Finite-sample bootstrap inference in GARCH models with heavy-tailed innovations

Richard Luger*

Georgia State University, Atlanta, GA, USA

August 9, 2011

Abstract

A general method is proposed for the construction of valid simultaneous confidence sets in the context of stationary GARCH models. The proposed method proceeds by numerically inverting the conventional likelihood ratio test. In order to hedge against the risk of a spurious rejection, candidate points that are rejected by the conventional test undergo a finite-sample parametric bootstrap test. A projection technique is then exploited to produce conservative confidence sets for general functions of the parameters. A simulation study illustrates the performance of the parametric bootstrap approach in the context of a GARCH model with heavy-tailed and skewed innovations. That model is then used in an empirical application to construct simultaneous confidence intervals for multi-step ahead volatility forecasts for the returns on a major stock market index.

Keywords: Conditional heteroskedasticity; Heavy tails; Weak identification; Parametric bootstrap; Monte Carlo p-value; Projection technique.

* *Corresponding address:* Department of Risk Management and Insurance, Georgia State University, P.O. Box 4036, Atlanta, GA 30302-4036, United States. Tel.: +1 404 413 7489; fax: +1 404 413 7499. *E-mail address:* rluger@gsu.edu (R. Luger).

1 Introduction

The autoregressive conditional hetroskedasticity (ARCH) model of Engle (1982) and the generalized ARCH (GARCH) model of Bollerslev (1986) specify explicitly how conditional variances evolve over time. The motivation for such specifications stems from the increasingly important need to model and forecast the risk in financial time series, for example. The GARCH class of models prove to be extremely useful in practice because they are relatively easy to estimate compared to, say, stochastic volatility models and they can account for many well-known empirical features such as the volatility clustering and excess kurtosis of financial returns. Furthermore, it is straightforward to generate volatility forecasts from these models. Such forecasts are obviously very important in financial risk management.

To set the stage, consider a financial return y_t that evolves over time according to the GARCH(1,1) model:

$$y_t = \mu + \sigma_t \varepsilon_t, \tag{1}$$

$$\sigma_t^2 = \omega + a_1 (y_{t-1} - \mu)^2 + b_1 \sigma_{t-1}^2, \tag{2}$$

where $\omega > 0$, $a_1 \geq 0$, and $b_1 \geq 0$, and $\{\varepsilon_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and unit variance. The model in (1) posits a constant expected return each period equal to μ. The conditional variance of y_t given its past, however, changes over time according to the recursion in (2). When $b_1 = 0$, the conditional variance specification becomes an ARCH(1) model. Lee and Hansen (1994) refer to the GARCH(1,1) model as the "workhorse of the industry," which speaks to the popularity of that specification. A reason for this popularity can be found in Hansen and Lunde (2005), where 330 volatility models are compared in terms of their ability to forecast the conditional variance in an out-of-sample setting using six different loss functions. The main conclusion is that they find no evidence that the GRACH(1,1) model is outperformed by the host of other models.

The estimation of GARCH models is usually performed by the method of maximum

likelihood once a distribution for the the innovation term ε_t in (1) is specified. The most basic specification is to assume that the innovation terms are standard normal. But often times a GARCH model with normally distributed innovations is not sufficient to capture the heavy-tail behavior of financial returns. It is for that reason that non-normal innovation distributions are routinely used. These include the Student-t distribution, the double exponential distribution, and the generalized error distribution among others; see Bollerslev (1987), Nelson (1991), and the references in Zivot (2009). More evidence showing that financial returns can be very heavy-tailed is found in Mittnik, Rachev, and Paolella (1998) and Mittnik and Rachev (2000).

An important appeal of the method of maximum likelihood is that the resulting estimators have well-known optimality properties under regularity conditions. Crowder (1976) gives one such set of regularity conditions for the maximum likelihood estimates in models with dependent observations to be consistent and asymptotically normally distributed. Another approach called quasi- (or pseudo-) maximum likelihood proceeds by maximizing the normal likelihood function, even though the true innovation distribution may not be normal. Gouriéroux, Monfort, and Trognon (1984) and Bollerslev and Wooldridge (1992) show that the resulting estimators are nevertheless consistent, again under certain regularity conditions. The consistency of the quasi-maximum likelihood estimator of the parameters of GARCH models is shown in Weiss (1986), Lee and Hansen (1994), Lumsdaine (1996), Berkes, Horváth, and Kokoszka (2003), Jensen and Rahbek (2004a, 2004b), and Francq and Zakoïan (2004, 2007). In each of those cases, however, the proofs crucially depend on the assumption that the innovation distribution has a finite moment beyond order two.

When the true innovations are relatively heavy-tailed, Hall and Yao (2003) show that the asymptotic distribution of the quasi-maximum likelihood estimator in GARCH models is a multivariate stable distribution, which is particularly difficult to estimate directly using standard parametric methods. As Hall and Yao emphasize, standard bootstrap methods are generally not consistent in settings where the asymptotic distribution of a statistic is non-normal; see also Mammen (1992), Athreya (1987a,b), Knight (1989), and

Hall (1990). For that reason, they suggest a subsample bootstrap procedure that yields asymptotic confidence sets for the GARCH parameters. In their simulations, however, Hall and Yao report problems with the coverage accuracy of their marginal confidence intervals for the GARCH(1,1) parameters even when the sample size is as large as 1000. Those problems are further exacerbated when the true innovation distribution is skewed.

In this paper, a simulation-based procedure is proposed for the construction of valid finite-sample confidence sets in the context of stationary GARCH models. In fact, the model is the same as in Hall and Yao (2003) who also maintain the assumption of a stationary GARCH process. An important difference via-à-vis the *quasi*-maximum likelihood approach of Hall and Yao, however, is that here the method is one of maximum likelihood where the assumed innovation distribution is part of the maintained model specification. The distribution of the innovation terms in the GARCH model may be any parametric distribution (skewed or not) of the user's choice without any requirements on the finiteness of its moments beyond order two. The central assumption exploited here is that the model can be simulated once the parameters have been specified. The proposed method uses a parametric bootstrap procedure based on the likelihood ratio statistic to obtain finite-sample tests of any hypothesis that completely sets the vector of model parameters to specified values. Closely related procedures are proposed by Dufour and Jouini (2006) for vector autoregressive models and by Dufour and Valéry (2009) for Gaussian stochastic volatility models.

The exactness of the bootstrap test proposed here rests on the assumption that the simulated samples are in their stationary distribution. The problem of obtaining stationary drawings depends on the length of the burn-in sample, which is under the user's control. This is a well-known problem in the literature on Markov chain Monte Carlo methods for which there exists many different diagnostics to check convergence to stationarity; Robert and Casella (2010) provide a pragmatic introduction. In the spirit of how the number of replications in bootstrap procedures are typically chosen, the user may increase the size of the burn-in sample until the inference results do not change or simply set it to an arbitrarily large value. The latter is the approach adopted in the reported

simulation experiments and the empirical application. A key difference though with other (asymptotically justified) bootstrap procedures is that, if the simulated samples are in their stationary distribution, exact p-values can be obtained even with a very small number of bootstrap replications. This is achieved through the technique of Monte Carlo tests, originally proposed by Dwass (1957) and extended by Dufour (2006); see Dufour and Khalaf (2001) for a comprehensive review of Monte Carlo test techniques.

Another remarkable feature is that the proposed bootstrap procedure remains valid whether the model is identified or not. For instance, the true value of a_1 in (2) may be zero in which case the data are homoskedastic and b_1 is not statistically identified. This stands in sharp contrast to standard asymptotic methods that assume that the true parameter values lie in the interior of the parameter space. Indeed when some parameters are not identified under the null hypothesis, the results of Davies (1977, 1987) show that the asymptotic distribution of the likelihood ratio statistic is non-standard and quite complicated.

The proposed method constructs simultaneous confidence sets for the parameters of the GARCH model by numerically "inverting" the likelihood ratio test. This involves collecting the parameter values from a grid of admissible candidate values that are not rejected by the test. This method to confidence set construction is obviously computer intensive. The computational cost would be dramatically reduced if the confidence sets were derived from the usual likelihood ratio test since its (asymptotic) critical value need not be simulated. But then of course the probability of a Type I error—a spurious rejection of the null hypothesis—may not be under control, especially in small samples. Given those considerations, a hybrid algorithm is proposed where the exact parametric bootstrap test is performed only for candidate points that are rejected by the conventional likelihood ratio test. This ensures that none of the considered points are spuriously rejected. The results of a simulation study show that not much power is lost with the hybrid approach. Therefore the confidence sets derived in this manner are only slightly more conservative than those that would be obtained by inverting only the asymptotic likelihood ratio test, if it always had the correct size. In return for the computational effort, the proposed

algorithm yields confidence sets that satisfy the nominal coverage constraint for any choice of the innovation distribution no matter the true parameter values nor the actual sample size.

The paper is organized as follows. Section 2 describes the proposed simulation-based inference method. It begins by presenting the statistical framework and the maintained assumptions. It then presents the parametric bootstrap procedure based on the likelihood ratio statistic that delivers exact tests of any hypothesis that completely sets the vector of model parameters to specified values. Section 2 then moves on to explain the hybrid approach to confidence set construction for the model parameters and how to derive projection-based confidence sets for general functions of the parameters and the observed data sample, following the approach of Abdelkhalek and Dufour (1998). Section 3 presents the results of the simulation experiments that compare the size and power behavior of the bootstrap approach with that of the likelihood ratio test based on the usual asymptotic distribution. The simulation experiments are conducted in the context of a GARCH model with heavy-tailed and skewed innovations. That model is then used in Section 4 to construct simultaneous confidence intervals for multi-step ahead volatility forecasts for monthly, weekly, and daily returns on the Standard & Poor's 500 stock market index. Section 5 concludes.

2 Simulation-based inference

2.1 Statistical framework

Consider a random variable of interest y_t that follows a GARCH(p,q) process, where $p \geq 0$ and $q \geq 0$.

Assumption 1. The process $\{y_t, -\infty < t < \infty\}$ follows a GARCH model of the form:

$$y_t = \mu + \sigma_t \varepsilon_t, \tag{3}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} a_i (y_{t-i} - \mu)^2 + \sum_{j=1}^{p} b_j \sigma_{t-j}^2, \tag{4}$$

where $\omega > 0$, $a_i \geq 0$, and $b_j \geq 0$ are unknown parameters, and the process starts indefinitely far in the past.

