conditional on different information sets, for example, estimation uncertainty becomes relevant for the asymptotic distribution, as we saw in Section 4.1. In these situations, asymptotic irrelevance of parameter estimation uncertainty can be attained by imposing that Assumption 3 holds with $\pi = 0$ (see West, 1996 for a discussion).

5.2 Bootstrap likelihood ratio tests

The asymptotic test can be complemented with and compared to the bootstrap test that was proposed in Section 4, with a weight function that is now identically one.

For the likelihood ratio test, we also explore the possibility of improving the accuracy of the bootstrap confidence interval CI in (9). We do so by use of the so-called ‘bootstrap- t ’ approach (see, e.g., Efron and Tibshirani, 1993). The procedure deviates from the one described in Section 4 in that it requires to calculate for each artificial sample b , both the test statistic LR_n^b ⁷ and an estimate $\hat{\sigma}_n^b$ of its standard deviation. The relevant standard deviation is σ/\sqrt{n} , where σ is the square root of the asymptotic variance defined in (13), and an estimate can be obtained by kernel-based estimators, such as the Newey-West (1987) estimator. A bootstrap- t confidence interval for LR_n with $(1 - \alpha)100\%$ nominal coverage is then computed as

$$CI - t = [LR_n - u^*(1 - \alpha/2)\hat{\sigma}_n, LR_n - u^*(\alpha/2)\hat{\sigma}_n], \quad (14)$$

where $u^*(1 - \alpha/2)$ and $u^*(\alpha/2)$ are the $1 - \alpha/2$ and the $\alpha/2$ quantiles of the empirical distribution of $\frac{LR_n^b - LR_n}{\hat{\sigma}_n^b}$, $b = 1, \dots, B$ and $\hat{\sigma}_n$ is an estimate of the standard deviation of LR_n . A theoretical result (e.g., Shao and Tu, 1995) proves that the bootstrap- t confidence interval (14) is more accurate⁸ than the bootstrap confidence interval (9) or the confidence interval implied by the asymptotic normal approximation in Theorem 1. The higher order accuracy of the bootstrap- t confidence interval is an asymptotic result. In practice, its superior performance in finite samples will likely depend on the quality of the estimator for the variance. It is thus a worthwhile exercise to contrast the performance of the proposed tests in samples of the sizes typically available in practice.

6 Monte Carlo experiment

In this section, we analyze and compare the size and power of the likelihood ratio tests proposed in the previous section. Due to the inherent difficulties in finding plausible non-nested models that satisfy the null hypothesis, we restrict attention to the case where the two density forecasts are both normal but their specification for the conditional mean depends on a common autoregressive term and on an exogenous variable, which is different for the two densities. This situation could arise in practical applications when, for example, two economic theories postulate that different information sets have predictive content for the variable of interest. Let X_t be the variable of interest, and Z_{1t}

⁷ LR_n^b is equivalent to WLR_n^b in (8) with $w(\cdot)$ identically equal to one.

⁸Higher accuracy of a confidence interval means that its coverage level is closer to the nominal level $(1 - \alpha)100\%$.

and Z_{2t} be the exogenous variables. Let $\Omega_t = \{x_{t-j}, z_{1t+1-j}, z_{2t+1-j}; j \geq 0\}$ be the information set at time t . We consider the following specifications for the true conditional density and for density forecasts f and g .

$$\begin{aligned}
 DGP & : X_t|\Omega_t \sim N(\rho^*x_{t-1} + \alpha^*z_{1t} + \beta^*z_{2t}, 1) \equiv h \\
 Forecast1 & : X_t|\Omega_t \sim N(\rho x_{t-1} + \alpha z_{1t}, 1) \equiv f \\
 Forecast2 & : X_t|\Omega_t \sim N(\rho x_{t-1} + \beta z_{2t}, 1) \equiv g
 \end{aligned} \tag{15}$$

The variables Z_{1t} and Z_{2t} are independent $N(0, 1)$ random variables. It can be easily shown that the probability limits for the parameters of f and g , i.e., the parameters that minimize the KLIC between h and each density, are respectively $(\rho^*, \alpha^*)'$ and $(\rho^*, \beta^*)'$. The relative distance of f and g from the truth, as measured by LR^* in (10), can then be shown to equal $LR^* = (\beta^*)^2 - (\alpha^*)^2$. A parameterization for the DGP such that $\alpha^* = \beta^*$ will thus satisfy the null hypothesis, and it will be used to investigate the size of the tests. We obtain a power curve by keeping $\alpha^* = 0.1$ fixed and increasing β^* .

In the experiment, we consider a range of different sizes for the in-sample (R) and the out-of-sample (n) parts and perform 1000 Monte Carlo iterations for each pair (R, n) and (α^*, β^*) . We consider a total of 15 different combinations of in-sample (R) and out-of-sample (n) sizes: $R = 50, 100, 150$ and $n = 25, 50, 75, 100, 150$. This range and relative proportion of in-sample and out-of-sample sizes seems broad enough to represent the typical situation of macroeconomic forecasting. To represent the larger sample sizes that may arise in some financial applications, we further consider the two pairs $(R, n) = (500, 250)$ and $(R, n) = (650, 350)$. For each iteration, the parameters of f and g are estimated by ML on the first sample of size R . The density forecasts for period $R + 1$ are then formed as $f : N(\hat{\rho}_R x_R + \hat{\alpha}_R z_{1R+1}, 1)$ and $g : N(\hat{\rho}_R x_R + \hat{\beta}_R z_{2R+1}, 1)$. Each density is evaluated at the realized value for the variable x_{R+1} and the first observed loglikelihood ratio is obtained as $d_{R+1} = \log g(x_{R+1}|\Omega_R; \hat{\rho}_R, \hat{\beta}_R) - \log f(x_{R+1}|\Omega_R; \hat{\rho}_R, \hat{\alpha}_R)$. The sample is then augmented by including observation x_{R+1} and the procedure is repeated on the sample of size $R + 1$ to obtain the loglikelihood ratio d_{R+2} , and so forth. The recursion generates a total of n log-likelihoods d_{R+1}, \dots, d_{R+n} that are averaged to obtain LR_n as in (12). The three likelihood ratio tests proposed in Sections 5 are then performed and their rejection frequencies calculated over the Monte Carlo iterations. We refer to the three tests as ‘asymptotic LR test’, ‘bootstrap LR test’ and ‘bootstrap-t LR test’. As noted in Section 5.1, when the forecasts are conditional on different information sets the asymptotic variance is (13) as long as Assumption 3 holds with $\pi = 0$. We implicitly impose this requirement, and use

(13) in the computations for the asymptotic LR test. The impact of ignoring estimation uncertainty will emerge from the analysis of the power curves for decreasing π . Table 1 reports the empirical size of the three tests for nominal size .05.

[TABLE 1 HERE]

The asymptotic LR test is oversized for an out-of-sample size $n \leq 50$. A mild tendency to overreject is still present for the two bootstrap tests when the size of the out-of-sample is small, but they are overall better sized than the asymptotic test for all combinations of R and n . All tests have good size for an out-of-sample $n \geq 75$.

Figures 2-5 show the power curves for a selection of in-sample and out-of-sample pairs.

[FIGURES 2-5 HERE]

On the horizontal axis, instead of reporting the increasing distance between the values of the coefficients α^* and β^* , we choose to report the corresponding difference in R^2 from the regressions that define density forecasts f and g .⁹ For the size study we let $\alpha^* = \beta^* = 0.1$, which implies an equal R^2 for the two models of 0.36, and for the power study we let β^* increase, so that the difference in R^2 for the two forecasts varies between 0 and 0.5. The graphs are presented in order of increasing out-of-sample size. A pattern that emerges from the figures is that the three tests have different power when the out-of-sample size n is small, but the power tends to become equal across the three tests for larger n . For $n = 25$ (Figure 2), the asymptotic LR test has higher power than the bootstrap tests, at the price of high size distortions. For the same value of n , the bootstrap-t LR test has slightly lower power than the bootstrap LR test, but this divergence in power between the bootstrap tests disappears for $n \geq 75$. Notice also that all power curves become steeper as the out-of-sample size increases, a sign of the consistency of the tests. A final conclusion that emerges from the comparison of the two panels in each figure is that the size of the in-sample seems not to affect the properties of the tests. The size and power differences seem to be driven only by the size of the out-of-sample. This is a hopeful indication that imposing the condition that the in-sample size grows faster than the out-of-sample (i.e., Assumption 3 with $\pi = 0$) does not affect the properties of the tests, while considerably simplifying the estimation of the asymptotic variance of the likelihood ratio test. This in turn would suggest that ignoring parameter estimation uncertainty does not significantly alter the properties of the likelihood ratio tests in our framework.

