Econometric Reviews

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: https://www.tandfonline.com/loi/lecr20

A bootstrap approach for generalized Autocontour testing Implications for VIX forecast densities

João Henrique G. Mazzeu, Gloria González-Rivera, Esther Ruiz & Helena Veiga

To cite this article: João Henrique G. Mazzeu, Gloria González-Rivera, Esther Ruiz & Helena Veiga (2020): A bootstrap approach for generalized Autocontour testing Implications for VIX forecast densities, Econometric Reviews, DOI: 10.1080/07474938.2020.1761150

To link to this article: https://doi.org/10.1080/07474938.2020.1761150

ECONOMETRIC REVIEWS
https://doi.org/10.1080/07474938.2020.1761150

Taylor & Francis
Taylor & Francis Group

A bootstrap approach for generalized Autocontour testing Implications for VIX forecast densities

João Henrique G. Mazzeu[a], Gloria González-Rivera[b], Esther Ruiz[a], and Helena Veiga[a,c]

[a]Department of Statistics, Universidad Carlos III de Madrid, Getafe, Madrid, Spain; [b]Departments of California, Riverside, California, USA; [c]BRU-IUL, Instituto Universitário de Lisboa, Lisboa, Portugal

ABSTRACT

We propose an extension of the Generalized Autocontour tests for dynamic specification (evaluation) of *in-sample* (*out-of-sample*) conditional densities. The new tests are based on probability integral transforms computed from bootstrap conditional densities that incorporate parameter uncertainty without relying on parametric assumptions of the error distribution. Their finite sample distributions are well approximated using standard asymptotic distributions while they are easy to implement and provide information about potential sources of misspecification. We apply the new tests to the Heterogeneous Autoregressive and the Multiplicative Error models of the VIX index and find strong evidence against the parametric assumptions of the conditional densities.

KEYWORDS

HAR model; model evaluation; Multiplicative Error Model; PIT

JEL CLASSIFICATION

C22; C52; C53; C58

1. Introduction

Density forecasting is a very important area of research in the analysis of economic and financial time series. A problem often faced by forecasters is testing the correct specification of a conditional forecast density. Corradi and Swanson (2006a), Bierens and Wang (2017) and Rossi and Sekhposyan (2019) contain excellent reviews of the literature on testing for the specification of univariate conditional densities. Many popular tests for conditional forecast densities are based on testing a joint hypothesis of uniformity and independence of the probability integral transforms (PITs). These tests, introduced by Diebold et al. (1998) in the econometric literature, have the advantage of being preferred regardless of the forecaster's loss function. Among these tests, in the context of diffusion processes, Hong and Li (2005) compare the joint nonparametric density of PITs at different lags with the product of two independent U(0,1) random variables. The main disadvantages of this test are that one needs to choose the bandwidth parameter and the test converges at a nonparametric rate. Alternatively, Corradi and Swanson (2006b) construct Kolmogorov-type conditional distribution tests in the presence of both dynamic misspecification and parameter estimation uncertainty; see Corradi and Swanson (2006a) for a generalization to an out-of-sample framework. However, the limiting distribution of these tests is not nuisance parameters free and, consequently, they propose bootstrap techniques in order to obtain valid critical values. Furthermore, they assume that the conditional distribution depends on a finite number of observable values of the variable of interest, excluding moving average or GARCH models; see Perera and Silvapulle (2018). Note that while the tests proposed by Corradi and Swanson (2006a, 2006b) allow for dynamic miss-specification of the conditional moments, in this

CONTACT Esther Ruiz ortega@est-econ.uc3m.es Department of Statistics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Madrid, Spain

Supplemental data for this article is available online at https://doi.org/10.1080/07474938.2020.1761150

paper, we propose tests that are robust to the distributional miss-specification of the conditional distribution of the errors. Very recently, Rossi and Sekhposyan (2019) propose a test based on PITs with the density evaluated at the estimated parameter values.

González-Rivera and Sun (2015) propose the generalized autocontour (G-ACR) tests for independence and uniformity of PITs, which have standard convergence rates and limiting distributions that deliver superior power. Furthermore, G-ACR tests are computationally easy to implement as they are based on a counting process and do not require either a transformation of the original data or an assessment of the Kolmogorov goodness of fit. However, as most tests based on PIT's, G-ACR tests are based on assuming a particular specification of the conditional density (often normality) while, in practice, there are applications in which the density does not have a known closed-form expression. Therefore, when a given predictive density model is rejected, it is difficult to disentangle whether the rejection can be attributed to the assumed functional form of the error distribution or to the specification of the conditional moments.