The parameters of the model in (3) and (4) for y_t are collected in the vector $\vartheta = (\mu, \omega, a_1, ..., a_q, b_1, ..., b_p)'$. Let $\mathcal{I}_{t-1} = \{y_{t-i}, i \geq 1\}$ denote the past of y_t. Assumption 2 below imposes restrictions on the innovation term ε_t in (3) which will imply that $E[y_t|\mathcal{I}_{t-1}] = \mu$ and $E[y_t^2|\mathcal{I}_{t-1}] = \sigma_t^2$. The conditional expectation of y_t is assumed constant for simplicity. So with the next assumption, the model in (3) specifies a time-varying conditional variance for y_t, which evolves over time according to the GARCH recursion in (4).

Assumption 2. The random variables ε_t, for $-\infty < t < \infty$, are i.i.d. according to a known continuous distribution with zero mean and unit variance. Furthermore, ε_t is independent of $\mathcal{I}_{t-1}$.

When $p = 0$ the process described by (3) and (4) reduces to an ARCH(q) process, and when $p = q = 0$ the y_t's are i.i.d. with mean μ and variance equal to ω. Note that the present framework allows $q = 0$. In that case, the process governing y_t is homoskedastic and the parameters $b_1, ..., b_p$ are not statistically identifiable. To see this, consider the GARCH(1,1) version of (3) and (4). When $a_1 = 0$ that model can be written as $\sigma_t^2 = \delta + b_1(\sigma_{t-1}^2 - \delta)$, where $\delta = \omega/(1 - b_1)$. So if the recursion is started with $\sigma_1^2 = \delta$, then $\sigma_t^2 = \delta$ for all subsequent t; i.e., b_1 is not identified. Ma, Nelson, and Startz (2007) show that the usual t-test has massive size distortions in non- or weakly identified GARCH(1,1) models. More generally, Dufour (1997) shows that Wald-type confidence sets based on asymptotic standard errors can arbitrarily deviate from their nominal level

in models with parameters which are not identifiable on certain subsets of the parameter space.

Let $g(\varepsilon_t;\eta)$ denote the known density function of ε_t in Assumption 2, which may depend on some unknown parameters collected in η. Regroup all the model parameters as $\theta = (\vartheta', \eta')'$ and write the conditional density of y_t given $\mathcal{I}_{t-1}$ as

$$f(y_t|\mathcal{I}_{t-1};\theta) = \frac{1}{\sigma_t(\theta)} \times g\left(\frac{y_t - \mu}{\sigma_t(\theta)};\eta\right), \tag{5}$$

where the skedastic function $\sigma_t^2(\theta)$ is defined by (4). The first term on the right-hand side of (5) is a Jacobian factor which reflects the fact that the derivative of ε_t with respect to y_t is $1/\sigma_t(\theta)$.

A very common specification for the innovation terms is to assume that they are normal with mean zero and unit variance. Since the normal distribution is completely determined by those two moment parameters, η is empty in that case. It is well known that a GARCH model with conditionally normal innovations results in a heavy-tailed unconditional distribution. However, the degree of kurtosis induced by the time-varying conditional variance often does not capture all the excess kurtosis present in financial returns. To circumvent that problem, Bollerslev (1987) suggests a variant where ε_t follows a standardized Student-t distribution with v degrees of freedom, in which case $v > 2$ is treated as the only unknown parameter appearing in η. That specification allows the innovation terms to have heavier tails and lets the data determine just how heavy they are. As already mentioned in the introduction, several other conditional distributions have been considered in the literature to also capture skewness in addition to excess kurtosis in the distribution of returns. The simulation study and the empirical illustration presented below consider a generalization of the Student-t distribution with an additional shape parameter that imparts asymmetry to the distribution of innovations.

Assumption 3. The process $\{y_t, -\infty < t < \infty\}$ is strictly stationary.

Bollerslev (1986), Nelson (1990), and Bougerol and Picard (1992) show that a sufficient

condition for Assumption 3 to hold is

$$\sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1. \tag{6}$$

The necessary and sufficient condition for strict stationarity in fact allows the GARCH process to be integrated. In the GARCH(1,1) case for instance, that condition is $E[\log(a_1\varepsilon_t^2+b_1)] < 0$ which does not rule out $a_1+b_1 = 1$ (IGARCH). However, as Christoffersen (2003, p. 24) emphasizes, the integrated model yields counterfactual longer-horizon forecasts. The condition in (6) ensures that the GARCH forecasts will eventually revert to their long-run mean value; see the empirical results in Section 4 below. The admissible region of the parameter space for ϑ satisfying the stationarity restriction in (6) and the non-negativity restrictions in Assumption 1, and the set of admissible values for η shall be denoted by Θ.

Since Assumption 2 fully specifies the distribution of the innovations, it is quite natural to estimate the model parameters by maximizing the sample likelihood function. Let $Y = (y_1, ..., y_T)'$ represent the observed sample, and denote by $m = \max(p, q)$ the number of observations lost for the initialization of the recursion in (4). The sample log-likelihood function given the initial values can then be written as

$$\mathcal{L}(\theta; Y) = \sum_{t=m+1}^{T} \ell_t(\theta), \tag{7}$$

where

$$\ell_t(\theta) = -\log(\sigma_t(\theta)) + \log\left(g\left(\frac{y_t - \mu}{\sigma_t(\theta)}; \eta\right)\right)$$

is the contribution to the sample log-likelihood made by observation t.

Assumption 4. The GARCH model in (3) and (4) is estimated by the method of maximum likelihood so that the the unconstrained maximum likelihood estimate (MLE), $\hat{\theta}$, satisfies

$$\hat{\theta}(Y) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathcal{L}(\theta; Y),$$

where $\mathcal{L}(\theta; Y)$ is the sample log-likelihood function in (7). Furthermore, Θ is assumed to be compact.

The restriction that Θ is a compact (i.e. closed and bounded) set needs to be imposed at the estimation stage. For instance, if the elements of θ are generically denoted by $\theta_1, ..., \theta_n$, then Θ may be some hypercube of the form

$$\Theta = [\underline{\theta}_1, \overline{\theta}_1] \times [\underline{\theta}_2, \overline{\theta}_2] \times ... \times [\underline{\theta}_n, \overline{\theta}_n],$$

where $\underline{\theta}_i$ and $\overline{\theta}_i$ are the lower and upper bounds, respectively, for the ith parameter. The compactness of Θ serves to bound the grid used in the construction of the proposed confidence sets. Note that the MLE need not lie in the interior of Θ. Indeed, all that is required by Assumption 4 is that $\hat{\theta}$ be a (local or global) maximum of the sample log-likelihood over the compact set Θ. The MLE must be a stationary solution but it may occur on the no-ARCH boundary of Θ, as it would when the data are homoskedastic ($a_1 = ... = a_q = 0$).

Two common ways of initializing the values of σ_t^2 in the GARCH recursion in (4) are as follows. The first is to use the sample variance of y_t. The second is to use the model-implied unconditional variance. In the GARCH(1,1) case for instance, the unconditional variance is given by $\sigma^2 = \omega/(1-a_1-b_1)$, which only makes sense when the parameters are restricted to belong to Θ as in Assumption 4. Either one of those GARCH initialization schemes may be used here.

2.2 Parametric bootstrap test

This section presents the bootstrap test used in the construction of confidence sets for the parameters of the GARCH model. Recall that a confidence set for a parameter vector can be interpreted as the result of a collection of tests for each possible value of the parameter vector. The confidence set simply reports all the values of the parameter vector which cannot be rejected at a given nominal level; see Theorem 3.5.1 in Lehmann and Romano (2005). Specifically, the $(1-\alpha)$-level confidence set $C_{1-\alpha}(\theta, Y)$ for the parameter vector θ is the result from tests of the following null hypothesis that completely sets the vector of parameters θ to a specified value, θ_0:

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \neq \theta_0, \tag{8}$$

where $\theta_0 \in \Theta$.

An appeal of the method of maximum likelihood is the well-known optimality of the resulting estimates under certain regularity conditions. Crowder (1976) gives one set of sufficient regularity conditions for the MLE in models with dependent observations to be consistent and asymptotically normally distributed. As Engle, Bollerslev, and Nelson (1994) point out, a common practice in empirical work is to proceed under the assumption that the necessary regularity conditions are satisfied. When those conditions actually hold, the conventional likelihood ratio (LR) statistic

$$LR(\theta_0, Y) = 2\big(\mathcal{L}(\hat{\theta}(Y); Y) - \mathcal{L}(\theta_0; Y)\big), \tag{9}$$

where $\mathcal{L}(\theta; Y)$ is the log-likelihood of Y evaluated at θ, asymptotically follows a chi-square distribution with degrees of freedom equal to the number of parameters in θ, if H_0 is true.

Dufour and Torrès (2003) make a nice connection between maximum likelihood estimators and Hodges-Lehmann estimators, where the latter approach derives a point estimator by narrowing a confidence set estimator; see Hodges and Lehmann (1963, 1983). To see this connection, consider the test at level α of H_0 that rejects the null hypothesis if $LR(\theta_0) > c_\alpha$, where c_α is the critical value for an α-level test. The associated confidence set at level $1-\alpha$ is given by $C_{1-\alpha}(\theta, Y) = \{\theta_0 \in \Theta : LR(\theta_0, Y) \leq c_\alpha\}$. That confidence set is monotonic in the sense that if $\alpha_1 > \alpha_2$, then $C_{1-\alpha_1}(\theta, Y) \subseteq C_{1-\alpha_2}(\theta, Y)$ with probability 1. Since (9) can only be non-negative and $LR(\hat{\theta}, Y) \equiv 0$, the MLE $\hat{\theta}$ cannot be rejected at any level $\alpha \in (0, 1]$. This means that $C_{1-\alpha}(\theta, Y)$ can never be an empty set, since it always contains at least the MLE $\hat{\theta}$. Moreover, if $\mathcal{L}(\theta; Y)$ has a unique maximum over Θ, then the MLE $\hat{\theta}$ is the only value that cannot be rejected at any level; i.e., it is the unique Hodges-Lehmann point estimate of θ. Related test inversion methods are studied in Bolduc, Khalaf, and Yélou (2010) and Dufour, Khalaf, and Kichian (2010).

A potential concern when using the likelihood ratio statistic is that its actual finite-sample distribution may not be well approximated by the asymptotic chi-square distribution. The concern in this case would be that the asymptotic test over-rejects and the derived confidence set fails to cover relevant regions of the parameter space. In order to

hedge that risk of a Type I error, a finite-sample parametric bootstrap test is proposed that exploits the fact that the model can easily be simulated given the value of θ_0 appearing in (8). A closely related procedure is proposed in Dufour and Valéry (2009) for Gaussian stochastic volatility models.