⁹The R^2 from each regression can be shown to equal $R_f^2 = 1 - (1 - \rho^{*2})(\beta^{*2} + 1)/(\alpha^{*2} + \beta^{*2} + 1)$ and $R_g^2 = 1 - (1 - \rho^{*2})(\alpha^{*2} + 1)/(\alpha^{*2} + \beta^{*2} + 1)$ and thus the difference is given by $R_f^2 - R_g^2 = (1 - \rho^{*2})(\alpha^{*2} - \beta^{*2})/(\alpha^{*2} + \beta^{*2} + 1)$

7 Empirical application

In this section, the tests proposed in the paper are utilized to compare one-step-ahead univariate density forecasts for the S&P500 obtained from GARCH-type models with different distributional assumptions. The data are daily U.S. returns on the S&P500 from 1/1/1990 to 8/3/2001 obtained from Datastream. The return series is derived from the price index data, p_t , as $x_t = 100 \log(p_t/p_{t-1})$, so that x_t represents the continuously compounded return (in percent) on the index. We model the returns as a GARCH(1,1) process

$$\begin{aligned}x_t &= \mu + \varepsilon_t, \\ \varepsilon_t | \Omega_{t-1} &\sim f(0, \sigma_t), \\ \sigma_t^2 &= w + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2\end{aligned}\tag{16}$$

This model specification will generate density forecasts with time-varying variance and functional form specified by f . When the model (16) is used for prediction, the accuracy of the forecast critically depends on the distribution chosen for the standardized residuals. It is widely acknowledged that a normality assumption for the conditional distribution of the disturbances ε_t does not account for the excess kurtosis that characterizes the residuals from a fitted GARCH model¹⁰. A way to incorporate excess kurtosis in the model is the use of fat-tailed distributions for the disturbances. Popular examples are the Student's t (proposed by Bollerslev, 1987) and the generalized error distribution (GED), utilized by Nelson, 1991. A further characteristic that one might want to account for is the possible skewness of returns, that can be captured, for example, by a skewed- t distribution (see Hansen, 1994). In the following, we will refer to density forecasts generated by model (16) with the three different specifications for the disturbances as t -GARCH, GED -GARCH and $skewt$ -GARCH. The goal is to do pairwise comparisons to select the model that yields the most accurate density forecasts, and then isolate the regions of the distribution where such outperformance takes place. We are able to apply the tests proposed in the paper to the non-nested pairs (t -GARCH, GED -GARCH) and ($skewt$ -GARCH, GED -GARCH). The comparison of out-of-sample performance of t -GARCH and $skewt$ -GARCH is not possible due to the nestedness of the two models.

We proceed as follows. We divide the available sample of $T = 2928$ observations in two parts, using the first R data for estimation and leaving the remaining n observations for out-of-sample evaluation. We consider a range of sizes for n that varies between $n = 100$ (and thus $R = 2828$) and $n = 1500$

¹⁰Evidence of non-normality of financial assets' returns has been documented, *inter alia*, by Mandelbrot (1963), Fama (1965), Bollerslev (1987).

(which corresponds to $R = 1428$), using increments of 100 in n . The first set of density forecasts is obtained by estimating the parameters of (16) for each of the three distributional assumptions on the sample of size R . This involves estimating the unconditional mean $\hat{\mu}$, the GARCH parameters $(\hat{\omega}, \hat{\alpha}, \hat{\beta})$ and the shape parameters of each distribution by QML. The t -GARCH density forecast for the variable X_{R+1} is, for example, a Student's t with mean $\hat{\mu}$, variance $\sigma_{R+1}^2 = \hat{\omega} + \hat{\alpha}\varepsilon_R^2 + \hat{\beta}\sigma_R^2$ and shape parameter as estimated on the sample of size R . For each of the three density forecasts, we then evaluate the log of the density at the realized value of x_{R+1} . This generates the first set of three out-of-sample log-likelihoods. The procedure is then repeated on the samples of sizes $R+1, \dots, T-1$, and it yields a sequence of n out-of-sample log-likelihoods.

Before proceeding with the tests, we consider the sequence of volatility forecasts implied by the three different models, that are plotted in Figure 6 for the case of $n = 1500$.

[FIGURE 6 HERE]

The volatility forecasts implied by the different distributional assumptions are virtually indistinguishable. As a consequence, any approach to model selection that relies on the comparison on volatility forecast accuracy is likely doomed to fail. The tests proposed in the paper, on the other hand, will help detect any superior predictive ability that is solely due to the different distributional assumptions of the three models.

The first set of results compares the accuracy of t -GARCH and GED -GARCH density forecasts.

[FIGURES 7-9 HERE]

The LR_n statistic is computed by letting g be the GED density and f the Student's t density. A value of LR_n that is significantly less than zero will thus mean that the t -GARCH density forecasts outperform the GED -GARCH forecasts. Figure 7 plots the p-values of the asymptotic LR test for increasing out-of-sample size n . Equal performance of the two density forecasts is rejected at a 95% confidence level for values of n greater than 1000. For these sample sizes, the LR_n is negative (Figure 8), indicating that t -GARCH forecasts are more accurate than GED -GARCH forecasts. This conclusion is confirmed by the bootstrap-t LR test, reported in Figure 8. The figure shows the value of LR_n for different n and the relative bootstrap 95% confidence interval. For values of n greater than 1000, LR_n is significantly negative.

To detect in what regions of the unconditional distribution of the returns the t -GARCH density forecasts outperform the GED -GARCH forecasts, we consider the weighted likelihood ratio tests

for the four weight functions shown in Figure 1. The four panels in Figure 9 plot the values of the weighted likelihood ratio statistic WLR_n for the different weight functions, together with 95% bootstrap confidence intervals. The value of WLR_n is significantly negative for the weight function w_1 , which represents the center of the unconditional distribution. This indicates that the t -GARCH density forecasts significantly outperform the GED -GARCH forecasts for values of returns that fall near the center of the unconditional distribution. In other words, t -GARCH density forecasts are better than GED -GARCH forecasts at predicting returns in ‘normal’ days, while the two densities are equally accurate in predicting ‘extreme’ events. We now investigate whether allowing the best model (t -GARCH) to incorporate skewness will lead to further improvements in its performance relative to the GED -GARCH forecast model.

The second set of results compares the performance of $skewt$ -GARCH and GED -GARCH density forecasts. Figures 10-12 are the equivalent of Figures 7-9 for the new pair of density forecasts.

[FIGURES 10-12 HERE]

As expected, given the nestedness of $skewt$ and t distributions, Figures 10 and 11 lead to the same conclusion highlighted by Figures 7 and 8. The $skewt$ -GARCH density forecasts outperform the GED -GARCH forecasts. Figure 12 reveals that the $skewt$ -GARCH outperforms the GED -GARCH in both the center and the right tail of the unconditional distribution of returns. That is, the $skewt$ -GARCH is better at predicting returns in normal days and in days when returns are relatively high (the definition of what constitutes a normal day and a high return is necessarily imprecise, given the arbitrary choice of weight function). In conclusion, a t -GARCH or $skewt$ -GARCH density forecasts are seen to be better approximations to the true conditional density of the S&P 500 returns than a GED -GARCH density forecast. A similar finding has been documented by Bollerslev, Engle and Nelson (1994) in a non-predictive setting.

8 Conclusion

The paper proposed a number of tests that can be used to compare conditional density forecasts, or in the more general context of model selection based on out-of-sample forecasting performance. The tests can be utilized in both a univariate and a multivariate setting, even though in the paper we focused for simplicity on the univariate case. We restricted attention to the case of density forecasts derived from conditional parametric models that are non-nested or overlapping, with known or estimated parameters. One of the main goals of the paper was to incorporate loss functions

into the evaluation of density models, an issue that has not yet been explored in the literature. To this purpose, we proposed a class of loss functions defined over the distance between the a density forecast and the true density. We pointed out how our discussion of loss functions for density forecasting evaluation differs from the treatment in the point forecasting literature, where the loss functions are defined over the forecast error. The class of loss functions considered led to the development of ‘weighted likelihood ratio’ tests for the comparison of density forecasts. We showed how these tests can be utilized to isolate the performance of competing density forecasts in different regions of the unconditional distribution of the variable of interest. Loosely speaking, the tests can help distinguish, for example, the relative performance of the models in ‘normal’ days from the performance in days when the variable takes on ‘extreme’ values. We proposed asymptotic and bootstrap weighted likelihood ratio tests, and the asymptotic test was shown to be consistent and to have correct asymptotic size. The familiar likelihood ratio test for non-nested hypothesis of Vuong (1989) was seen to be a special case of a weighted likelihood ratio test, for a weight function identically equal to one. When the parameters of the density forecasts are estimated by maximum likelihood, we further pointed out that the asymptotic likelihood ratio test becomes particularly easy to compute. The performance of the asymptotic and bootstrap likelihood ratio tests in finite samples was analyzed through a Monte Carlo simulation that considered the case where the alternative density forecasts have the same functional form, but use different exogenous variables in their conditional mean specification. We found the asymptotic likelihood ratio test to be oversized for very small out-of-sample sizes, while the bootstrap tests had good size and power for all combinations of in-sample and out-of-sample sizes. The tests proposed in the paper were finally used in an empirical application aiming to compare density forecasts obtained from GARCH-type models with different distributional assumptions for the standardized residuals. The data considered was the series of S&P 500 daily returns and the models used for constructing density forecast were GARCH(1,1) with Student’s t , generalized error distribution (GED) and skewed Student’s t ($skewt$) disturbances. We concluded that density forecasts from a t -GARCH and from a $skewt$ -GARCH models are more accurate than density forecasts from a GED -GARCH model, and the superior performance was seen to occur in ‘normal days’. The $skewt$ -GARCH forecasts also outperformed the GED -GARCH forecasts on days when the returns on the S&P 500 are large and positive.