In this paper, we propose an extension of the G-ACR tests for (in-sample) the dynamic specification of a density model and for (out-of-sample) the evaluation of forecast densities. Our contribution lies on computing the PITs from a bootstrapped conditional density so that no assumption on the functional form of the forecast error density is needed. Alternative tests to those proposed in this article for the case of unknown conditional density functions have been suggested in Ait-Sahalia et al. (2009) and Altissimo and Mele (2009) who propose tests based on kernel estimates of the densities with nonparametric rates. Bhardwaj et al. (2008) construct a conditional Kolmogorov test with a parametric rate by simulating at each moment of time paths that have a common starting value. The only restrictions required on the error density are those needed to guarantee that the estimator of the parameters of the conditional moments is consistent and asymptotically normal. Our proposed residual bootstrap procedure is different from that proposed by Manzan and Zerom (2008) who suggest using PITs based on non-parametric bootstrap replicates obtained from kernel densities instead of the residual bootstrap. The non-parametric bootstrap depends crucially on several smoothing parameters. Furthermore, they test separately for uniformity, using a Kolmogorov-Smirnov test, and for independence, using a Lagrange Multiplier test for linear dependence, while we propose a unique test for the adecuacy of the forecast densities. It is also important to point out that, the bootstrap procedure proposed in this paper allows for the incorporation of parameter uncertainty and can be extended to multivariate systems. We show that the finite sample distributions of the bootstrapped G-ACR (BG-ACR) tests are well approximated using standard asymptotic distributions. The proposed approach is very easy to implement and particularly useful to evaluate forecast densities when the error distribution is unknown. Furthermore, it is possible to use graphical devices that are visual aids to the formal results of the tests. When the PITs are uniform and independent, the pairs of PITs are expected to be evenly spread over the unit cube. When this is not the case, the visual aid should confirm the formal results of the test about rejection of the null. In general, it is not possible to identify the source of misspecification when the null hypothesis is rejected because the alternative is just written as the negation of the null. The graphical tool may offer some clues. We present some experiments that confirm that when the rejection of the null is due to mis-specified linear dependence, we observe the pairs of PITs lined up across the diagonals of the unit cube. When the rejection of the null is due to mis-specified conditional heteroscedasticity, the pair of PITs tend to concentrate around the vertexes of the unit cube.

Our second contribution is the implementation of the proposed BG-ACR tests to evaluate the specification of the Heterogeneous Autoregressive (HAR) model and of the Multiplicative Error Model (MEM), proposed to represent the dynamic evolution of the daily forward-looking market volatility index (VIX) from the Chicago Board Options Exchange (CBOE). After implementing

the BG-ACR tests, we show that the HAR and MEM specifications are both rejected if they do not incorporate conditional heteroscedasticity of the VIX itself. Furthermore, we also show that normality of the errors of the HAR model and the semiparametric Gamma distribution (GSNP) of the errors of the MEM model are also rejected.

The rest of the paper is organized as follows. Section 2 briefly describes the G-ACR tests. Section 3 contains the main contribution with the description of the new proposed BG-ACR tests and the analysis of their in-sample performance. In section 4, we analyze their out-of-sample performance. In section 5, we offer an empirical application to illustrate the advantages of the BG-ACR tests by testing for the adequacy of the HAR and MEM models to obtain forecast densities of the VIX index. Finally, we conclude in section 6.

2. The Generalized-AutoContouR (G-ACR) test

We briefly describe the G-ACR test proposed by González-Rivera and Sun (2015) to facilitate the reading of the forthcoming sections and to make the exposition self-contained.

Let $\{y_t\}_{t=1}^{T}$ be a strictly stationary univariate random process with finite marginal variance and conditional density function $f_t(y_t|Y_{t-1})$, where $Y_{t-1} = (y_1, ..., y_{t-1})$ is the information set available at time $t-1$. A conditional density model is constructed by specifying the conditional mean, conditional variance or other conditional moments of interest, and making distributional assumptions on the functional form of $f_t(y_t|Y_{t-1})$. Based on the conditional model, the researcher might construct a density forecast denoted by $g_t(y_t|Y_{t-1})$ and obtain a sequence of PITs of $\{y_t\}_{t=1}^{T}$ w.r.t. $g_t(y_t|Y_{t-1})$ as given by $u_t = \int_{-\infty}^{y_t} g_t(v_t|Y_{t-1})\, dv_t$. If $g_t(y_t|Y_{t-1})$ coincides with the true conditional density, $f_t(y_t|Y_{t-1})$, then the sequence of PITs, $\{u_t\}_{t=1}^{T}$, must be i.i.d. $U(0, 1)$; see Rosenblatt (1952).

Following González-Rivera and Sun (2015), for lag $k = 1, 2, ...,$ we define

$$\mathrm{G-ACR}_{k,\alpha_i} = \{(u_t, u_{t-k}) \subset \Re^2 | 0 \le u_t \le \sqrt{\alpha_i} \quad \text{and} \quad 0 \le u_{t-k} \le \sqrt{\alpha_i}, s.t. : u_t \times u_{t-k} \le \alpha_i\}, \quad (1)$$

and the indicator series I_t^{k,α_i} that takes value one if $(u_t, u_{t-k}) \in \mathrm{G-ACR}_{k,\alpha_i}$ and zero otherwise, where α_i is the population probability level. Consider the following statistic

$$t_{k,\alpha_i} = \frac{\sqrt{T-k}(\hat{\alpha}_{k,i} - \alpha_i)}{\sigma_{\alpha_i}}, \quad (2)$$

where $\hat{\alpha}_{k,i} = \frac{\sum_{t=k+1}^{T} I_t^{k,\alpha_i}}{T-k}$ is the sample proportion of PIT pairs within the $\mathrm{G-ACR}_{k,\alpha_i}$ cube and $\sigma_{\alpha_i}^2 = \alpha_i(1-\alpha_i) + 2\alpha_i^{3/2}(1-\alpha_i^{1/2})$. This variance can be derived taking into account that the indicator variable, I_t^{k,α_i}, is a Bernoulli random variable. The variable (u_t, u_{t-k}) falls inside $\text{G-ACR}_{k,\alpha_i}$ with probability α_i. Furthermore, I_t^{k,α_i} follows an Moving Average process whose order depends on k. Finally, González-Rivera and Sun (2015) show that, under the null hypothesis of u_t being i.i.d. U(0,1), t_{k,α_i} is asymptotically standard normal distributed.