Consider a simulated realization from the GARCH model in (3) and (4) evaluated at $\theta_0 \in \Theta$, say $(\tilde{y}_{-H}, ..., \tilde{y}_0, \tilde{y}_1, ..., \tilde{y}_T)'$, where H is a large number, and let $\widetilde{Y} \equiv \widetilde{Y}(\vartheta_0, \tilde{\varepsilon}(\eta_0)) = (\tilde{y}_1, ..., \tilde{y}_T)'$ represent the last T values so that the size of the retained portion of the simulated sample is the same as the original one. By making the starting value of t far enough in the past, the joint distribution of $\tilde{y}_1$ through $\tilde{y}_T$ can be made arbitrarily close to the stationary distribution. The values of $\tilde{y}_t$ for non-positive t are discarded. In the following it will be assumed that the burn-in size H is sufficiently large so that $\widetilde{Y}$ is in its stationary distribution. The next result is an immediate implication of the stationarity assumption and forms the building block for obtaining finite-sample inferences about θ.

Proposition 1. If $\widetilde{Y}$ is in its stationary distribution, then, under Assumptions 1–4 and the null hypothesis in (8), the likelihood ratio statistic defined in (9) is distributed like

$$LR(\theta_0, \widetilde{Y}) = 2\big(\mathcal{L}(\hat{\theta}(\widetilde{Y}); \widetilde{Y}) - \mathcal{L}(\theta_0; \widetilde{Y})\big),$$

where $\hat{\theta}(\widetilde{Y})$ is the maximum likelihood estimate of θ based on $\widetilde{Y}$.

Note that the distributional result in Proposition 1 is a finite-sample one in that it holds for any sample size, T. Of course, the assumption that $\widetilde{Y}$ is in its stationary distribution becomes more tenable when the burn-in size H is very large. Choosing H here is essentially the same problem that arises in the implementation of Markov chain Monte Carlo algorithms, where the distribution of Gibbs sampled chains should correspond to the stationary target distribution. This is a well-known problem for which there exists many different diagnostics to check convergence to stationarity; see Robert and Casella (2010) for a pragmatic exposition and many references. Alternatively, in the spirit of bootstrap procedures one may simply increase the value of H to a point where the inference results do not change. Yet another possibility is to treat the parameters needed to initialize the

GARCH recursion as additional (nuisance) parameters. The downside to that approach is that it increases the dimension of θ, which in turn increases the computational burden when constructing confidence sets.

For the result in Proposition 1 to hold, it is important that the numerical maximization of $\mathcal{L}(\theta; \widetilde{Y})$ to get $\hat{\theta}(\widetilde{Y})$ be performed in exactly the same manner as was done to get $\hat{\theta}(Y)$ with the original sample. In particular, this means that the only difference is that instead of the original sample Y the simulated sample $\widetilde{Y}$ is used to compute and numerically maximize the log-likelihood function, and that all other aspects of the optimization problem remain unchanged (e.g. numerical optimization method, initial parameter values, parameter bounds, GARCH initialization scheme, step size, stopping rule, etc).

The result in Proposition 1 paves the way for a one-sided test of H_0 in (8) using the technique of Monte Carlo (MC) tests; see Dufour and Khalaf (2001) for a detailed introduction to this technique.

Algorithm 1 (*LR bootstrap test of* H_0). Step 1: Generate B simulated samples $\widetilde{Y}_1, ..., \widetilde{Y}_B$ under H_0. Step 2: Calculate the simulated statistics

$$LR(\theta_0, \widetilde{Y}_i) = 2\big(\mathcal{L}(\hat{\theta}(\widetilde{Y}_i); \widetilde{Y}_i) - \mathcal{L}(\theta_0; \widetilde{Y}_i)\big), \quad i = 1, ..., B,$$

according to Proposition 1. Step 3: Compute the MC p-value as

$$p_B(\theta_0) = \frac{1 + \sum_{i=1}^{B} \mathbb{I}\big[LR(\theta_0, \widetilde{Y}_i) \geq LR(\theta_0, Y)\big]}{B + 1}, \tag{10}$$

where $\mathbb{I}[A]$ is the indicator function of event A.

Proposition 2. If Assumptions 1–4 and the null hypothesis H_0 in (8) hold, and $\widetilde{Y}_1, ..., \widetilde{Y}_B$ are i.i.d. samples in their stationary distribution, then by selecting B such that $\alpha(B+1)$ is an integer, $\Pr[p_B(\theta_0) \leq \alpha] = \alpha$; i.e., the critical region $p_B(\theta_0) \leq \alpha$ has size α.

Upon recognizing that $(LR(\theta_0, Y), LR(\theta_0, \widetilde{Y}_1), ..., LR(\theta_0, \widetilde{Y}_B))$ is a vector of exchangeable random variables with zero probability of ties, the proof of the result in Proposition 2 about the exactness of the MC p-value is immediate from Proposition 2.2 in Dufour

(2006). Unlike standard bootstrap p-values, those in (10) computed according to the MC technique take the sample size and the number of replications explicitly into consideration.

2.3 Computation of confidence sets

The computation of the p-value in (10) can be computationally expensive since a total of B numerical optimizations need to be performed to find $LR(\theta_0, \widetilde{Y}_i)$, $i = 1, ..., B$. So clearly B should be kept as small as possible. For instance, $B = 19$ parametric bootstrap replications are sufficient for a test at the nominal 5% level according to Proposition 2. It is also obvious that inverting the bootstrap test to establish a confidence set for θ can quickly become intractable, since a simulation-based p-value needs to be computed at every point over a grid of candidate values. In contrast, inverting the usual likelihood ratio test in (9) is computationally inexpensive since its (asymptotic) critical value, say c_α^{asy}, need not be simulated. But then of course the risk of a Type I error may not be under control. These considerations suggest the following hybrid approach to decide whether a candidate point $\theta_0 \in \Theta$ gets included in the confidence set $C_{1-\alpha}^{hyb}$.

Algorithm 2 (*Computation of confidence sets*). Step 1: Generate a set of grid points $\{\theta_0^{(1)}, ..., \theta_0^{(M)}\}$, where each $\theta_0^{(i)} \in \Theta$. For $\theta_0^{(i)}$, $i = 1, ..., M$, repeat the following steps. Step 2: If $LR(\theta_0^{(i)}, Y) \leq c_\alpha^{asy}$ then $\theta_0^{(i)}$ is included in $C_{1-\alpha}^{hyb}$. Step 3: Else if $LR(\theta_0^{(i)}, Y) > c_\alpha^{asy}$, then $\theta_0^{(i)}$ is included in $C_{1-\alpha}^{hyb}$ only if $p_B(\theta_0^{(i)}) > \alpha$. Otherwise $\theta_0^{(i)}$ is rejected.

This approach yields computational savings by only performing the parametric bootstrap test for candidate points that are rejected by the conventional likelihood ratio test. The points that pass the conventional likelihood ratio test in Step 2 immediately become part of the confidence set, and those that are rejected in Step 2 but pass the test in Step 3 are also included. So the finite-sample test in Step 3 ensures that none of the considered points are spuriously rejected. This can be seen by noting that a candidate point is rejected only if it is rejected by both the conventional likelihood ratio test in Step 2 *and*

the finite-sample test in Step 3. Under the conditions of Proposition 2, it follows that

$$\begin{aligned}\Pr[\theta_0^{(i)} \text{ rejected} \,|\, \theta = \theta_0^{(i)}] &= \Pr[p_B(\theta_0^{(i)}) \leq \alpha \,|\, \theta = \theta_0^{(i)}] \Pr[LR(\theta_0^{(i)}, Y) > c_\alpha^{asy} \,|\, \theta = \theta_0^{(i)}] \\ &= \alpha \Pr[LR(\theta_0^{(i)}, Y) > c_\alpha^{asy} \,|\, \theta = \theta_0^{(i)}] \\ &\leq \alpha,\end{aligned}$$

since the exactness of the MC p-value in Step 3 is independent of the outcome from Step 2 for all $\theta_0^{(i)} \in \Theta$. So whether the asymptotic LR test based on c_α^{asy} is reliable or not, the test procedure in Algorithm 2 maintains the overall level α.

Let $C_{1-\alpha}^{asy} = \{\theta_0^{(i)} \in \Theta, i = 1, ..., M : LR(\theta_0^{(i)}, Y) \leq c_\alpha^{asy}\}$ and $C_{1-\alpha}^{boot} = \{\theta_0^{(i)} \in \Theta, i = 1, ..., M : p_B(\theta_0^{(i)}) > \alpha\}$ denote the confidence sets derived from the conventional and bootstrap LR tests, respectively. Both these sets are finite and hence compact. Note that a candidate point is included in $C_{1-\alpha}^{hyb}$ according to Algorithm 2 if it passes the conventional LR test *or* its bootstrap version, so that

$$C_{1-\alpha}^{hyb} = C_{1-\alpha}^{asy} \cup C_{1-\alpha}^{boot}, \tag{11}$$

which is also a compact set. Given the duality between confidence sets and hypothesis testing, the relation in (11) makes clear that the trade-off with the hybrid approach is that it can entail a power loss. However, the simulation evidence presented in Section 3 shows that the loss in power relative to the conventional LR test is quite small. Therefore performing Step 3 in Algorithm 2 makes the resulting confidence set estimate only slightly more conservative than it would be if the conventional LR test always had the correct size no matter the true parameter values nor the sample size.

It remains to discuss how to generate the grid of relevant points used in Algorithm 2. This problem also comes up in Chernozhukov, Hansen, and Jansson (2009). To resolve it, they suggest a random walk Metropolis-Hastings algorithm to generate an adaptive grid over relevant regions of the parameter space. They report, however, that when the confidence set is disconnected that algorithm explores one region but fails to jump to other regions. A simpler approach that was found here to be far more robust is to perform a basic stochastic search over $\mathcal{B} \subseteq \Theta$, a subset of the admissible region of the parameter space.

If the subset $\mathcal{B}$ represents *a priori* information such that $\theta \in \mathcal{B}$ with probability 1, then $\theta \in C_{1-\alpha}^{hyb} \Leftrightarrow \theta \in C_{1-\alpha}^{hyb} \cap \mathcal{B}$. In that case restricting the search to $\mathcal{B}$ does not modify the level of the derived confidence interval; i.e., $\Pr[\theta \in C_{1-\alpha}^{hyb}] = \Pr[\theta \in C_{1-\alpha}^{hyb} \cap \mathcal{B}] \geq 1-\alpha$. In most realistic cases, $\mathcal{B}$ will depend on sample information. For instance, suppose the subset $\mathcal{B}(\hat{\theta})$ is constructed by simulating from a uniform distribution in a neighborhood of $\hat{\theta}$, which should be sufficiently wide to ensure that $C_{1-\alpha}^{hyb} \subseteq \mathcal{B}(\hat{\theta})$. Setting the limits of that neighborhood and choosing the number of simulated points will inevitably involve some fine tuning in any given application. A simple way to check the adequacy of $\mathcal{B}(\hat{\theta})$ is presented in the next section.