9 Proofs

Proof. Theorem 1. Since $WLR_n = \iota n^{-1} \sum_{t=R}^{T-1} q_{t+1}(\hat{\beta}_t)$ and $WLR^* = \iota E q_{t+1}(\beta^*)$, we have that $\sqrt{n}(WLR_n - WLR^*) = \iota \sqrt{n} [n^{-1} \sum_{t=R}^{T-1} q_{t+1}(\hat{\beta}_t) - E q_{t+1}(\beta^*)]$. We will make use of Theorem 4.1 of West (1996) to show that

$$\sqrt{n} [n^{-1} \sum_{t=R}^{T-1} q_{t+1}(\hat{\beta}_t) - E q_{t+1}(\beta^*)] \xrightarrow{D} N(0, \Omega), \quad (17)$$

from which it follows that $\sqrt{n}(WLR_n - WLR^*) \xrightarrow{D} N(0, \iota \Omega')$. To be able to apply Theorem 4.1, we must first show that Assumptions 1-4 of West (1996) (which we will call W1-W4) are satisfied by the vector $q_{t+1}(\beta)$.

W1-(a) requires $q_{t+1}(\beta)$ to be measurable and twice continuously differentiable in a neighborhood of β^* , which is implied by Assumption 1-(a).

W1-(b) requires the matrix $\nabla_{\beta}^2 q_{t+1}(\beta) = \begin{pmatrix} \nabla_{\gamma}^2 \log g(X_{t+1} | \Omega_t, \gamma) & 0 \\ 0 & \nabla_{\theta}^2 \log f(X_{t+1} | \Omega_t, \theta) \end{pmatrix}$ to be bounded by a variable with finite expectation, which follows from the boundedness of the two diagonal components imposed by Assumption 1-(b).

W2 assumes that the parameter estimates $\hat{\beta}_t$ can be written as $\hat{\beta}_t - \beta^* = B(t)A(t)$, where $B(t) \xrightarrow{a.s.} B$, B a matrix of rank $k = k_1 + k_2$, and $A(t) = t^{-1} \sum_{s=1}^t a_s(\beta^*)$, with $E a_s(\beta^*) = 0$. This follows directly from Assumption 1-(c), by letting $B(t) \equiv \begin{pmatrix} B^g(t) & 0 \\ 0 & B^f(t) \end{pmatrix}$ and B and $a_t(\beta^*)$ to be as defined in (7).

W3-(a) imposes a bound on the fourth moments of $a_t(\beta^*)$, $q_{t+1}(\beta^*)$ and $\nabla_{\beta} q_{t+1}(\beta^*)$. Existence of the fourth moments of $a_t(\beta^*)$ is directly implied by Assumption 1-(d). The boundedness on the fourth moments of $q_{t+1}(\beta^*)$ and $\nabla_{\beta} q_{t+1}(\beta^*)$ is implied by the existence of the fourth moments of each component of the two vectors. For illustration, we only prove this claim for the second component of $q_{t+1}(\beta^*)$, which equals $w(X_{t+1}) \log f(X_{t+1} | \Omega_t, \theta^*)$. From Assumption 1-(d), it follows that there exists a $d > 1$ such that $E |\log f(X_{t+1} | \Omega_t, \theta^*)|^{4d} < \infty$. Consider $E |w(X_{t+1}) \log f(X_{t+1} | \Omega_t, \theta^*)|^{4d'}$, with $d' = \frac{d}{1+\varepsilon}$, for some $\varepsilon > 0$. Since $w(X_{t+1}) \geq 0$, and by applying Hölder's inequality, we have

$$\begin{aligned} E |w(X_{t+1}) \log f(X_{t+1} | \Omega_t, \theta^*)|^{4d'} &= E |w(X_{t+1})|^{4d'} |\log f(X_{t+1} | \Omega_t, \theta^*)|^{4d'} \\ &\leq (E |w(X_{t+1})|^{4d' \frac{1+\varepsilon}{\varepsilon}})^{\frac{\varepsilon}{1+\varepsilon}} (E |\log f(X_{t+1} | \Omega_t, \theta^*)|^{4d'(1+\varepsilon)})^{\frac{1}{1+\varepsilon}} \\ &= (E |w(X_{t+1})|^{\frac{4d}{\varepsilon}})^{\frac{\varepsilon}{1+\varepsilon}} (E |\log f(X_{t+1} | \Omega_t, \theta^*)|^{4d})^{\frac{1}{1+\varepsilon}} < \infty, \end{aligned}$$

since the first term is finite because of boundedness of $w(\cdot)$, and the second term is bounded by Assumption 1-(d).

W3-(b) assumes the vector $[q_{t+1}(\beta^*)', \text{vec}(\nabla_{\beta} q_{t+1}(\beta^*))', a_t(\beta^*)']'$ to be strong mixing of size $-3d/(d-1)$. This follows from $\{X_t\}$ being strong mixing of size $-3d/(d-1)$ by Assumption 1-(e), and from Lemma 2.1 of White and Domowitz (1984), showing that measurable functions of mixing processes are mixing of the same size.

W3-(c) requires $[q_{t+1}(\beta^*)', \text{vec}(\nabla_{\beta} q_{t+1}(\beta^*))', a_t(\beta^*)']'$ to be covariance stationary, which is directly implied by Assumption 1-(f).

W3-(d) assumes S_{qq} defined in (7) to be positive definite. This is ensured by Assumption 2, requiring the two components of $q_{t+1}(\beta^*)$ to be distinct.

W4 coincides with Assumption 3.

We can thus apply Theorem 4.1 of West (1996) to prove (17).

The second part of the theorem follows straightforwardly from the fact that $\sqrt{n}(WLR_n - WLR^*) \xrightarrow{D} N(0, \sigma^2)$, with $\sigma^2 > 0$ due to positive definiteness of Ω , and from Slutsky's Theorem.

Theorem 2. When $w(x) = 1$, for all x , and the parameter estimates are QMLEs, it follows that $F \equiv E \begin{pmatrix} \nabla_{\gamma} \log g(X_{t+1}|\Omega_t; \gamma^*) & 0 \\ 0 & \nabla_{\theta} \log f(X_{t+1}|\Omega_t; \theta^*) \end{pmatrix} = \mathbf{0}$. In this case, thus, the matrix Ω defined in (7) reduces to S_{qq} , and the asymptotic variance of the likelihood ratio test becomes $\sigma^2 = \iota S_{qq} \iota' = \sum_{j=-\infty}^{+\infty} \iota \delta_{qq}(j) \iota' = \sum_{j=-\infty}^{+\infty} \delta_{dd}(j)$. ■

References

- [1] Andrews, D. W. K. (1991): ‘Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation’, *Econometrica*, 59, 817-858.
- [2] Artzner, P., Delbaen, F., Eber, J. M. and Heath, D. (1997): ‘Thinking Coherently’, *Risk*, 10, 68-71.
- [3] Bauwens, L., Giot, P., Grammig, J. and Veredas, D. (2000): ‘A Comparison of Financial Duration Models via Density Forecasts’, CORE Discussion Paper.
- [4] Berkowitz, J. (2000): ‘The Accuracy of Density Forecasts in Risk Management’, forthcoming *Journal of Business and Economic Statistics*.
- [5] Bollerslev, T. (1987): ‘A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return’, *Review of Economics and Statistics*, 69, 542-547.
- [6] Bollerslev, T., Engle, R. F. and Nelson, D. B. (1994): ‘ARCH Models’, in *Handbook of Econometrics*, vol. 4, Chapter 49, Engle, R. F. and McFadden, D. (eds.), Elsevier Science B. V., Amsterdam, The Netherlands.
- [7] Christoffersen P. F. and Diebold, F. X. (1997): ‘Optimal Prediction under Asymmetric Loss’, *Econometric Theory*, 13, 808-817.
- [8] Christoffersen P. F. and Diebold, F. X. (1996): ‘Further Results on Forecasting and Model Selection under Asymmetric Loss’, *Journal of Applied Econometrics*, 11, 561-571.
- [9] Clements, M. P., Smith, J. (2000): ‘Evaluating the Forecast Densities of Linear and Non-linear Models: Applications to Output Growth and Unemployment’, *Journal of Forecasting*, 19, 255-276.
- [10] Diebold, F. X., Gunther, T. A., Tay, A. S. (1998): ‘Evaluating Density Forecasts with Applications to Financial Risk Management’, *International Economic Review*, 39, 863-883.
- [11] Diebold, F. X., Mariano, R. S. (1995): ‘Comparing Predictive Accuracy’, *Journal of Business and Economic Statistics*, 13, 253-263.
- [12] Diebold, F. X., Tay, A.S., Wallis, K. F. (1999): ”Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters ”, in R. Engle and H. White (eds.), *Festschrift in Honour of C.W.J. Granger*, 76-90, Oxford University Press.