The t-statistic in (2) is constructed for a single fixed autocontour, α_i, and a single fixed lag, k. However, it can be generalized to a set of lags with a fixed autocontour or to several autocontours with a fixed lag. In the first case, for a fixed autocontour α_i, define $L_{\alpha_i} = (\ell_{1,\alpha_i}, ..., \ell_{K,\alpha_i})'$ which is a $K \times 1$ stacked vector with element $\ell_{k,\alpha_i} = \sqrt{T-k}(\hat{\alpha}_{k,i} - \alpha_i)$. For economic and financial data, conventional wisdom suggest using a small K as serial correlation among PITs is often the strongest at small lags; see Hong and Li (2005) for a similar argument. Under the null, González-Rivera and Sun (2015) show that $L'_{\alpha_i}\Lambda_{\alpha_i}^{-1}L_{\alpha_i}$ is asymptotically χ_K^2 distributed, where a typical element of the asymptotic covariance matrix, Λ_{α_i}, is given by:

$$\lambda_{j,k} = \begin{cases} \alpha_i(1-\alpha_i) + 2\alpha_i^{3/2}(1-\alpha_i^{1/2}), & j = k, \\ 4\alpha_i^{3/2}(1-\alpha_i^{1/2}), & j \neq k. \end{cases}$$

Alternatively, for a fixed lag k, define the vector $C_k = (c_{k,1}, ..., c_{k,C})'$ with $c_{k,i} = \sqrt{T-k}(\hat{\alpha}_{k,i} - \alpha_i)$. Once more, under the null, $C_k' \Omega_k^{-1} C_k$ has asymptotically a χ^2_C distribution, where a typical element of the asymptotic covariance matrix, Ω_k, is given by:

$$\omega_{i,j} = \begin{cases} \alpha_i(1-\alpha_i) + 2\alpha_i^{3/2}(1-\alpha_i^{1/2}), & i = j, \\ \alpha_i(1-\alpha_j) + 2\alpha_i\alpha_j^{1/2}(1-\alpha_j^{1/2}), & i < j, \\ \alpha_j(1-\alpha_i) + 2\alpha_j\alpha_i^{1/2}(1-\alpha_i^{1/2}), & i > j. \end{cases}$$

The expression of the covariances in $\lambda_{j,k}$ and $\omega_{i,j}$ can be derived using the same arguments about the distribution of the indicator variables explained above; see González-Rivera and Sun (2015) for more details.

If the researcher is interested in partial aspects of the densities, such as, a particular collection of quantiles, it is more informative to examine the L_{α_i} statistic, which incorporates information for all desired k lags. On the other hand, if he is interested in the whole distribution, C_k collects information on all desired C autocontours for a given fixed lag k.

The tests described above are based on an assumed predictive density $g_t(y_t|Y_{t-1})$. However, in practice, the parameters associated with the moments of this density need to be estimated. González-Rivera and Sun (2015) analyze the effects of parameter estimation on the asymptotic distribution of t_{k,α_i}, and consequently on L_{α_i} and C_k, and conclude that the corresponding adjustments to the asymptotic variance are model dependent and thus, difficult to calculate analytically. To overcome this drawback, they propose a fully parametric bootstrap procedure to approximate the asymptotic variance based on obtaining random extractions from the known error predictive density assumed under the null hypothesis.

The G-ACR tests can be implemented both in-sample and out-of-sample. González-Rivera and Sun (2015) show that, when testing the out-of-sample specification, the importance of parameter uncertainty will depend on both the forecasting scheme and the size of the estimation sample (T) relative to the forecast sample (H). When implementing the tests to check the correct specification of the out-of-sample forecast densities, parameter uncertainty will distort the test size as long as the proportion of the out-of-sample and in-sample sizes, H and T, respectively, is large. However, under the assumption of $\sqrt{T}$-consistent estimators, if $T \to \infty, H \to \infty$ and $H/T \to 0$, parameter uncertainty is asymptotically negligible and no adjustment to the test is needed.

Finally, note that, if any of the G-ACR tests rejects the null hypothesis, there is no indication about whether the rejection can be attributed to an inadequate assumption about the error distribution or to misspecification of the conditional moments. González-Rivera and Sun (2015) point out that the G-ACR tests are more powerful for detecting departures from the distributional assumption than for detecting misspecified dynamics.

3. In-sample bootstrap G-ACR (BG-ACR) tests

We propose a generalization of the G-ACR tests that allows testing for the specification of the conditional moments without making any particular assumption on the conditional distribution. We justify heuristically the asymptotic validity of the proposed procedure and carry out Monte Carlo experiments to establish its finite sample performance.