Before closing this section, it is important to make two remarks about Algorithm 2. The first one is that the randomness in $\widetilde{Y} \equiv \widetilde{Y}(\vartheta_0, \tilde{\varepsilon}(\eta_0))$ is held constant across the candidate points $\theta_0^{(i)}$, $i = 1, ..., M$. To see what is meant by that, suppose the simulated values $\tilde{\varepsilon}(\eta_0) = (\tilde{\varepsilon}_{-H}, ..., \tilde{\varepsilon}_0, \tilde{\varepsilon}_1, ..., \tilde{\varepsilon}_T)$ used to get $\widetilde{Y}$ are generated via the transformation $\tilde{\varepsilon}_t = G^{-1}(u_t; \eta_0)$, where u_t is a drawing from the $U(0,1)$ distribution and G^{-1} is the inverse function to G, the cumulative distribution function of ε_t. If G is strictly increasing, then the inverse function G^{-1} is a well-defined mapping from $[0,1]$ on to the support of G and $G^{-1}(u_t; \eta_0)$ is a drawing from the distribution of ε_t given η_0. The construction of confidence sets according to Algorithm 2 proceeds by varying $\theta_0^{(i)}$ only; i.e., the same draws of u_t are used for all values $\theta_0^{(i)}$ so that the bootstrap p-value $p_B(\theta_0^{(i)})$ does not depend of the sequence of u_t's. A simple way to ensure this is to reset the seed of the random number generator to the same value for each $\theta_0^{(i)}$, $i = 1, ..., M$.

The second remark about Algorithm 2 is that it is naturally parallelizable on a multi-core computer; i.e., the generated stochastic grid of points that comprise $\mathcal{B}(\hat{\theta})$ can be split into subgrids and those can be searched over in parallel across multiple cores. A parallel implementation of the proposed algorithm can dramatically speed up its execution time.

2.4 Projection-based confidence sets

From the joint confidence set, it is possible to derive conservative confidence sets and intervals for general functions of the parameter vector θ using the projection technique in Abdelkhalek and Dufour (1998); see also Dufour and Jasiak (2001), Dufour and Taamouti (2005), Coudin and Dufour (2009), and Chernozhukov, Hansen, and Jansson (2009) for other examples of this technique. To introduce the approach, consider a non-linear function $h(\theta) \equiv h(\theta, Y) \in \mathbb{R}^d$ of θ and the observed sample Y, treated as given. It is then clear that $\theta \in C_{1-\alpha}^{hyb} \Rightarrow h(\theta) \in h(C_{1-\alpha}^{hyb})$, where the image set is defined as $h(C_{1-\alpha}^{hyb}) = \{\psi \in \mathbb{R}^d : \exists\, \theta \in C_{1-\alpha}^{hyb}, h(\theta) = \psi\}$. Therefore

$$\Pr[h(\theta) \in h(C_{1-\alpha}^{hyb})] \geq \Pr[\theta \in C_{1-\alpha}^{hyb}] \geq 1 - \alpha, \tag{12}$$

meaning that $h(C_{1-\alpha}^{hyb})$ is a conservative confidence set for $h(\theta)$; i.e., one for which the level is at least $1 - \alpha$. In addition if $h(\cdot)$ is continuous, then $h(C_{1-\alpha}^{hyb})$ is compact; see Proposition 1 in Abdelkhalek and Dufour (1998).

Consider the problem of constructing confidence intervals for the individual elements of θ, say θ_i, $i = 1, ..., n$. Define the functions $h_i(\theta) = \theta_i$ for $i = 1, ..., n$. It then follows from (12) that

$$\Pr[\theta_i \in h_i(C_{1-\alpha}^{hyb}), i = 1, ..., n] \geq 1 - \alpha, \tag{13}$$

which shows that $h_i(C_{1-\alpha}^{hyb})$, $i = 1, ..., n$, are simultaneous confidence sets whose joint level is $1 - \alpha$; see Miller (1981) and Dufour (1989) for more on simultaneous confidence sets. Here $h_i(C_{1-\alpha}^{hyb})$ is the projection of $C_{1-\alpha}^{hyb}$ on to the θ_i-axis. Even for a scalar function $h_i(\theta)$, the projection-based confidence set does not necessarily take the form of an interval. This could be the case, for instance, when $C_{1-\alpha}^{hyb}$ is disconnected. To obtain a confidence *interval*, note that $h_i(\theta) \in h_i(C_{1-\alpha}^{hyb})$ implies that

$$h_i^L(C_{1-\alpha}^{hyb}) \leq h_i(\theta) \leq h_i^U(C_{1-\alpha}^{hyb}), \quad i = 1, ..., n,$$

where $h_i^L(C_{1-\alpha}^{hyb}) = \inf\{h_i(\theta_0) : \theta_0 \in C_{1-\alpha}^{hyb}\}$ and $h_i^U(C_{1-\alpha}^{hyb}) = \sup\{h_i(\theta_0) : \theta_0 \in C_{1-\alpha}^{hyb}\}$. Therefore

$$\Pr[h_i^L(C_{1-\alpha}^{hyb}) \leq h_i(\theta) \leq h_i^U(C_{1-\alpha}^{hyb}), i = 1, ..., n] \geq 1 - \alpha, \tag{14}$$

so that the intervals

$$\left[h_i^L(C_{1-\alpha}^{hyb}),\ h_i^U(C_{1-\alpha}^{hyb})\right], \quad i = 1, ..., n, \tag{15}$$

constitute valid confidence intervals (with joint level $1-\alpha$) for the functions $h_i(\theta)$, $i = 1, ..., n$. A distinguishing feature of the intervals in (15) is that they are *simultaneous* confidence intervals, as (14) shows.

The confidence intervals in (15) can be used to obtain an exact test of the null hypothesis of homoskedasticity against a GARCH alternative. For example, suppose the alternative hypothesis is a GARCH(1,1) process. Then the null under test is $H_0 : a_1 = 0$ against the one-sided alternative $H_1 : a_1 > 0$, where b_1 is free under both H_0 and H_1. Define $h_1(\theta) = a_1$ in this context. The decision rule is then to reject H_0 at level α if zero is not contained in the confidence interval $[h_1^L(C_{1-\alpha}^{hyb}), h_1^U(C_{1-\alpha}^{hyb})]$, otherwise there is not sufficient evidence to reject it at that level of significance. This procedure yields a conservative LR test of H_0 against H_1 with level α, even though b_1 is not identified under the null hypothesis.

The derived intervals in (15) also provide a simple way to check whether the search set $\mathcal{B}(\hat{\theta})$ is sufficiently wide: if

$$\left[h_i^L\big(\mathcal{B}(\hat{\theta})\big),\ h_i^U\big(\mathcal{B}(\hat{\theta})\big)\right] \subset \left[h_i^L\big(C_{1-\alpha}^{hyb}\big),\ h_i^U\big(C_{1-\alpha}^{hyb}\big)\right], \text{ for } i = 1, ..., n, \tag{16}$$

then $\mathcal{B}(\hat{\theta})$ is too small. In words, the condition in (16) means that the lower and upper limit points $h_i^L(\mathcal{B}(\hat{\theta}))$ and $h_i^U(\mathcal{B}(\hat{\theta}))$ are not rejected and the projected search interval $\left[h_i^L(\mathcal{B}(\hat{\theta})), h_i^U(\mathcal{B}(\hat{\theta}))\right]$ is thus a strict subset of $\left[h_i^L(C_{1-\alpha}^{hyb}), h_i^U(C_{1-\alpha}^{hyb})\right]$. When that happens, the limits of $\mathcal{B}(\hat{\theta})$ should be extended at least until (16) no longer holds for any i. The empirical application in Section 4 illustrates the use of this diagnostic check.

As a further illustration of (12), suppose one wishes to obtain simultaneous confidence intervals for multiple-step ahead volatility forecasts from (4) starting at time T. As Francq and Zakoïan (2010, p. 11) note, there is no general agreement concerning the definition of "volatility" in the literature. Sometimes it refers to a conditional standard deviation and sometimes to a conditional variance. Following Francq and Zakoïan, volatility here refers to the conditional standard deviation. Let $\sigma_{T+\kappa}(\theta) = \sqrt{\sigma_{T+\kappa}^2(\theta)}$ denote the κ-steps

ahead volatility forecast given $\mathcal{I}_T$. In the GARCH(1,1) case, the 1-step ahead forecast for the conditional variance is

$$\sigma^2_{T+1}(\theta) = \omega + a_1(y_T - \mu)^2 + b_1\sigma^2_T(\theta),$$

since $\sigma^2_T(\theta)$ is a function of the model parameters and the observed sample. In turn, the κ-steps ahead forecast can be calculated from the well-known recursion

$$\sigma^2_{T+\kappa}(\theta) = \omega + (a_1 + b_1)\sigma^2_{T+\kappa-1}(\theta), \quad \text{for } \kappa \geq 2,$$

which will converge to the model-implied unconditional value $\sigma^2(\theta) = \omega/(1 - a_1 - b_1)$ as κ increases, under the maintained stationarity assumption (i.e. the restriction $a_1 + b_1 < 1$ ensures mean reversion). The same approach used to get (15) implies that the intervals

$$\left[\sigma^L_{T+\kappa}(C^{hyb}_{1-\alpha}),\ \sigma^U_{T+\kappa}(C^{hyb}_{1-\alpha})\right], \quad \kappa = 1, ..., K, \tag{17}$$

constitute valid simultaneous confidence intervals (with joint level $1-\alpha$) for the K volatility forecasts. It should be emphasized that the commonly used delta method (cf. Lai and Xing 2008, p. 149) does not yield simultaneous confidence intervals such as those advocated here.

3 Size and power comparisons

The key part of the proposed inference method is the parametric bootstrap test. Indeed, the coverage probability and width of the confidence set depends on the exactness and power of the proposed test. In this section, the empirical size and power of the bootstrap LR approach is compared to that of the conventional LR test based on the usual asymptotic chi-square distribution.