- [13] Duffie, D. and Pan, J. (1996): ‘An Overview of Value at Risk’, *Journal of Derivatives*, 4, 13-32.
- [14] Efron, B. and Tibshirani, R. J. (1993): *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- [15] Fama, E. F. (1965): ‘The Behavior of Stock-Market Prices’, *Journal of Business* 38, 34-105.
- [16] Fisher, R. A. (1932): *Statistical Methods for Research Workers*.
- [17] Granger, C. W. J., Pesaran, M. H. (2000): ‘A Decision-Theoretic Approach to Forecast Evaluation’, in *Statistics and Finance: An Interface*, W. S. Chan, W. K. Li and H. Tong (eds.), Imperial College Press, London.
- [18] Hansen, B. (1994): ‘Autoregressive Conditional Density Estimation’, *International Economic Review*, 35, 705-730.
- [19] Hong, Y. (2000): ‘Evaluation of Out-of-Sample Density Forecasts with Applications to S&P 500 Stock Prices’, manuscript.
- [20] Hong, Y. and White, H. (2000): ‘Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence’, manuscript.
- [21] Künsch, H. R. (1989): ‘The Jackknife and the Bootstrap for General Stationary Observations’, *Annals of Statistics*, 17, 1217-1241
- [22] Liu, R. Y. and Singh, K. (1992): ‘Moving Blocks Jackknife and Bootstrap Capture Weak Dependence ’ in *Exploring the Limits of the Bootstrap*, R. LePage and L. Billiard (eds), Wiley, New York
- [23] Mandelbrot, B. (1963): ‘The Variation of Certain Speculative Prices’, *Journal of Business*, 26, 395-419.
- [24] McCracken, M. W. (2000): ‘Robust out-of-sample Inference’, *Journal of Econometrics*, 99, 195-223.
- [25] Nelson, D. (1991): ‘Conditional Heteroskedasticity in Asset Returns: A New Approach’, *Econometrica*, 59, 347-370.
- [26] Newey, W. K. and West, K. D. (1987): ‘A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix”, *Econometrica*, 55, 703-708.

- [27] Patton, A. J. (2001a): ‘Modelling Time-Varying Exchange Rate Dependence using the Conditional Copula’, UCSD working paper n. 2001-09
- [28] Patton, A. J. (2001b): ‘On the Importance of Skewness and Asymmetric Dependence in Stock Returns for Asset Allocation’, UCSD manuscript.
- [29] Pesaran, M. H., Skouras, S. (2000): ‘Decision-Theoretic Methods for Forecast Evaluation’, University of Cambridge manuscript.
- [30] Politis, D. and Romano, J. (1994): ‘The Stationary Bootstrap’, *Journal of the American Statistical Association*, 89, 1303-1313
- [31] Rivers, D. and Vuong, Q. (1999): ‘Model Selection Tests for Nonlinear Dynamic Models’, University of Southern California manuscript
- [32] Shao, J. and Tu, D. (1995): *The Jackknife and Bootstrap*, Springer-Verlag, New York
- [33] Soderlind, P. and Svensson, L. (1997): ‘New Techniques to Extract Market Expectations from Financial Instruments’, *Journal of Monetary Economics*, 40, 383-429.
- [34] Tay, A. S. and Wallis, K. F. (2000): ‘Density Forecasting: A Survey’, *Journal of Forecasting*, 19, 235-254.
- [35] Vuong, Q. H. (1989): ‘Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses’, *Econometrica*, 57, 307-333.
- [36] Weigend, A. S., Shi, S. (2000): ‘Predicting Daily Probability Distributions of S&P500 Returns’, *Journal of Forecasting*, 19, 375-392.
- [37] Weiss, A. A. (1996): ‘Estimating Time Series Models Using the Relevant Cost Function’, *Journal of Applied Econometrics*, 11, 539-560.
- [38] West, K. D. (1996): ‘Asymptotic Inference about Predictive Ability’, *Econometrica*, 64, 1067-1084.
- [39] White, H. (2000): ‘A Reality Check for Data Snooping’, *Econometrica*, 68, 1097-1126.
- [40] White, H. and Domowitz, I. (1984): ‘Nonlinear regression with dependent observations’, *Econometrica*, 52, 143-162.

Table 1

Size of nominal .05 tests

A. LR asymptotic test					
R	n				
	25	50	75	100	150
50	0.162	0.095	0.059	0.042	0.047
100	0.158	0.160	0.076	0.078	0.064
150	0.197	0.167	0.086	0.053	0.065
B. LR bootstrap test					
R	n				
	25	50	75	100	150
50	0.059	0.058	0.037	0.033	0.033
100	0.072	0.065	0.047	0.052	0.050
150	0.092	0.073	0.052	0.036	0.054
C. LR bootstrap-t test					
R	n				
	25	50	75	100	150
50	0.080	0.063	0.044	0.047	0.046
100	0.066	0.073	0.058	0.069	0.055
150	0.094	0.077	0.064	0.048	0.061

Notes: Each panel reports the empirical size of the three likelihood ratio tests discussed in Section 5. Entries represent the rejection frequencies over 1000 Monte Carlo replications of the null hypothesis $H_0 : E[\log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*)] = 0$, where the density forecasts f , g and the DGP are defined in (15). The nominal size is .05. Each cell corresponds to a pair of in-sample and out-of-sample sizes (R, n) .

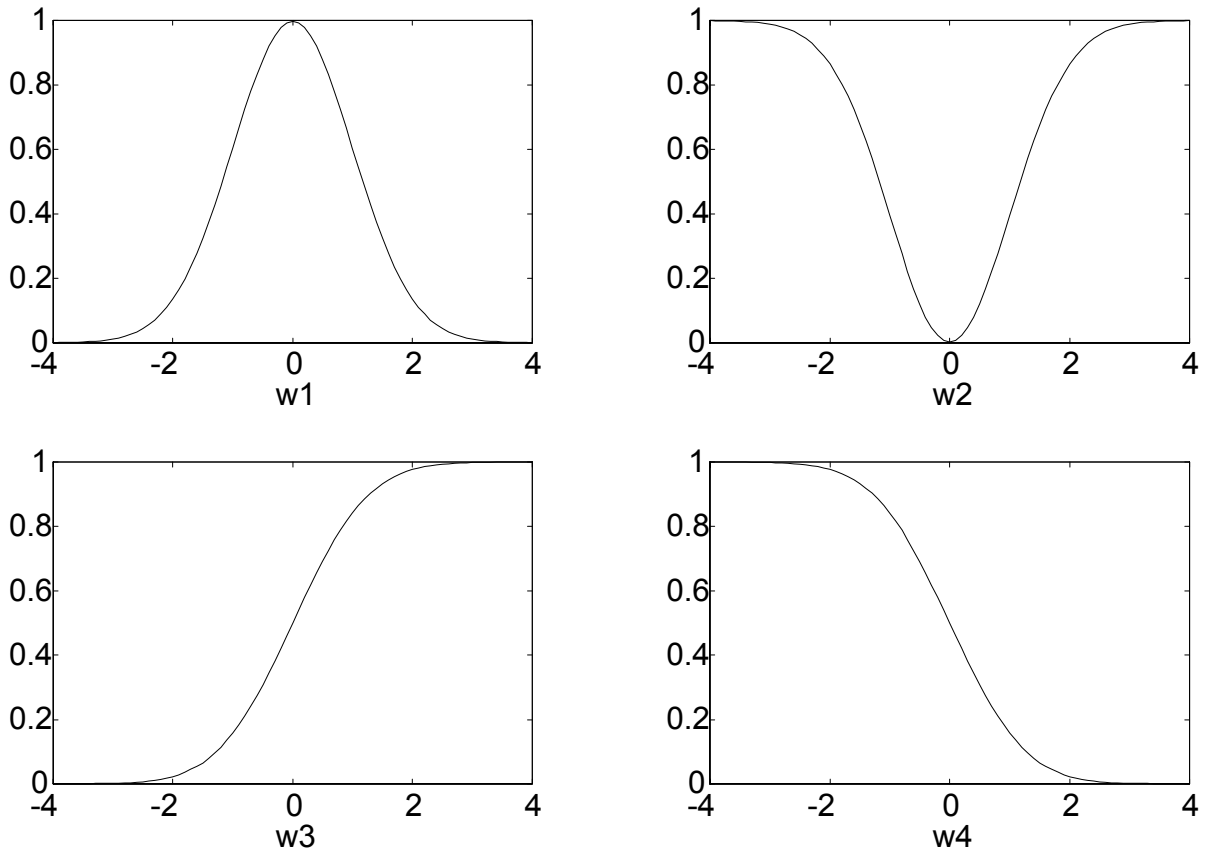


Figure 1: Weight functions for the weighted likelihood ratio test. $w_1(x) = \phi(x)$, $w_2(x) = 1 - 2.5\phi(x)$, $w_3(x) = \Phi(x)$, $w_4(x) = 1 - \Phi(x)$, where ϕ and Φ are, respectively, the standard normal pdf and cdf.