3.1. Description of the BG-ACR tests

Consider the following parametric location-scale model for a univariate strictly stationary finite variance series of interest, y_t, $t = 1, ..., T$,

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

where μ_t and σ_t^2 are the conditional mean and variance of y_t, which are parametric functions of the information set, Y_{t-1}, and ε_t is an independent white noise process with distribution F_ε, such that $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = 1$. The parameters governing μ_t, σ_t^2 and F_ε guarantee stationarity and satisfy the conditions required for their estimators to be consistent and asymptotically normal. For example, Francq and Zakoian (2004) give conditions for the strong consistency and asymptotic normality of the Gaussian-Quasi-Maximum-Likelihood (G-QML) estimator of the ARMA-GARCH model and Mika and Saikkonen (2011) when both the conditional mean and the conditional variance are nonlinear. From now on, we consider the G-QML estimator of the parameters of the conditional mean and variance.

Without loss of generality and to illustrate the procedure, we consider the following popular AR(1)-GARCH(1,1) model

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + a_t, \\ a_t &= \varepsilon_t \sigma_t, \\ \sigma_t^2 &= \omega_0 + \omega_1 a_{t-1}^2 + \omega_2 \sigma_{t-1}^2, \end{aligned} \tag{4}$$

where $|\phi_1| < 1, \omega_1 + \omega_2 < 1, \omega_0 > 0$ and $\omega_1, \omega_2 \geq 0$ to guarantee the stationarity of y_t and the positiveness of the conditional variance. Note that the proposed procedure to obtain in-sample bootstrap conditional densities, and the consequent BG-ACR statistics to evaluate them, can be applied to any other parametric specifications of the conditional mean and conditional variance (and any other higher moments) as far as a consistent and asymptotically normal estimator of the parameters is available.

Next, we describe the proposed bootstrap algorithm to obtain in-sample one-step-ahead bootstrap conditional densities of y_t in the context of the AR(1)-GARCH(1,1) model in (4). The algorithm is based on the residual bootstrap algorithms of Pascual et al. (2004, 2006) for the construction of forecast densities in linear ARMA and GARCH models, respectively.

Step 1. Obtain the residuals. Obtain the G-QML estimates of the parameters: $\hat{\phi}_0$, $\hat{\phi}_1$, $\hat{\omega}_0, \hat{\omega}_1$ and $\hat{\omega}_2$. Obtain the standardized residuals $\hat{\varepsilon}_t = \frac{\hat{a}_t}{\hat{\sigma}_t}, t = 2, ..., T$, where $\hat{a}_t = y_t - \hat{\phi}_0 - \hat{\phi}_1 y_{t-1}, \hat{\sigma}_2^2 = \hat{\omega}_0/(1 - \hat{\omega}_1 - \hat{\omega}_2)$ and $\hat{\sigma}_t^2 = \hat{\omega}_0 + \hat{\omega}_1 \hat{a}_{t-1}^2 + \hat{\omega}_2 \hat{\sigma}_{t-1}^2$, for $t = 3, ..., T$. Denote by $\hat{F}_{\hat{\varepsilon}}$ the empirical distribution of the centered and scaled residuals.

Step 2. Obtain bootstrap replicates of parameter estimates. For $t = 3, ..., T$, obtain recursively a bootstrap replicate of y_t that mimics the dynamic dependence of the original series as follows

$$\sigma_t^{*2(b)} = \hat{\omega}_0 + \hat{\omega}_1 a_{t-1}^{*2(b)} + \hat{\omega}_2 \sigma_{t-1}^{*2(b)}, \tag{5}$$

$$\begin{aligned} a_t^{*(b)} &= \varepsilon_t^{*(b)} \sigma_t^{*(b)}, \\ y_t^{*(b)} &= \hat{\phi}_0 + \hat{\phi}_1 y_{t-1}^{*(b)} + a_t^{*(b)}, \end{aligned} \tag{6}$$

where $a_2^{*(b)} = \hat{a}_2, \sigma_2^{*2(b)} = \hat{\sigma}_2^2, y_2^{*(b)} = y_2$ and $\varepsilon_t^{*(b)}$ are random extractions with replacement from $\hat{F}_{\hat{\varepsilon}}$. Estimate the parameters by G-QML using $\left\{ y_t^{*(b)} \right\}_{t=3}^{T}$, obtaining $\hat{\phi}_0^{*(b)}$, $\hat{\phi}_1^{*(b)}$, $\hat{\omega}_0^{*(b)}, \hat{\omega}_1^{*(b)}$ and $\hat{\omega}_2^{*(b)}$.

Step 3 Obtain in-sample bootstrap one-step-ahead predictive densities. For $t = 3, ..., T$, obtain in-sample one-step-ahead estimates of volatilities and observations:

$$\sigma_t^{**2(b)} = \hat{\omega}_0^{*(b)} + \hat{\omega}_1^{*(b)}(y_{t-1} - \hat{\phi}_0^{*(b)} - \hat{\phi}_1^{*(b)} y_{t-2})^2 + \hat{\omega}_2^{*(b)} \sigma_{t-1}^{**2(b)}, \tag{7}$$

$$y_t^{**(b)} = \hat{\phi}_0^{*(b)} + \hat{\phi}_1^{*(b)} y_{t-1} + \sigma_t^{**(b)} \hat{\varepsilon}_t^{*(b)}, \tag{8}$$

where $\sigma_2^{**2(b)} = \hat{\omega}_0^{*(b)}/(1 - \hat{\omega}_1^{*(b)} - \hat{\omega}_2^{*(b)})$.