To set the stage, recall that a well-known stylized fact about financial asset returns is that they exhibit excess kurtosis which is at odds with models that assume a normal distribution of returns; see also the evidence in Section 4 below. While the commonly used Student-t distribution can capture heavy tails, it still restricts returns to be symmetrically

distributed around their means. In order to allow for possible return asymmetries, the ε_t innovation terms in (3) are assumed to be i.i.d. random variables following the skewed Student-t distribution of Hansen (1994), which introduces an additional parameter that imparts asymmetry. Specifically, Hansen's skewed Student-t distribution is defined as

$$g(z;v,\lambda) = \begin{cases} bc\left(1+\frac{1}{v-2}\left(\frac{bz+a}{1-\lambda}\right)^2\right)^{-(v+1)/2}, & \text{if } z < -a/b, \\ bc\left(1+\frac{1}{v-2}\left(\frac{bz+a}{1+\lambda}\right)^2\right)^{-(v+1)/2}, & \text{if } z \geq -a/b, \end{cases} \tag{18}$$

where the constants a, b, and c are themselves defined as

$$a = 4\lambda c\frac{v-2}{v-1}, \quad b^2 = 1+3\lambda^2-a^2, \quad c = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi(v-2)}\Gamma(\frac{v}{2})},$$

and the parameters $v > 2$ and $-1 < \lambda < 1$ represent the degrees of freedom and the asymmetry of the distribution, respectively. When $\lambda = 0$, (18) reduces to the traditional Student-t distribution, and a positive value for λ implies right skewness. When modeling the distribution of financial returns it is usually the case that there is a greater probability of observing large negative returns than large positive ones. In that case the distribution is skewed to the left, which corresponds to a negative value for λ in (18).

Jondeau and Rockinger (2003) show that a random variable distributed according the the skewed Student-t distribution in (18) has mean zero and unit variance. They further show that the associated cumulative distribution function (cdf) is defined by

$$G(z;v,\lambda) = \begin{cases} (1-\lambda)T\left(\frac{bz+a}{1-\lambda}\sqrt{\frac{v}{v-2}};v\right), & \text{if } z < -a/b, \\ (1+\lambda)T\left(\frac{bz+a}{1+\lambda}\sqrt{\frac{v}{v-2}};v\right)-\lambda, & \text{if } z \geq -a/b, \end{cases}$$

where $T(x;v)$ is the cdf of the standard Student-t distribution with v degrees of freedom, and that the inverse of this cdf is given by

$$G^{-1}(u;v,\lambda) = \begin{cases} \frac{1}{b}\left((1-\lambda)\sqrt{\frac{v-2}{v}}T^{-1}\left(\frac{u}{1-\lambda};v\right)-a\right), & \text{if } u < \frac{1-\lambda}{2}, \\ \frac{1}{b}\left((1+\lambda)\sqrt{\frac{v-2}{v}}T^{-1}\left(\frac{u+\lambda}{1+\lambda};v\right)-a\right), & \text{if } u \geq \frac{1-\lambda}{2}. \end{cases}$$

This last expression is useful for generating a random draw z from the skewed Student-t distribution with parameters v and λ, since it suffices to first generate $u \sim U(0,1)$ and to then compute $z = G^{-1}(u; v, \lambda)$. See Section 5.2.3 in Jondeau, Poon, and Rockinger (2007) for a textbook discussion of Hansen's skewed Student-t distribution.

The simulation study examines the GARCH(1,1) version of (3) and (4) with the random innovations ε_t begin draws from the skewed Student-t distribution in (18). The parameters μ and ω are set as $\mu = 0$ and $\omega = 0.01$. In the size comparisons, the volatility parameters a_1, b_1 are varied over the pairs $(0.07, 0.92)$, $(0.03, 0.92)$, and $(0.03, 0.96)$ to examine the differential effects of changes in the ARCH parameter and/or the GARCH parameter. The degrees-of-freedom parameter in (18) varies as $v = 3, 4, 5$ and the parameter governing the shape of that distribution varies as $\lambda = 0, -0.4, -0.8$; i.e., from symmetric to more and more left skewed. The simulation results are based on 1000 replications of each data-generating process (DGP). The computations were done in Fortran using IMSL. Specifically, maximum likelihood estimation was done with the Nelder-Mead simplex algorithm.

For a given nominal 5% level, Table 1 shows the empirical size of the LR test based on the asymptotic chi-square distribution (LR_{asy}), the proposed bootstrap version (LR_{boot}) in Algorithm 1 based on $B = 19$, and the proposed hybrid version (LR_{hyb}) used in Algorithm 2 also based on $B = 19$. The top three panels show the results for a sample of size $T = 120$ and the last one is for $T = 520$. These sample sizes correspond to the 10 years of monthly returns and weekly returns data, respectively, used in the empirical application in Section 4. The picture that emerges from Table 1 is that LR_{boot} and LR_{hyb} behave as expected with empirical sizes respecting the nominal 5% level constraint, no matter the sample size and the DGP parameter configuration. On the other hand, LR_{asy} can have marked size distortions when T is small and those tend to be exacerbated when: (i) the degrees of freedom decrease (implying heavier tails) and (ii) when the innovation distribution becomes more skewed. In the top panel for example where $a_1 = 0.07, b_1 = 0.92$, the empirical size of LR_{asy} is about 22% when $v = 5, \lambda = -0.8$. And that distortion more than doubles to about 53% when v decreases from 5 to 3.

Table 1: Empirical rejection rates under the DGP

	$v=3$			$v=4$			$v=5$		
	$\lambda=0$	-0.4	-0.8	$\lambda=0$	-0.4	-0.8	$\lambda=0$	-0.4	-0.8
$T=120$, $a_1=0.07$, $b_1=0.92$									
LR_{asy}	4.4	8.9	52.9	6.2	6.6	29.5	7.9	8.1	22.5
LR_{boot}	5.1	5.2	5.0	4.5	5.1	5.1	5.0	5.1	4.8
LR_{hyb}	3.4	4.2	4.8	4.0	3.9	4.7	4.3	3.9	4.0
$T=120$, $a_1=0.03$, $b_1=0.92$									
LR_{asy}	6.9	5.7	28.5	3.1	4.1	17.2	2.4	2.7	14.7
LR_{boot}	4.0	5.0	5.1	5.0	4.9	5.2	5.1	4.8	4.9
LR_{hyb}	3.1	3.6	4.6	2.6	3.1	4.8	2.0	2.2	4.0
$T=120$, $a_1=0.03$, $b_1=0.96$									
LR_{asy}	4.1	11.3	57.1	4.2	7.3	35.2	3.2	4.5	22.1
LR_{boot}	4.5	5.1	5.3	5.0	5.2	4.8	4.6	4.7	5.1
LR_{hyb}	2.6	4.0	5.0	3.3	3.4	4.1	2.5	2.9	4.2
$T=520$, $a_1=0.03$, $b_1=0.96$									
LR_{asy}	2.8	7.1	34.0	4.5	3.6	12.3	5.1	3.2	7.8
LR_{boot}	4.6	4.8	4.8	4.4	4.6	4.7	5.0	4.5	4.9
LR_{hyb}	2.2	4.1	3.9	3.0	2.8	3.1	2.6	2.4	4.0

Notes: The entries are the empirical size of each test as the values in the null hypothesis are varied. The DGP is a GARCH(1,1) with parameters a_1, b_1. The innovations are drawn from a skewed Student-t distribution with degrees of freedom v and asymmetry parameter λ. When $\lambda = 0$ the distribution is symmetric and lower values of λ imply greater left skewness.

Table 2: Empirical rejection rates under departures from the DGP

	LR_{asy}			LR_{boot}			LR_{hyb}		
$T = 120$									
	$b_1 = 0.80$	0.85	0.89	0.80	0.85	0.89	0.80	0.85	0.89
$a_1 = 0.10$	3.9	17.2	77.1	5.0	23.3	78.6	2.8	13.7	70.8
0.05	18.2	5.8	31.7	22.6	10.5	37.6	15.3	5.3	28.8
0.01	60.3	29.5	22.8	65.0	36.6	30.3	58.7	26.7	19.8
$T = 240$									
	$b_1 = 0.80$	0.85	0.89	0.80	0.85	0.89	0.80	0.85	0.89
$a_1 = 0.10$	4.0	42.1	99.4	5.1	45.9	98.1	3.0	35.7	97.8
0.05	36.6	9.0	64.1	41.0	14.7	67.9	32.1	7.3	59.4
0.01	89.0	54.5	40.8	91.0	61.2	48.1	87.3	51.7	37.1

Notes: The GARCH(1,1) parameters in the DGP are set as $a_1 = 0.10$ and $b_1 = 0.80$, so the entries in positions [1,1] are the empirical size of each test. The other entries are the empirical power as the values of a_1, b_1 in the null hypothesis depart from their DGP counterparts.

How are the small-sample size distortions of LR_{asy} related to the volatility parameters? Comparing the top two panels of Table 1, one can see that as the ARCH parameter a_1 decreases from 0.07 to 0.03 that the size distortions diminish. The reason for this is that as a_1 goes to zero the data become homoskedastic. Indeed, this is not surprising since as Dufour (1997) points out, the likelihood function is flat on the non-identification subset. On the flip side, as the GARCH parameter increases from 0.92 in the second panel to 0.96 in the third panel of Table 1 (holding a_1 fixed at 0.03), the size distortions amplify. In the third panel under $v = 3, \lambda = -0.8$ the size of LR_{asy} attains an astonishing 57%. Even when the sample size grows to $T = 520$ (Table 1, panel 4), the behavior of LR_{asy} still doesn't agree with its asymptotic distribution when $v = 3$ and $\lambda = -0.8$.

Table 2 provides a comparison of the empirical rejection rates of LR_{asy}, LR_{boot}, and LR_{hyb} when the null hypothesis under test departs from the DGP. The parameters of the DGP are set as $\mu = 0$, $\omega = 0.01$, $a_1 = 0.10$, $b_1 = 0.80$, $v = 5$, and $\lambda = 0$. In the null hypothesis under test only the volatility parameters vary as $a_1 = 0.10, 0.05, 0.01$ and $b_1 = 0.80, 0.85, 0.89$. The sample size is $T = 120$ and 240. So Table 2 shows the resulting six 3×3 arrays; the two in the left (middle, right) column show the results for LR_{asy}

(LR_{boot}, LR_{hyb}). In each case, the [1,1] entry represents the situation where the null being tested is true. And indeed the empirical rejection rates are seen not to violate the nominal 5% level constraint. In each of the cases along the main diagonal, the persistence of the GARCH model (as measured by $a_1 + b_1$) is held equal to 0.90. As the null under test departs from the DGP, power increases when either a_1 or b_1 depart from their true values in the [1,1] case. But the power increases much more slowly if the value of $a_1 + b_1$ in the null hypothesis equals that in the DGP. Of course, power increases as the sample size increases. This is seen by comparing the top and bottom panels. The overall conclusion that emerges from Table 2 is that the power loss with LR_{hyb} relative to LR_{asy} is quite small, and that LR_{boot} dominates the two other tests. These results are all the more remarkable considering that LR_{boot} and LR_{hyb} use only $B = 19$ bootstrap replications, and it is also clear that LR_{boot} should be preferred when the computational cost can be afforded.