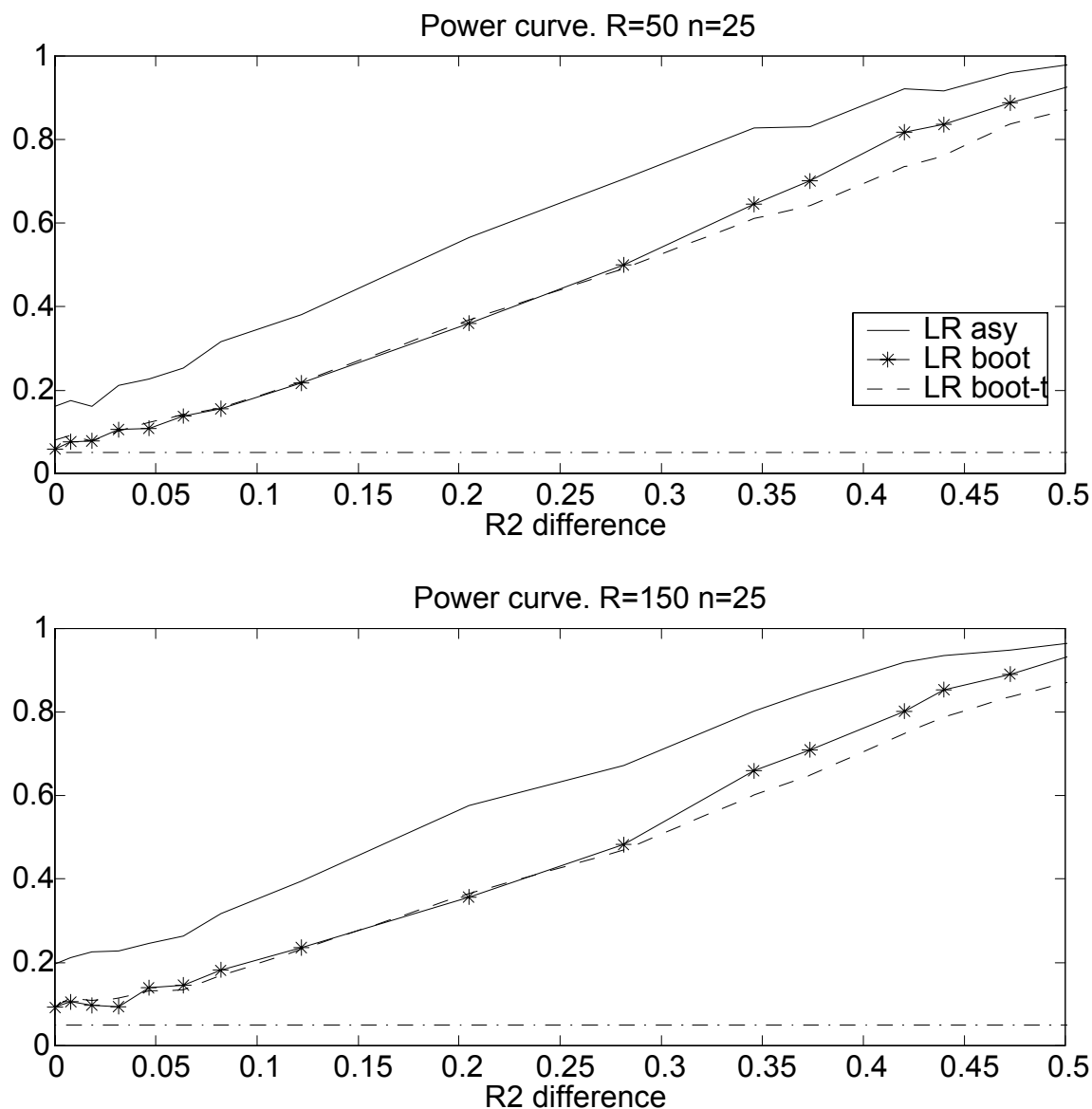


Figure 2: Power curves of LR asymptotic test, LR bootstrap test and LR bootstrap-t test in the Monte Carlo experiment discussed in Section 6. Each curve represents the rejection frequencies over 1000 Monte Carlo replications of the null hypothesis $H_0 : E[\log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*)] = 0$, where the density forecasts f , g and the DGP are defined in (15). The horizontal axis shows the difference in R^2 from the regressions defining the two density forecasts f and g . Both figures consider an out-of-sample size of $n = 25$. The upper panel is for in-sample size $R = 50$ and the lower panel for $R = 150$.

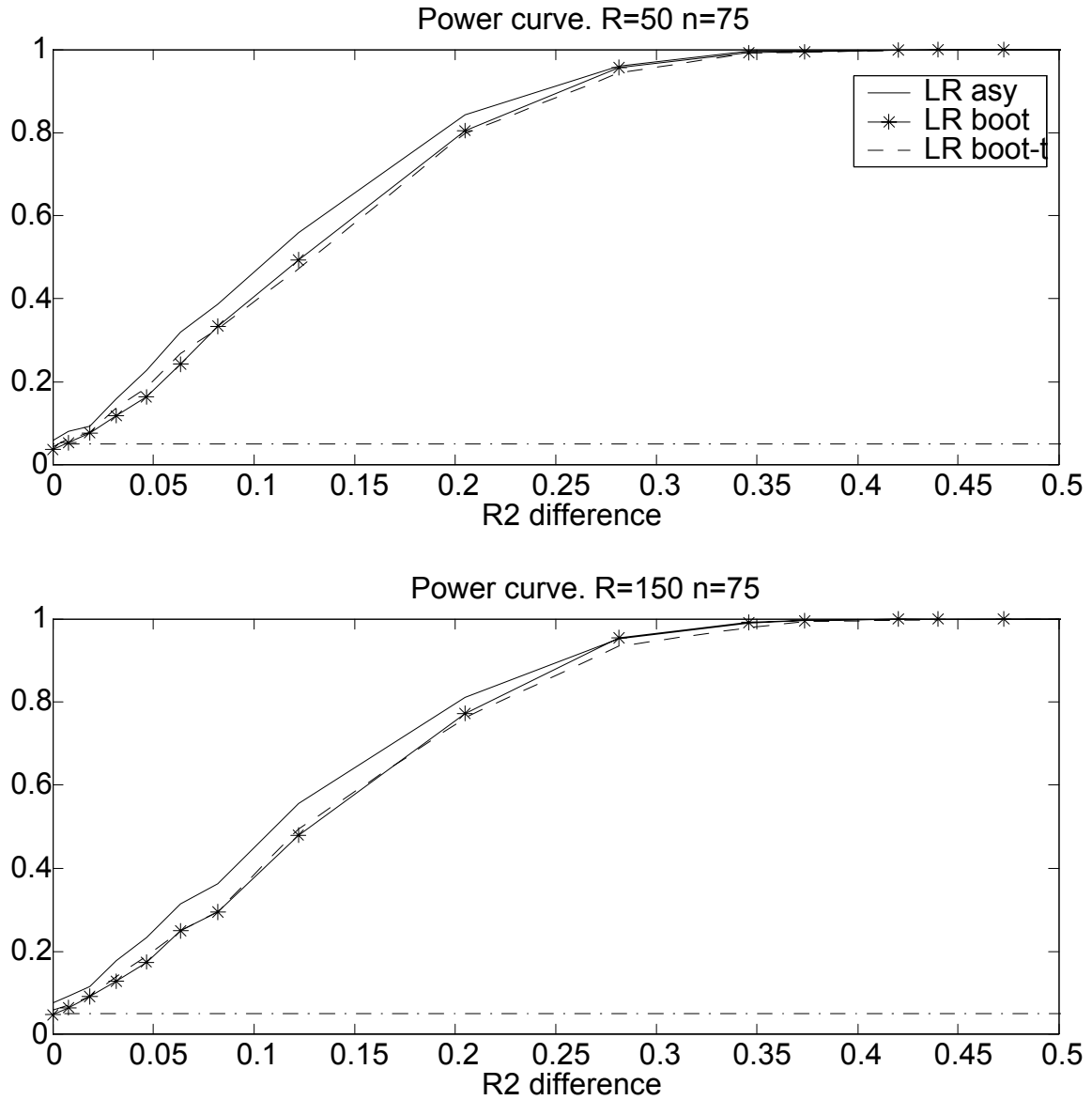


Figure 3: Power curves of LR asymptotic test, LR bootstrap test and LR bootstrap-t test in the Monte Carlo experiment discussed in Section 6. Each curve represents the rejection frequencies over 1000 Monte Carlo replications of the null hypothesis $H_0 : E[\log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*)] = 0$, where the density forecasts f , g and the DGP are defined in (15). The horizontal axis shows the difference in R^2 from the regressions defining the two density forecasts f and g . Both figures consider an out-of-sample size of $n = 75$. The upper panel is for in-sample size $R = 50$ and the lower panel for $R = 150$.