Step 4. Repeat steps 2 and 3 for $b = 1, ..., B^{(1)}$.

The residual-bootstrap procedure described above is based on separating the two constituents of the forecast errors, i.e. the estimation error and the innovation error. First, in step 2, we obtain replicates of y_t^* that are not conditional on Y_{t-1}. In (5), σ_t^{*2} depends on a_{t-1}^{*2} and in (6), y_t^* depends on y_{t-1}^*. Therefore, independent replicates of y_t are generated to estimate the parameter estimator sample distribution. There is a large literature on implementing the residual-bootstrap for estimating the parameter sample distribution; see, for example, Politis (2003). Second, in step 3, the bootstrap replicates, σ_t^{**2} and y_t^{**}, in (7) and (8) respectively, are obtained incorporating the parameter uncertainty through the bootstrap estimates of the parameters but always conditional on the original data $\{y_1, ..., y_{t-1}\}$. In this sense, the simulation scheme resembles that proposed by Bhardwaj et al. (2008) in the context of diffusion processes in which they simulate future paths that have a common starting value at time t. However, Bhardwaj et al. (2008) simulate the innovations from a normal variable. In our algorithm, at each moment of time, $t = 3, ..., T$, we generate $B^{(1)}$ bootstrap replicates of y_t conditional on Y_{t-1} incorporating parameter uncertainty and avoiding any specific assumption about the distribution of ε_t. In order to decide the number of bootstrap replicates that guarantees an appropriate estimate of the predictive density, one can implement the procedure proposed by Andrews and Buchinsky (2000).

In-sample PITs can be easily computed as follows

$$u_t = \frac{1}{B^{(1)}} \sum_{b=1}^{B^{(1)}} \mathbf{1}(y_t^{**(b)} < y_t), \tag{9}$$

where $\mathbf{1}(.)$ is the indicator function which takes value 1 when the argument is true and zero otherwise. After computing the corresponding indicators, I_t^{k,α_i}, and sample proportions, $\hat{\alpha}_{k,i}$, the $t_{k,\alpha_i}^*, L_{\alpha_i}^*$ and C_k^* statistics are calculated and the asymptotics of Section 2 are applied.[1]

In order to illustrate how the proposed procedure works, we have generated a time series of size $T = 5000$ from the following homoscedastic AR(1) model:

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \tag{10}$$

with $\phi_1 = (0.5, 0.95)$ and i.i.d. ε_t either N(0,1), or centered and standardized Student-5, or $\chi_{(5)}^2$. In each case, an AR(1) model is fitted to the artificial series with the parameters estimated by G-QML. Then, in-sample PITs are computed (i) assuming normal errors as in González-Rivera and Sun (2015) and (ii) implementing the bootstrap algorithm described above based on $B^{(1)} = 999$

[1]Following the suggestion of González-Rivera and Sun (2015), the variance of $\hat{\alpha}_{k,i}$ is approximated using a bootstrap procedure that takes into account parameter uncertainty. $B^{(2)}$ bootstrap replicates, $\{y_t^{*(b)}\}_{t=1}^T$ are generated as in (6) and $\hat{\alpha}_{k,i}^{*(b)}$ is obtained using the bootstrap series as if they were the original series.