4 Application to volatility forecasting

The proposed inference method is illustrated with an application to 10 years worth of monthly, weekly, and daily returns on a major U.S. stock market index: the Standard & Poor's (S&P) 500 index, which represents the prices of 500 large market capitalization stocks actively traded in the United States.

Table 3: Summary statistics: S&P 500 index returns

	Mean	Std. dev.	Max.	Min.	Skewness	Kurtosis
Monthly	-0.0015	0.0483	0.0897	-0.1856	-0.7871	4.0376
Weekly	-0.0003	0.0272	0.1135	-0.2008	-0.8841	10.3310
Daily	-6.46E-5	0.0138	0.1095	-0.0946	-0.1143	10.9414

Notes: The table reports the sample mean, standard deviation, maximum value, minimum value, and the coefficients of skewness and kurtosis. The sample covers the period from November 2000 to October 2010.

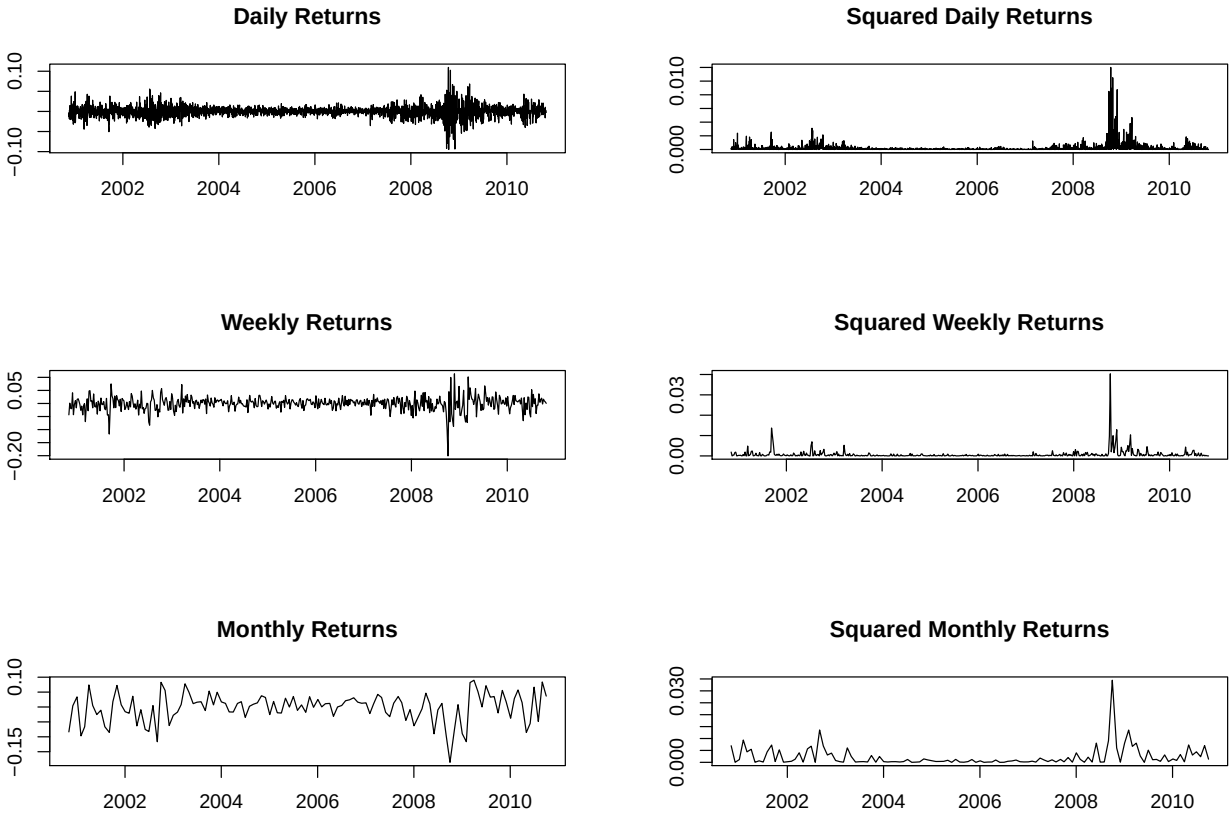

Figure 1: Time-series plots of the daily, weekly, and monthly log-returns on the S&P 500 stock market index (left column) and their squared values (right column) over a ten year period from November 2000 to October 2010.

The data consist of closing index prices adjusted for dividends and splits. The level data were converted to returns by taking the first difference over time of the natural logarithm of the daily (weekly, monthly) prices. The resulting time series cover the periods: (i) November 1, 2000 to October 10, 2010 for the monthly returns, (ii) November 6, 2000 to October 25, 2010 for the weekly returns, and (iii) November 16, 2000 to October 27, 2010 for the daily returns. The monthly, weekly, and daily returns comprise 120, 520, and 2500 observations, respectively. Figure 1 shows the three return series and their corresponding squared values. The presence of volatility clustering effects is evident in each series and the increased volatility during the financial crisis starting in 2008 is also quite remarkable. Table 3 reports the sample mean, standard deviation, maximum value, minimum value, and the coefficients of skewness and kurtosis for the three return series. It is interesting to note that in this sample monthly returns are relatively more left skewed and heavy tailed than daily returns, which appear to be more symmetrical and light-tailed.

Table 4: Projection-based confidence intervals: GARCH(1,1) model of S&P 500 index returns

	μ	ω	a_1	b_1	v	λ
Monthly returns, $T = 120$						
MLE	0.0070	4.42E-5	0.0893	0.8692	13.8476	-0.7139
SI	[-0.0149, 0.0149]	[1.51E-7, 9.99E-4]	[2.90E-6, 0.5893]	[0.3692, 0.9980]	[2.0001, 29.9987]	[-0.9899, 0.0998]
$CI^{asy}_{0.95}$	[-0.0080, 0.0146]	[1.52E-7, 0.0005]	[0.0533, 0.5086]	[0.4485, 0.9279]	[3.8184, 29.9908]	[-0.9635, -0.0244]
$CI^{hyb}_{0.95}$	[-0.0080, 0.0146]	[1.52E-7, 0.0009]	[0.0068, 0.5086]	[0.4485, 0.9923]	[2.0817, 29.9908]	[-0.9635, -0.0244]
Weekly returns, $T = 520$						
MLE	0.0008	9.87E-6	0.0586	0.9176	6.7949	-0.2216
SI	[-0.0149, 0.0149]	[1.45E-8, 9.99E-5]	[2.64E-6, 0.2086]	[0.7676, 0.9981]	[2.0013, 20.0000]	[-0.9899, 0.0999]
$CI^{hyb}_{0.95}$	[-0.0016, 0.0039]	[2.68E-6, 5.80E-5]	[0.0536, 0.2029]	[0.7683, 0.9187]	[3.7968, 19.9257]	[-0.4208, -0.0095]
Daily returns, $T = 2500$						
MLE	0.0002	1.04E-6	0.0735	0.9237	6.5881	-0.1442
SI	[-0.0014, 0.0014]	[4.11E-10, 9.99E-5]	[0.0435, 0.1035]	[0.8937, 0.9537]	[4.0000, 9.9999]	[-0.9899, -5.20E-5]
$CI^{hyb}_{0.95}$	[-0.0003, 0.0010]	[3.93E-7, 2.16E-6]	[0.0560, 0.0983]	[0.8982, 0.9404]	[6.3878, 9.6365]	[-0.1781, -0.0037]

Notes: The entries below the MLEs are simultaneous confidence intervals at the 0.95 level for the parameters of a GARCH(1,1) model with skewed Student-t innovations. The intervals indicated by $CI^{asy}_{0.95}$ are obtained by inverting the asymptotic LR test and those indicated by $CI^{hyb}_{0.95}$ are obtained according to the hybrid approach in Algorithm 2. For weekly and daily returns, the asymptotic and hybrid confidence intervals are identical.

The same GARCH(1,1) model with skewed Student-t innovations examined in the previous section was estimated using each return series in turn. The results are reported in Table 4. The table also shows projection-based confidence intervals of the form (15) for the individual parameters at the simultaneous 95% level. The limits of $\mathcal{B}(\hat{\theta})$ were fine tuned to ensure that (16) did not hold for any of the parameters. This can be ascertained from the lines labeled SI in Table 4 which show the resulting projected search intervals $\left[h_i^L(\mathcal{B}(\hat{\theta})), h_i^U(\mathcal{B}(\hat{\theta}))\right]$ for each parameter. The only exception occurs for the degrees-of-freedom parameter with monthly returns, for which it was assumed that values beyond 30 were inadmissible. Abdelkhalek and Dufour (1998) show that eliminating (truncating) inadmissible values from a confidence set does not modify its level. The simulation results presented in the previous section (see Table 1) revealed that the null behavior of the conventional likelihood ratio test is essentially like that of the finite-sample bootstrap test for this model once the sample size T reaches 520 observations and v is at least 5. So perhaps not surprisingly the asymptotic and hybrid confidence intervals are identical with the weekly and daily return series (in the second and third panels of Table 4).

The first four columns of Table 4 show the inference results for the μ, ω, a_1, and b_1 parameters. Since the confidence intervals for the ARCH parameter a_1 exclude zero, the null hypothesis of homoskedasticity is rejected at the 5% level in each case in favor of the GARCH(1,1) alternative. The persistence of shocks to the conditional variance, as measured by $a_1 + b_1$, is seen to increase with the return frequency. The results in the last two columns pertain to the degrees-of-freedom and shape parameters, v and λ respectively. The confidence intervals for v are wide and quite uninformative, which shows just how hard it is to pin down this parameter. On the other hand, the other parameters seem to be relatively better identified with narrower confidence intervals. Of course, these become narrower as the sample size increases. In light of the connection between maximum likelihood estimators and Hodges-Lehmann estimators previously discussed, it is interesting to observe the locations of the point estimates in their respective marginal confidence intervals.