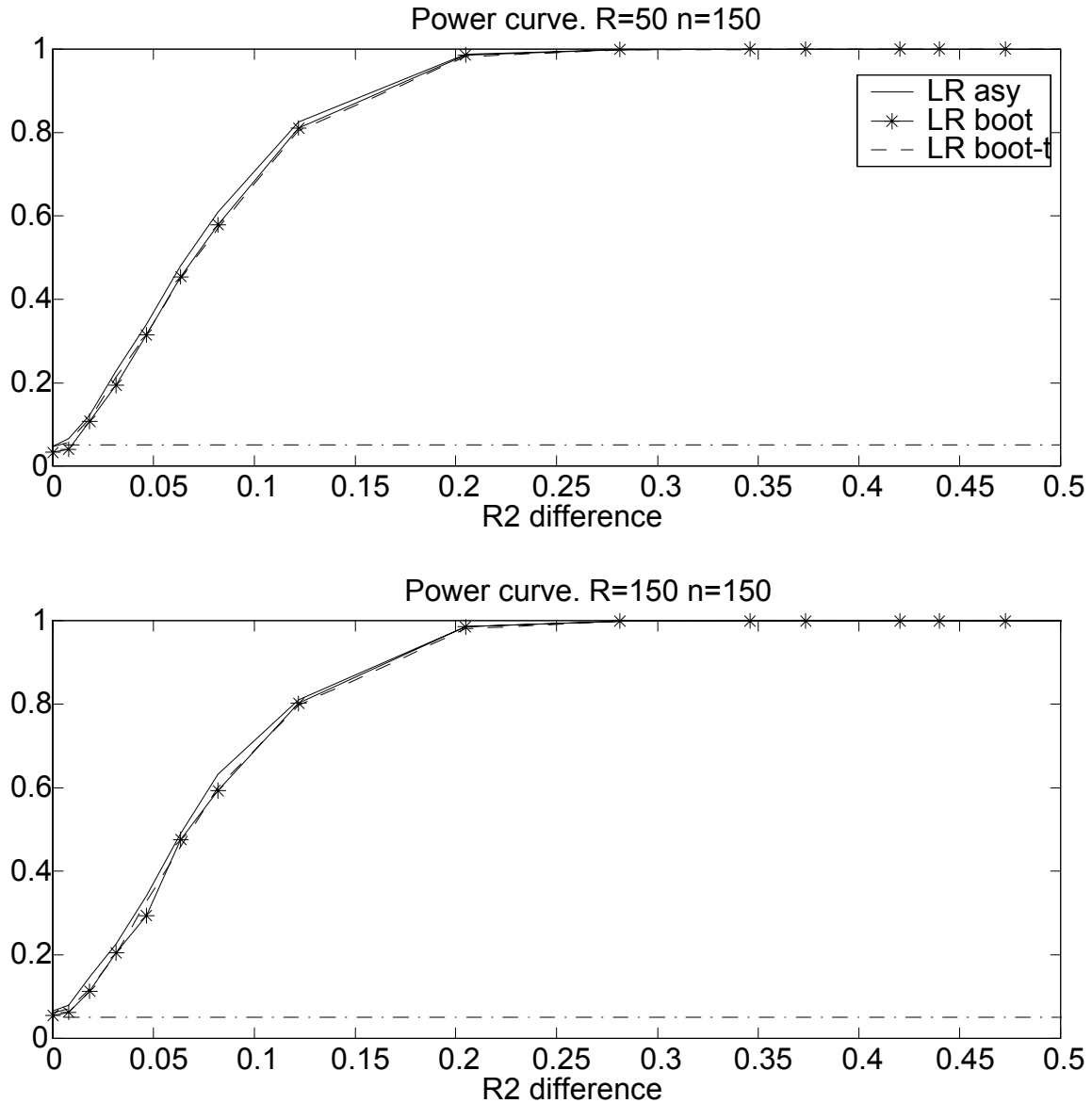


Figure 4: Power curves of LR asymptotic test, LR bootstrap test and LR bootstrap-t test in the Monte Carlo experiment discussed in Section 6. Each curve represents the rejection frequencies over 1000 Monte Carlo replications of the null hypothesis $H_0 : E[\log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*)] = 0$, where the density forecasts f , g and the DGP are defined in (15). The horizontal axis shows the difference in R^2 from the regressions defining the two density forecasts f and g . Both figures consider an out-of-sample size of $n = 150$. The upper panel is for in-sample size $R = 50$ and the lower panel for $R = 150$.

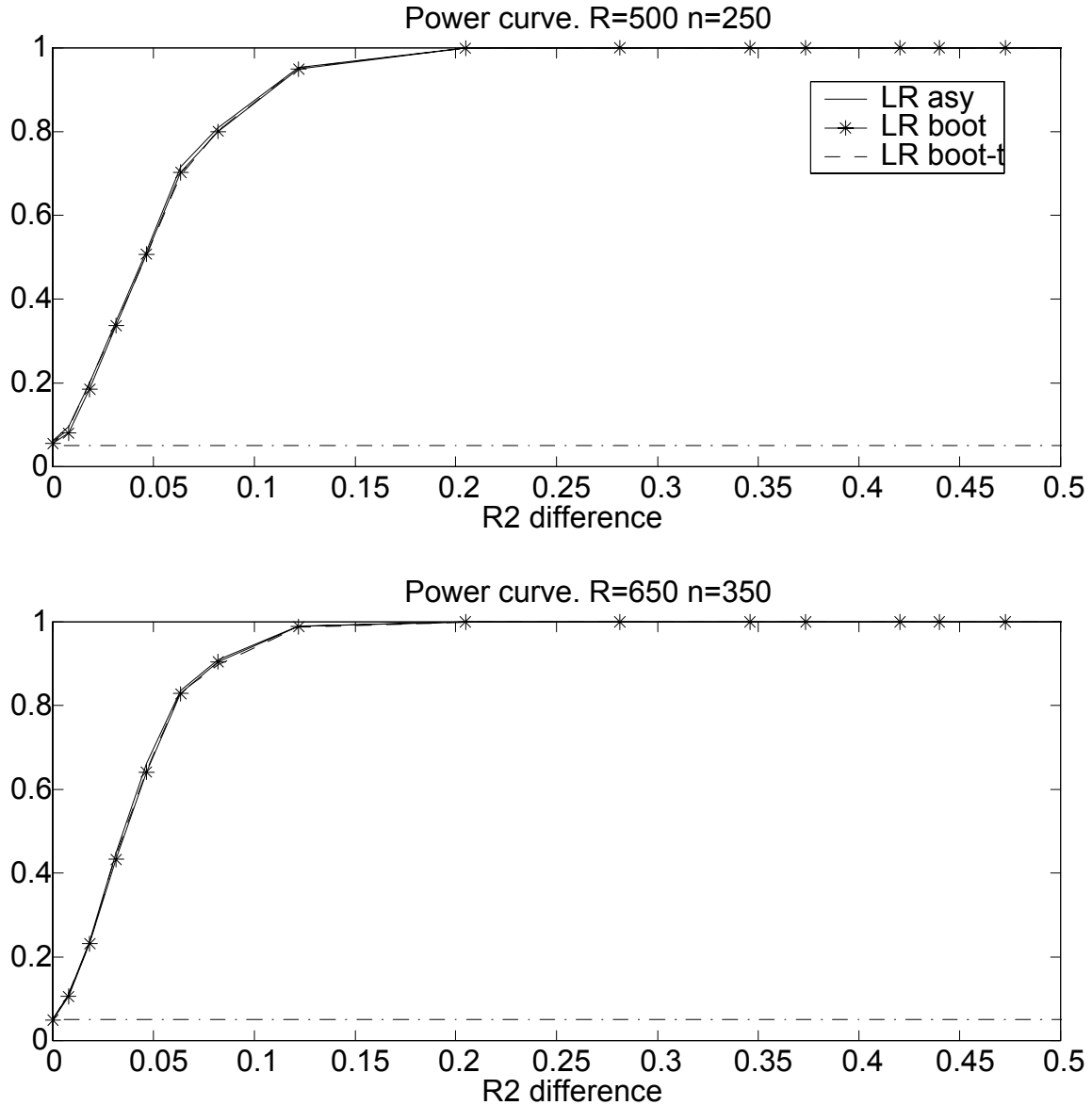


Figure 5: Power curves of LR asymptotic test, LR bootstrap test and LR bootstrap-t test in the Monte Carlo experiment discussed in Section 6. Each curve represents the rejection frequencies over 1000 Monte Carlo replications of the null hypothesis $H_0 : E[\log g(X_{t+1}|\Omega_t; \gamma^*) - \log f(X_{t+1}|\Omega_t; \theta^*)] = 0$, where the density forecasts f , g and the DGP are defined in (15). The horizontal axis shows the difference in R^2 from the regressions defining the two density forecasts f and g . The upper panel is for the pair of in-sample and out-of-sample sizes $R = 500$ and $n = 250$ and the lower panel is for $R = 650$ and $n = 350$.