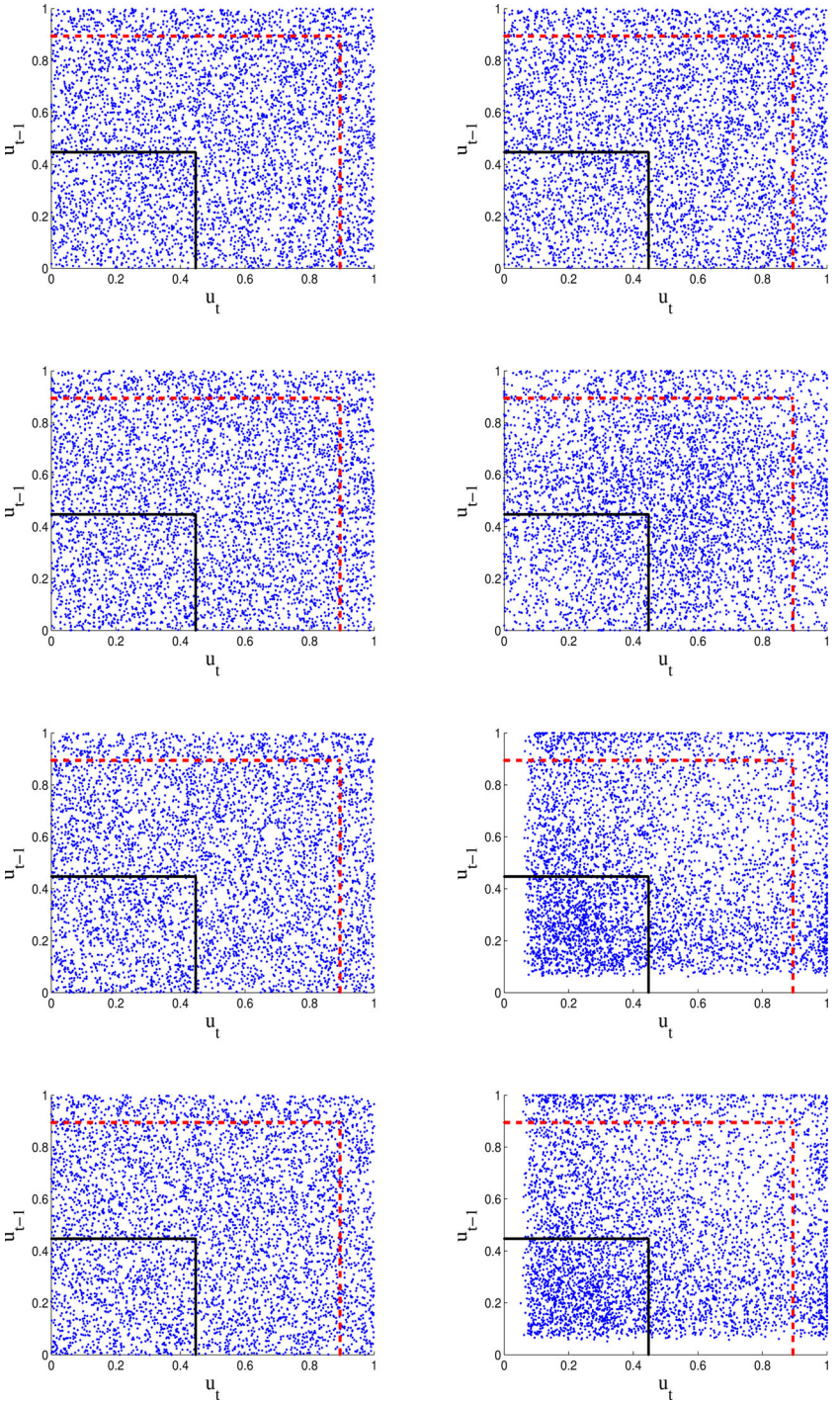

Figure 1. Pairs (u_t, u_{t-1}) and autocontours for the estimated AR(1) model with $T = 5000$. $ACR_{0.2,1}$ corresponds to the black (continuous) box and the $ACR_{0.8,1}$ to the red (discontinuous) box. The DGPs are the AR(1) model with: $\phi_1 = 0.5$ and $\varepsilon_t \sim N(0,1)$ (first row); $\phi_1 = 0.5$ and $\varepsilon_t \sim$ Student $-$ 5 (second row); $\phi_1 = 0.5$ and $\varepsilon_t \sim \chi^2_{(5)}$ (third row); and $\phi_1 = 0.95$ and $\varepsilon_t \sim \chi^2_{(5)}$ (fourth row). The PITs are computed using the bootstrap algorithm with $B^{(1)} = 999$ (first column), or assuming Gaussian errors (second column).

replicates. In Fig. 1, we plot the autocontours for $\alpha_i = 0.2$ and 0.8 together with the pairs (u_t, u_{t-1}) for the AR(1) model with $\phi_1 = 0.5$ and $\varepsilon_t \sim N(0,1)$ (first row); $\phi_1 = 0.5$ and $\varepsilon_t \sim$ Student-5 (second row); $\phi_1 = 0.5$ and $\varepsilon_t \sim \chi^2_{(5)}$ (third row); and $\phi_1 = 0.95$ and $\varepsilon_t \sim \chi^2_{(5)}$ (fourth row). Note that, when the PITs are computed using the bootstrap densities (first column), they are uniformly distributed on the surface regardless of the true error distribution of the underlying DGP. Therefore, they suggest that the fitted AR(1) model is adequate. However, when the PITs are computed as in the G-ACR procedure (second column), assuming normality, they are not uniformly distributed unless the errors are Gaussian. In this case, when the model is rejected, there is no indication about whether the rejection is coming from the misspecification of the conditional mean or from a misspecified functional form of the error distribution.

Consider now the following three DGPs, from which we generate three time series

$$y_t = 0.3y_{t-1} + 0.6y_{t-2} + \varepsilon_t, \tag{11}$$

$$y_t = \begin{cases} 0.5y_{t-1} + \varepsilon_t, & \text{for} \quad t < T/2, \\ 1 + 0.5y_{t-1} + \varepsilon_t, & \text{for} \quad t \geq T/2, \end{cases} \tag{12}$$

$$\begin{aligned} y_t &= 0.5y_{t-1} + \varepsilon_t\sigma_t, \\ \sigma_t^2 &= 0.05 + 0.5\varepsilon_{t-1}^2\sigma_{t-1}^2 + 0.45\sigma_{t-1}^2, \end{aligned} \tag{13}$$