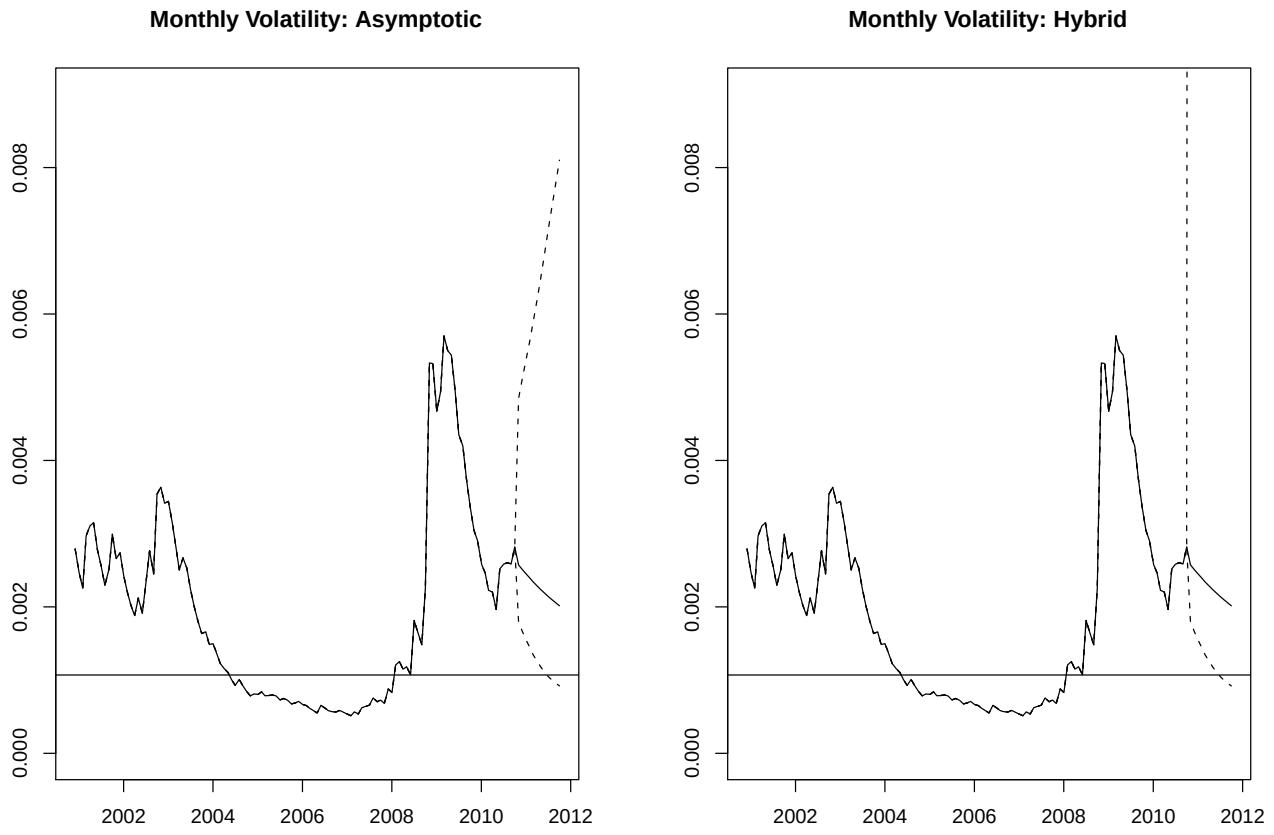

Figure 2: Time-series plots of the model-implied monthly volatility. The jagged line up to October, 2010 represents the in-sample fitted values. After that date are shown the out-of-sample volatility forecasts for the next 12 months along with simultaneous 95% confidence bands derived from the asymptotic LR test (left plot) and the hybrid approach in Algorithm 2 (right plot). The horizontal line is the model-implied unconditional standard deviation.

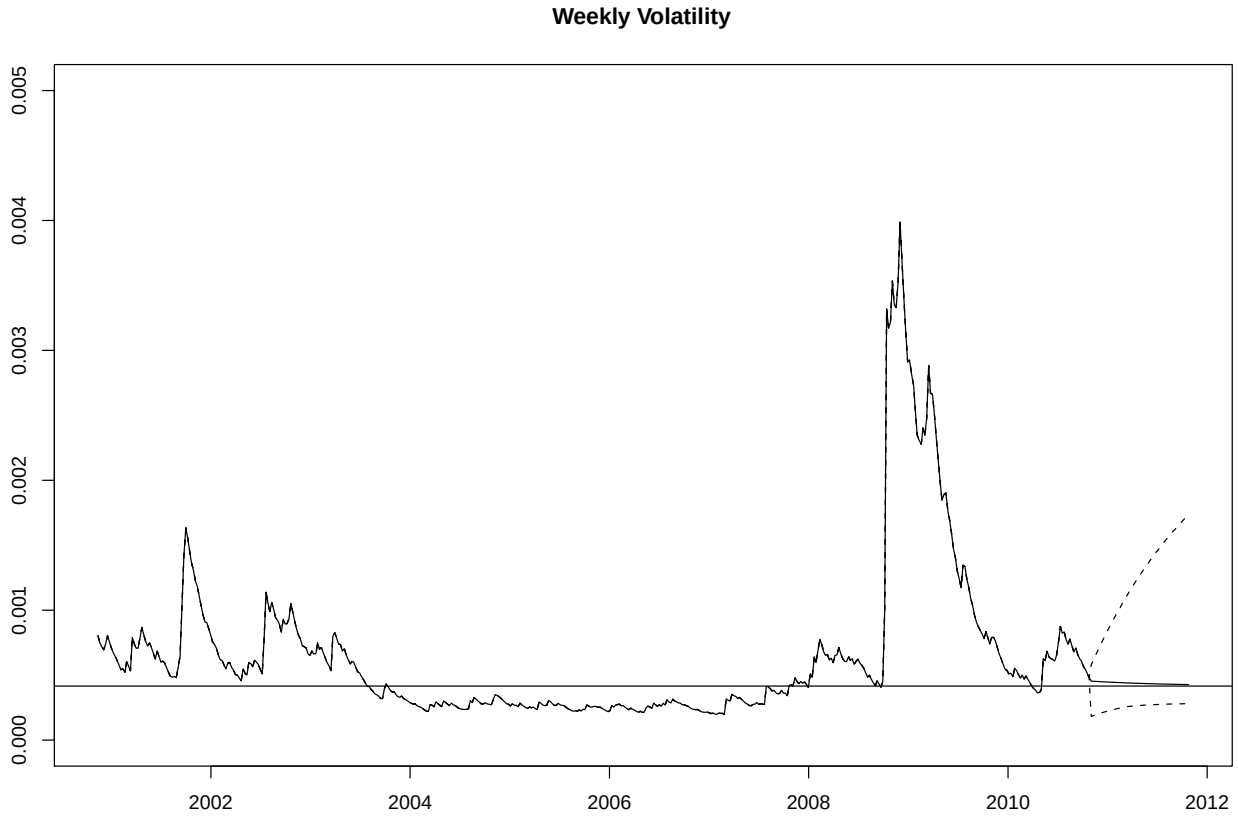

Figure 3: Time-series plots of the model-implied weekly volatility. The jagged line up to the week of October 25, 2010 represents the in-sample fitted values. After that date are shown the out-of-sample volatility forecasts for the next 52 weeks along with simultaneous 95% confidence bands derived from the LR test. (The asymptotic and hybrid approaches yield identical results in this case.) The horizontal line is the model-implied unconditional standard deviation.

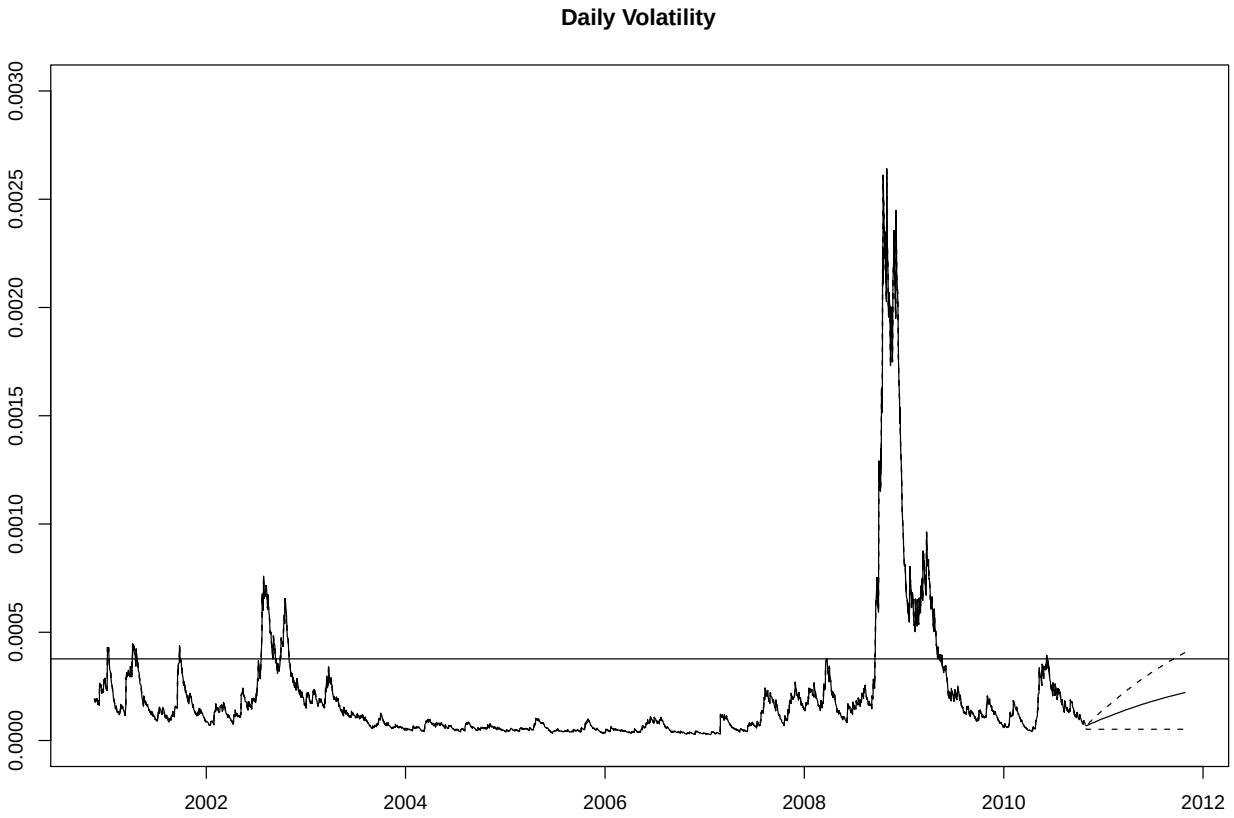

Figure 4: Time-series plots of the model-implied monthly volatility. The jagged line up to October 27, 2010 represents the in-sample fitted values. After that date are shown the out-of-sample volatility forecasts for the next 250 days along with simultaneous 95% confidence bands derived from the LR test. (The asymptotic and hybrid approaches yield identical results in this case.) The horizontal line is the model-implied unconditional standard deviation.