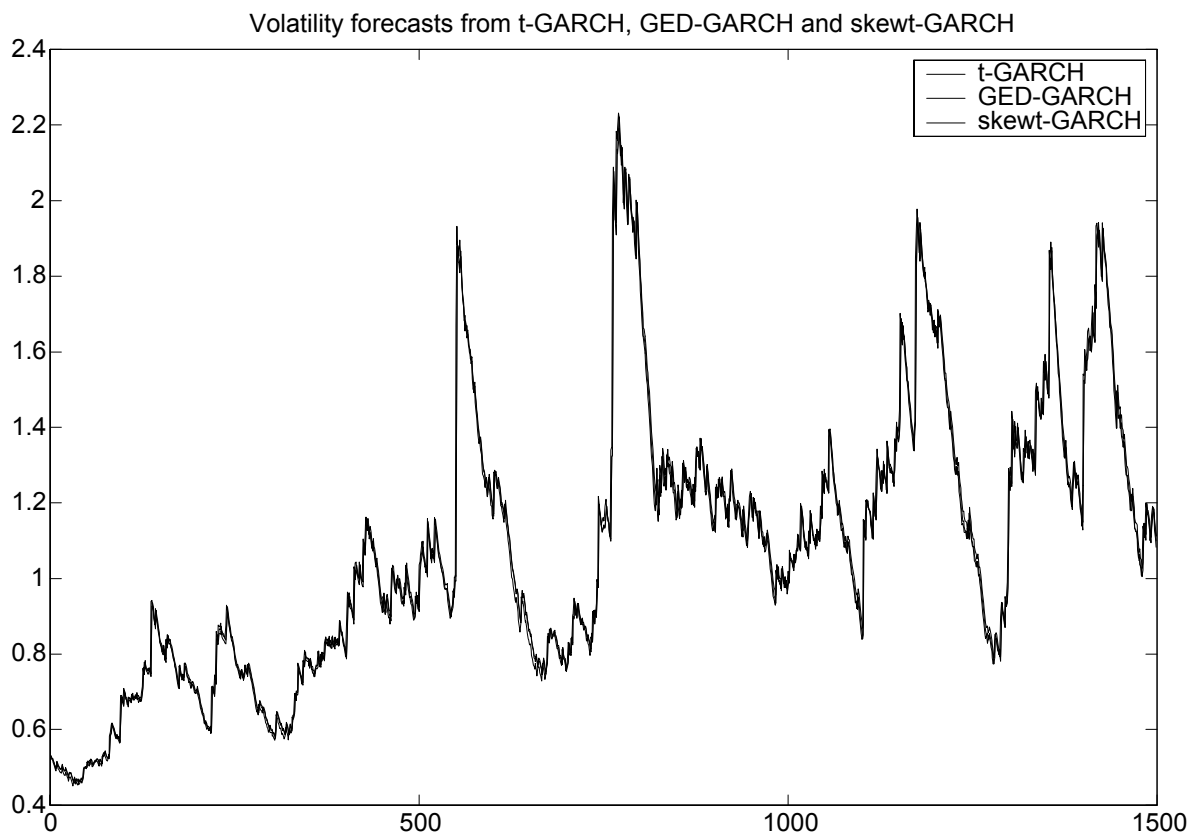


Figure 6: One-step-ahead forecasts of the daily volatility of the S&P500 returns implied by recursively estimated t -GARCH(1,1), GED -GARCH(1,1) and $skewt$ -GARCH(1,1) models.

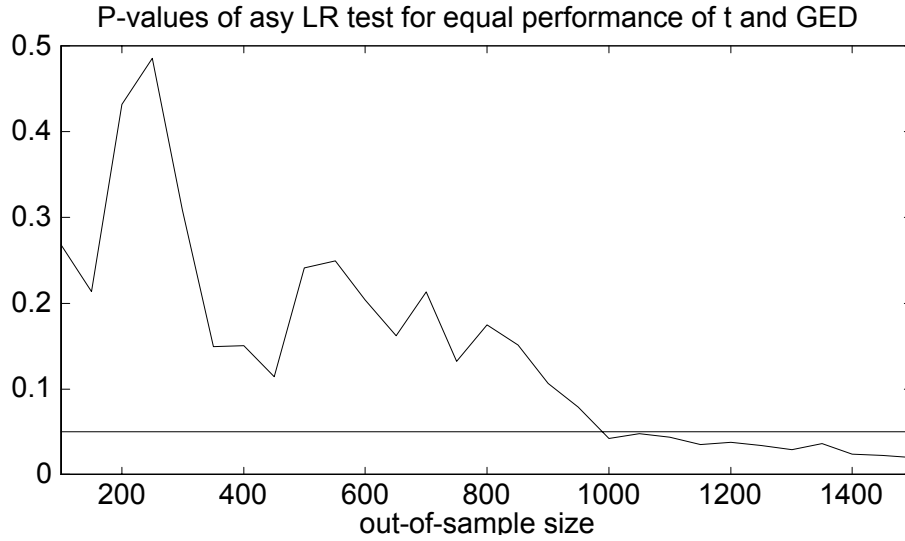


Figure 7: The figure shows p-values for the asymptotic LR test of the null hypothesis of equal performance of t -GARCH and GED -GARCH density forecasts. The horizontal axis indicates increasing out-of-sample size.

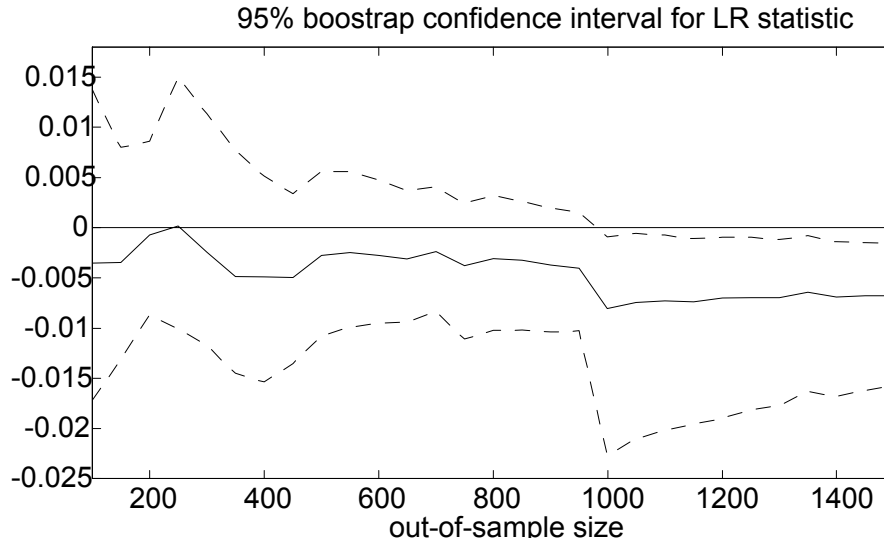


Figure 8: The solid line in the figure indicates the value of the test statistic LR_n , representing the out-of-sample mean of the likelihood ratio between a GED -GARCH density forecast and a t -GARCH forecast. A negative value indicates that t -GARCH outperforms GED -GARCH. The horizontal axis represents increasing out-of-sample size. The dashed line is the 95% bootstrap- t confidence interval for LR_n .

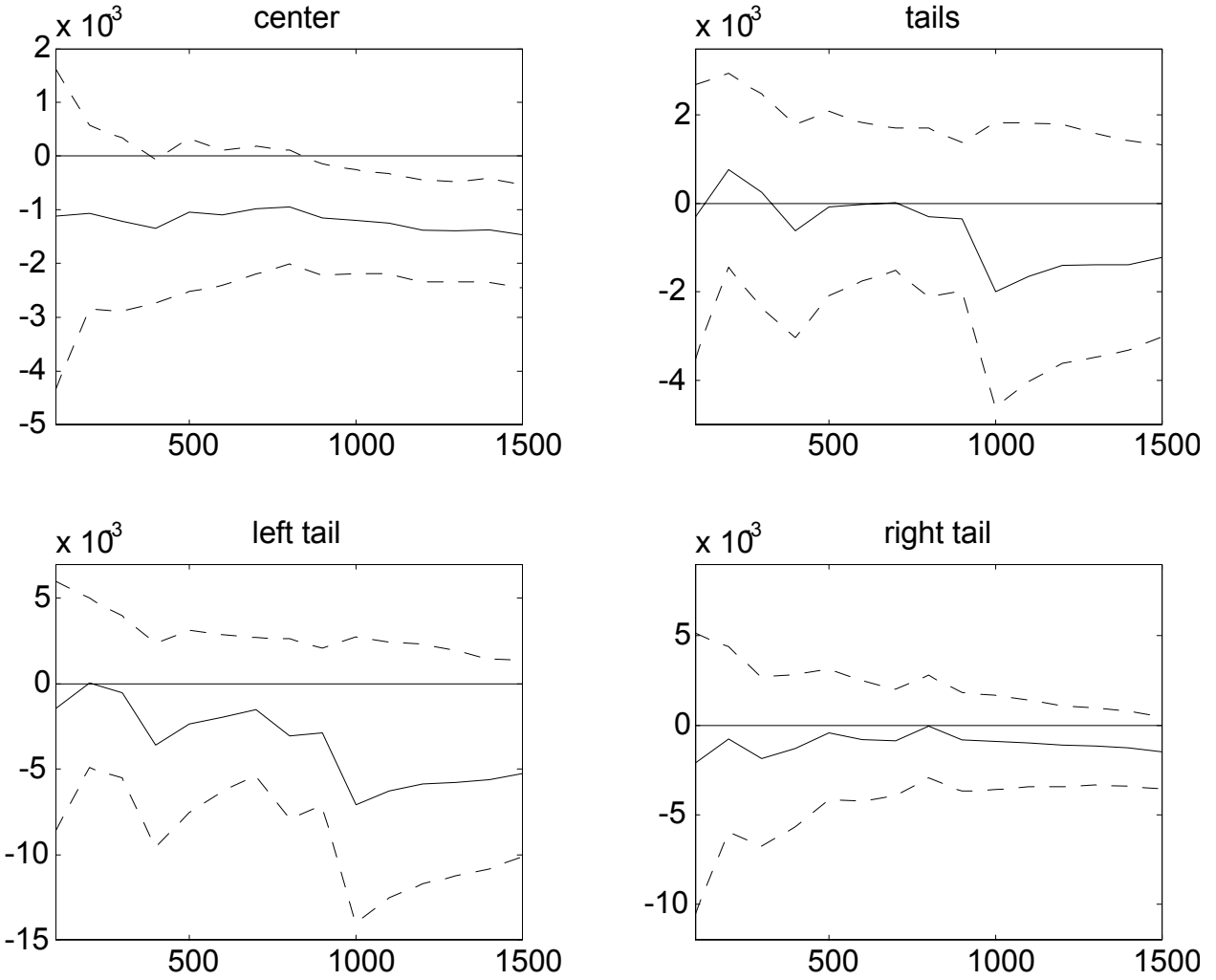


Figure 9: Comparing accuracy of t -GARCH and GED -GARCH density forecasts. The figure shows 95% bootstrap confidence intervals for the weighted likelihood ratio statistics WLR_n for increasing out-of-sample size. A negative value of the statistic indicates that t -GARCH outperforms GED -GARCH. The panels represent the four different weight functions shown in Figure 6; clockwise, the weight function used are w_1 , w_2 , w_3 and w_4 .