with ε_t defined as above. We fit an AR(1) model to each of the simulated series and estimate its parameters by G-QML. As above, we compute the PITs both assuming normal errors and using the proposed bootstrap procedure. In Fig. 2, we plot the autocontours for $\alpha_i = 0.2$ and 0.8 together with the pairs (u_t, u_{t-1}) when the DGP is the AR(2) model in (11) with $\chi^2_{(5)}$ errors (first row); the AR(1) model with structural break in the mean in (12) with $\varepsilon_t \sim \chi^2_{(5)}$ (second row); the GARCH model in (13) with normal errors (third row); and the GARCH model in (13) with $\chi^2_{(5)}$ errors (fourth row). We observe that, when the PITs are based on bootstrap densities (first column), they suggest the source of the misspecification. In the first row, when the AR(1) model is fitted to the AR(2) series, we observe a linear relation between the PITs, which tend to group around one of the diagonals of the unit-square. In the second row, when the DGP is the AR(1) model with a break in the mean, the PITs do not show any particular linear or non-linear relationship but they are concentrated on the top-right corner of the unit-square. Finally, when the DGP is the AR(1)-GARCH(1,1) model, we observe a non-linear relation between the PITs, which are more concentrated toward the four corners of the unit-square. In this last case, the autocontour plots are very similar regardless of the error distribution. Comparing the bootstrap-based PITs with those obtained using G-ACR assuming a normal density (second column), the rejection of the fitted AR(1) model is also evident. However, there is not an obvious indication of the source of the misspecification.

3.2. Asymptotic validity: some heuristic arguments

The asymptotic distributions of t^*_{k,α_i}, $L^*_{\alpha_i}$ and C^*_k depend on the asymptotic validity of the residual bootstrap algorithm. Next, we discuss the validity of such procedure.

First, for the bootstrap procedure to be asymptotically valid, it has to be valid for the distribution of the estimator of the conditional mean and variance parameters. In the context of stationary linear ARMA models, the validity of the residual-bootstrap advocated in this paper for the QML estimator was established early in the literature; see, for example, the survey by Kreiss and Lahiri (2012) and the references therein. In the context of nonlinear ARCH-GARCH models, Hall and Yao (2003) show that asymptotic normality of the parameter estimator is a requirement for the bootstrap to be asymptotically valid for the estimation of its sample distribution. As