Figures 2, 3, and 4 show the time-series plots of the model-implied volatilities with monthly, weekly, and daily returns, respectively. In each case, the jagged lines represent the in-sample fitted values. After the last in-sample dates are shown out-of-sample volatility forecasts for the next year along with simultaneous 95% confidence bands. In each figure, those bands connect the end points of the 12-months ahead, 52-weeks ahead, and 250-days ahead simultaneous 95% confidence intervals, computed according to (17). The horizontal line in each figure is the value of the model-implied unconditional standard deviation. As expected, the volatility forecasts are seen to converge to their unconditional value as the forecast horizon increases.

Figure 2 shows that the monthly volatility forecast confidence intervals are very wide, which is to be expected given the quite small sample size in that case. As Dufour (1997) emphasizes, accepting the possibility of extremely wide confidence intervals is simply a matter of logic and scientific rigor: the data may simply by uninformative about certain parameters. And that is precisely what Figure 2 illustrates with the monthly volatility forecasts. That figure also illustrates some differences between the asymptotic and hybrid confidence intervals, which, in line with the results in Table 4, suggests that the asymptotic test spuriously rejects some candidate points and hence fails to cover some parts of the parameter space in this small sample. In contrast, the confidence bands shown in Figures 3 and 4 for the weekly and daily volatility forecasts are much more informative and reveal some interesting shapes. For instance, the lower band of the weekly volatility forecasts in Figure 3 exhibits a kink and eventually parallels the unconditional volatility value.

5 Conclusion

In this paper, a parametric bootstrap procedure based on the likelihood ratio statistic has been proposed to test any hypothesis that completely sets the parameter values of a stationary GARCH model. When the simulated samples are in their stationary distribution, the proposed procedure yields a test of the null hypothesis with an exact finite-sample (Monte Carlo) p-value. This procedure has the advantage of being valid even when the

standard regularity conditions for asymptotic inference do not hold. For instance, the model parameters need not be identified and moments beyond order two of the innovation distribution need not be finite.

The proposed procedure is used in the construction of confidence sets for the GARCH model parameters. The construction method proceeds by numerically inverting the likelihood ratio test. In a first step, a set of admissible candidate points is generated by simulation. Next, if a given candidate point passes the likelihood ratio test using its asymptotic critical value, it then immediately becomes part of the confidence set. If the candidate point fails the conventional test, it then gets included in the confidence set only if it passes the parametric bootstrap test. This last step provides a hedge against the risk of a spurious rejection by the likelihood ratio test with asymptotic critical value. The results of simulation experiments revealed that the power loss with the hedging step is quite small, meaning that the resulting confidence set is not much bigger than it would be if the conventional test always had the correct size no matter the true parameter values nor the sample size. A projection technique was then advocated to produce conservative confidence sets for general functions of the parameters and the observed sample. As computer technology continues to evolve with time, computational savings will become less of a concern and the confidence sets may be constructed by skipping the asymptotic test altogether and using only the preferred parametric bootstrap with an even greater number of replications.

Although the proposed method has been presented in the context of a benchmark GARCH model (and illustrated with heavy-tailed and skewed innovations), it can be extended quite naturally to other GARCH specifications. For instance, the empirical finance literature has documented cases where negative returns tend to be followed by larger increases in volatility than do equally large positive returns. Asymmetric GARCH models that allow for such leverage effects include Nelson's (1991) Exponential GARCH model, which admits strictly stationary solutions under a certain parameter restriction.

Acknowledgments

I would like to thank the editor and two anonymous referees, as well as Lynda Khalaf for useful comments that helped improve the paper.

References

Abdelkhalek, T., Dufour, J.-M., 1998. Statistical inference for computable general equilibrium models, with application to a model of the Maroccan economy. Review of Economics and Statistics 80, 520–534.

Athreya, K.B., 1987a. Bootstrap of the mean in the infinite variance case. Annals of Statistics 15, 724–731.

Athreya, K.B., 1987b. Bootstrap of the mean in the infinite variance case. In Proceedings of the 1st World Congress of the Bernoulli Society, 2. VNU Scientific Press, Utrecht.

Berkes, I., Horváth, L., Kokoszka, P., 2003. GARCH processes: structure and estimation. Bernoulli 9, 201–227.

Bolduc, D., Khalaf, L., Yélou, C., 2010. Identification robust confidence set methods for inference on parameter ratios with application to discrete choice models. Journal of Econometrics 157, 317–327.

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

Bollerslev, T., 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. Review of Economics and Statistics 69, 542–547.

Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. In Handbook of Econometrics, Vol. IV (eds. Robert F. Engle and Daniel McFadden). Elsevier, Amsterdam.

Bollerslev, T., Wooldridge, J.M., 1992. Quais-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econometric Reviews 11, 143–172.

Bougerol, P., Picard, N., 1992. Stationarity of GARCH processes and some nonnegative time series. Journal of Econometrics 52, 115–127.

Chernozhukov, V., Hansen C., Jansson, M., 2009. Finite sample inference for quantile regression models. Journal of Econometrics 152, 93–103.

Christoffersen, P.F., 2003. Elements of Financial Risk Management. Academic Press, San Diego.

Coudin, E., Dufour, J.-M., 2009. Finite-sample distribution-free inference in linear median regressions under heteroscedasticity and non-linear dependence of unknown form. Econometrics Journal 12, 19–49.

Crowder, M.J., 1976. Maximum likelihood estimation with dependent observations. Journal of the Royal Statistical Society 38, 45–53.

Davies, R.B., 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64, 247–254.

Davies, R.B., 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74, 33–43.

Dufour, J.-M., 1989. Nonlinear hypotheses, inequality restrictions, and non-nested hypotheses: exact simultaneous tests in linear regressions. Econometrica 57, 335–355.

Dufour, J.-M., 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. Econometrica 65, 1365–1388.

Dufour, J.-M., 2006. Monte Carlo tests with nuisance parameters: a general approach to finite-sample inference and nonstandard asymptotics. Journal of Econometrics 133, 443–477.

Dufour, J.-M., Jasiak, J., 2001. Finite sample limited information inference methods for structural equations and models with generated regressors. International Economic Review 42, 815–843.

Dufour, J.-M., Jouini, T., 2006. Finite-sample simulation-based inference in VAR models with application to Granger causality testing. Journal of Econometrics 135, 229–254.

Dufour, J.-M., Khalaf, L., 2001. Monte Carlo test methods in econometrics. In Companion to Theoretical Econometrics. Basil Blackwell, Oxford, UK.

Dufour, J.-M., Khalaf, L., Kichian, M., 2010. On the precision of Calvo parameter estimates in structural NKPC models. Journal of Economic Dynamics and Control 34, 1582–1595.

Dufour, J.-M., Taamouti, M., 2005. Projection-based statistical inference in linear structural models with possibly weak instruments. Econometrica 73, 1351–1365.

Dufour, J.-M., Torrès, O., 2003. Hodges-Lehmann sample-split estimators. Université Lille 3 Working Paper.

Dufour, J.-M., Valéry, P., 2009. Exact and asymptotic tests for possibly non-regular hypotheses on stochastic volatility models. Journal of Econometrics 150, 193–206.

Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 28, 181–187.

Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1007.

Francq, C., Zakoïan, J.-M., 2004. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. Bernoulli 10, 605–637.

Francq, C., Zakoïan, J.-M., 2007. Quasi-maximum likelihood estimation in GARCH processes when some coefficients are equal to zero. Stochastic Processes and their Applications 117, 1265–1284.

Francq, C., Zakoïan, J.-M., 2010. GARCH Models: Structure, Statistical Inference and Financial Applications. John Wiley & Sons Ltd, Chichester, West Sussex, UK.

Gouriéroux, C., Monfort, A., Trognon, A., 1984. Pseudo maximum likelihood methods: theory. Econometrica 52, 681–700.

Hall, P., 1990. Asymptotic properties of the bootstrap for heavy-tailed distributions. Annals of Probability 18, 1342–1360.

Hall, P., Yao, Q., 2003. Inference in ARCH and GARCH models with heavy-tailed errors. Econometrica 71, 285–317.

Hansen, B., 1994. Autoregressive conditional density estimation. International Economic Review 35, 705–730.

Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH(1,1)? Journal of Applied Econometrics 20, 873–889.

Hodges, J.L., Lehmann, E.L., 1963. Estimates of location based on rank tests. Annals of Mathematical Statistics 34, 598–611.

Hodges, J.L., Lehmann, E.L., 1983. Hodges-Lehmann estimators. In Encyclopedia of Statistical Science, Vol. 3. Wiley, New York.

Jensen, S.T., Rahbek, A., 2004a. Asymptotic inference for nonstationary GARCH. Econometric Theory 20, 1203–1226.

Jensen, S.T., Rahbek, A., 2004b. Asymptotic normality of the QMLE estimator of ARCH in the nonstationary case. Econometrica 72, 641–646.

Jondeau, E., Poon, S.-H., Rockinger, M., 2007. Financial Modeling Under Non-Gaussian Distributions. Springer-Verlag, London.

Jondeau, E., Rockinger, M., 2003. Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. Journal of Economic Dynamics and Control 27, 1699–1737.

Knight, K., 1989. On the bootstrap of the sample mean in the infinite variance case. Annals of Statistics 17, 1168–1175.

Lai, T.L., Xing, H., 2008. Statistical Models and Methods for Financial Markets. Springer, New York.

Lee, S.-W., Hansen, B., 1994. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. Econometric Theory 10, 29–53.

Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses, 3rd edition. Springer, New York.

Lumsdaine, R.L., 1996. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. Econometrica 64, 575–596.

Ma, J., Nelson, C.R., Startz, R., 2007. Spurious inference in the GARCH(1,1) model when it is weakly identified. Studies in Nonlinear Dynamics and Econometrics 11, Article 1.

Mammen, E., 1992. When does bootstrap work? Asymptotic results and simulations. In Lecture Notes in Statistics, 77. Springer, New York.

Miller, R.G., 1981. Simultaneous Statistical Inference, 2nd ed. Springer-Verlag, New York.

Mittnik, S., Rachev, S.T., 2000. Stable Paretian Models in Finance. Wiley, New York.

Mittnik, S., Rachev, S.T., Paolella, M.S., 1998. Stable Paretian modeling in finance: some empirical and theoretical aspects. In A Practical Guide to Heavy Tails. Birkhauser, Boston.

Nelson, D.B., 1990. Stationarity and persistence in the GARCH(1,1) model. Econometric Theory 6, 318–334.

Nelson, D.B., 1991. Conditional heteroskedsticity in asset returns: a new approach. Econometrica 59, 347–370.

Robert, C.P., Casella, G., 2010. Introducing Monte Carlo Methods with R. Springer, New York.

Weiss, A., 1986. Asymptotic theory for ARCH models: estimation and testing. Econometric Theory 2, 107–131.

Zivot, E., 2009. Practical issues in the analysis of univariate GARCH models. In Handbook of Financial Time Series. Springer-Verlag, Berlin.