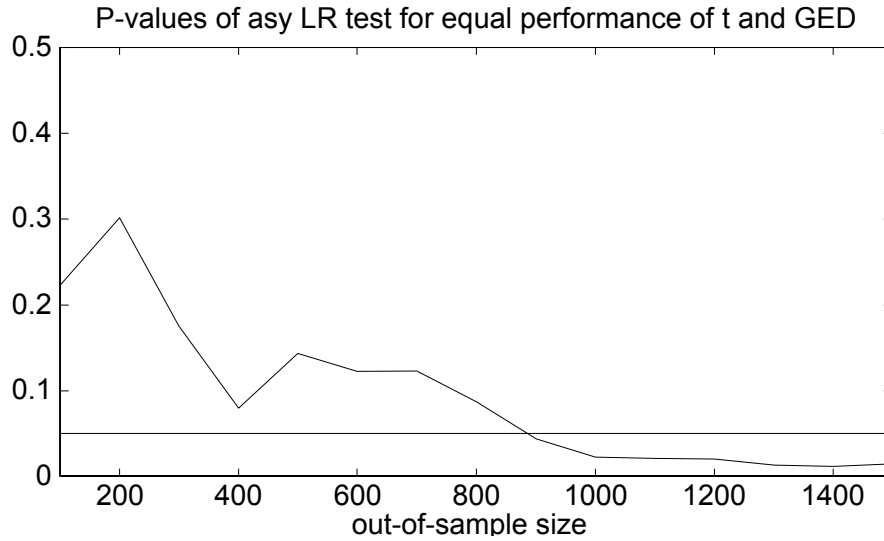


Figure 10: The figure shows p-values for the asymptotic LR test of the null hypothesis of equal performance of *skewt*-GARCH and *GED*-GARCH density forecasts. The horizontal axis indicates increasing out-of-sample size.

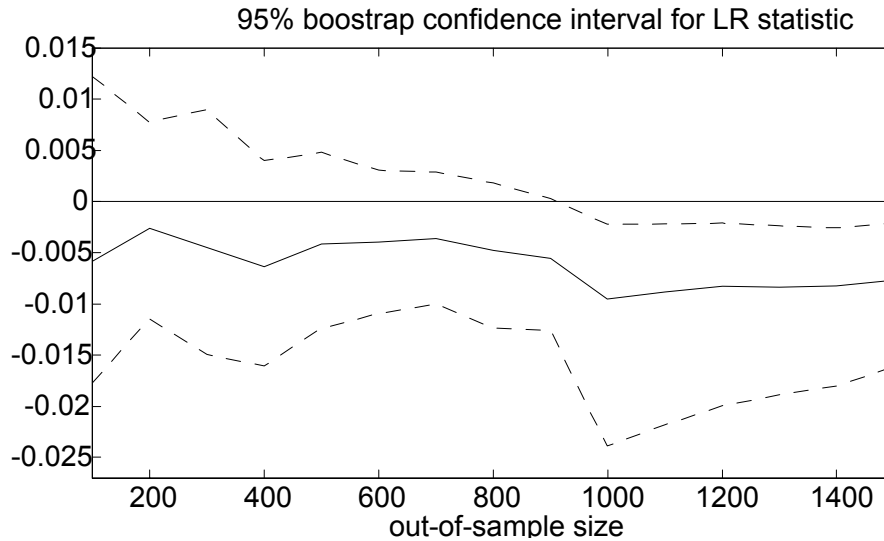


Figure 11: The solid line in the figure indicates the value of the test statistic LR_n , representing the out-of-sample mean of the likelihood ratio between a *GED*-GARCH density forecast and a *skewt*-GARCH forecast. A negative value indicates that *skewt*-GARCH outperforms *GED*-GARCH. The horizontal axis represents increasing out-of-sample size. The dashed line is the 95% bootstrap-*t* confidence interval for LR_n .

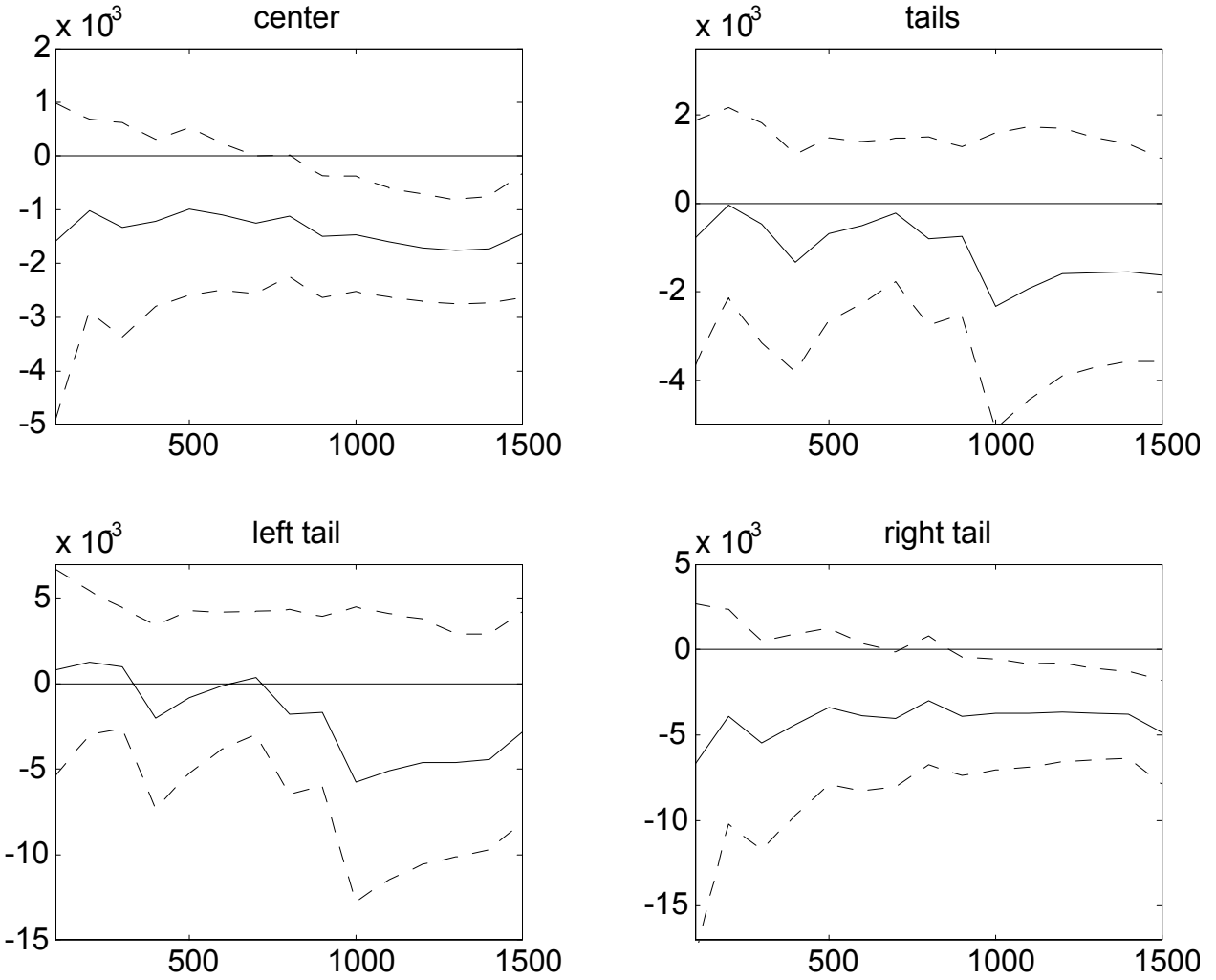


Figure 12: Comparing accuracy of *skewt*-GARCH and *GED*-GARCH. The figure shows 95% bootstrap confidence intervals for the weighted likelihood ratio statistics WLR_n for increasing out-of-sample size. A negative value of the statistic indicates that *skewt*-GARCH outperforms *GED*-GARCH. The panels represent the four different weight functions shown in Figure 6; clockwise, the weight function used are w_1 , w_2 , w_3 and w_4 .