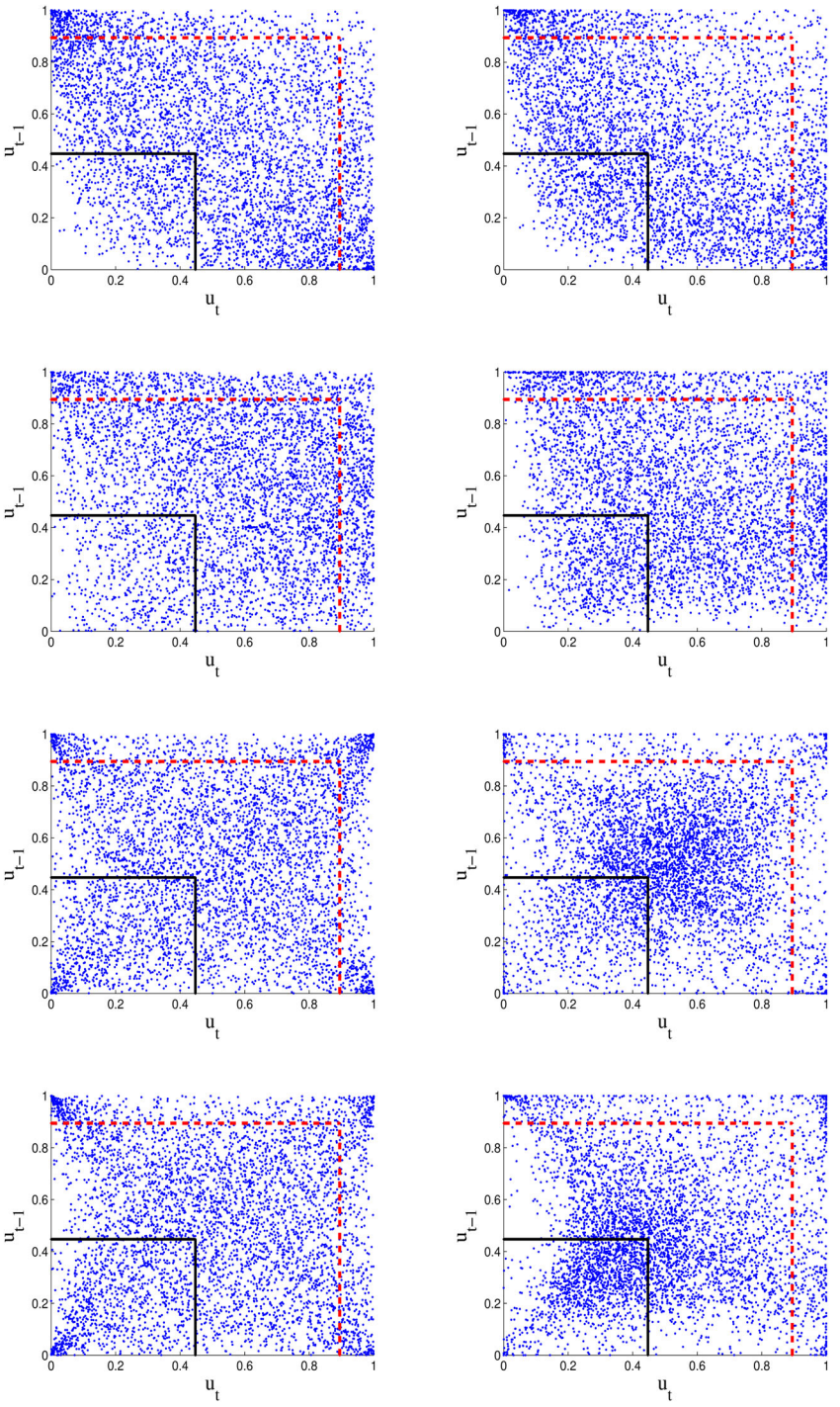

Figure 2. Pairs (u_t, u_{t-1}) and autocontours for estimated AR(1) model with $T = 5000$. $ACR_{0.2,1}$ corresponds to the black (continuous) box and the $ACR_{0.8,1}$ to the red (discontinuous) box. The DGPs are: AR(2) with $\varepsilon_t \sim \chi^2_{(5)}$ (first row); AR(1) model with break in the mean with $\varepsilon_t \sim \chi^2_{(5)}$ (second row); AR(1)-GARCH(1,1) model with $\varepsilon_t \sim$ N(0,1) (third row); and AR(1)-GARCH(1,1) model with $\varepsilon_t \sim \chi^2_{(5)}$ (fourth row). The PITs are computed using the bootstrap algorithm with $B^{(1)} = 999$ (first column), or assuming Gaussian errors (second column).

mentioned above, in this paper, we are implementing the G-QML estimator assuming that the conditions for its consistency and asymptotic validity are satisfied. Perera and Silvapulle (2018) proves the validity of the residual bootstrap for the G-QML estimator of the parameters of ARMA-GARCH models; see also Hidalgo and Zaffaroni (2007) who prove the first order validity of the residual bootstrap for the parameters of an ARCH(∞) process characterized by a particular decay in the ARCH parameters and Jeong (2017) who proves the asymptotic validity of the bootstrap procedure for the parameters of GARCH models.

Second, once the asymptotic validity of the bootstrap parameter estimator (step 2) is established, one needs to look at the validity of the boostrap procedure to construct predictive densities in step 3. In the case of stationary linear ARMA models, Pascual et al. (2004) shows that the bootstrap is asymptotically valid to obtain predictive densities. However, as far as we know, there is not a formal proof of the validity of the residual-bootstrap procedure to construct predictive densities of nonlinear GARCH models. Having said that, several related results can be called to support heuristically the validity of the residual-bootstrap proposed in this paper. First, in the context of the closely related MEM models, Perera and Silvapulle (2019), propose a similar boostrap procedure in which the bootstrap residuals are drawn with replacement from the assumed conditional distribution with the parameters substituted by the corresponding estimates. They formaly prove the asymptotic validity of this closely bootstrap procedure. Alternatively, in the context of ARCH models, Kless (2019) formaly proves the asymptotic validity of the bootstrap procedure proposed in this paper in the context of ARCH models when the bootstrap innovations are drawn from a kernel smoothed density instead of drawing them from the empirical distribution. The Monte Carlo experiments carried out in Kless (2019) show that the results are the same regardless of whether the bootstrap innovations are drawn from the smoothed kernel or from the empirical distribution. In any case, if the bootstrap procedure is asymptotically valid for the estimation of the parameters, using the arguments in Pascual et al. (2004) and Reeves (2005), one can establish its validity for the predictive densities.

3.3. Finite sample performance of in-sample tests

We perform Monte Carlo simulations to assess the finite sample properties of the proposed statistics. For the size assessment, the DPG is a linear AR(1). We consider a model far from the non-stationary region and another one near the non-stationary region with different error distributions. For the power assessment, we consider linear and non-linear alternatives. The number of Monte Carlo replicates is $R = 1000$ and the sample size $T = 50,\ 100,\ 300,\ 1000$ and 5000. The number of bootstrap replicates is $B^{(1)} = 1000$, except for $T = 5000$, when we use $B^{(1)} = 2000$. Finally, the number of bootstrap replicates used to compute the variance of $\hat{\alpha}_{k,i}, L^*_{\alpha_i}$ and C^*_k is $B^{(2)} = 500$.

To investigate the size of the tests, we consider as DGP the AR(1) in (10). For each Monte Carlo replicate, we compute the proportions $\hat{\alpha}_{k,i}$, for $k = 1, ..., 5$, and their bootstrap variances. Then, we compute the Monte Carlo averages and standard deviations of $\hat{\alpha}_{k,i}$, together with the averages of the bootstrap standard deviations and the percentage of rejections of the null hypothesis when the nominal size of the test is 5%. Table 1 reports the Monte Carlo results for $k = 1$ when $\phi_1 = 0.95$ and the errors are $\chi^2_{(5)}$. We observe that, even for $T = 50$, the Monte Carlo averages of $\hat{\alpha}_{k,i}$ are rather close to α_i and that, for moderate sample sizes, the average of the bootstrap standard deviations is a good approximation to the Monte Carlo standard deviation of $\hat{\alpha}_{k,i}$. For relatively small sample sizes, the bootstrap standard deviations tend to overestimate the empirical standard deviations of $\hat{\alpha}_{k,i}$, mainly for the largest quantiles. Consequently, the size of the t^*_{1,α_i} statistic is smaller than the nominal. As the sample size increases, the percentage of rejections becomes rather close to the 5% nominal level. Therefore, asymptotic normality is a good

Table 1. For each sample size, T, the table reports the Monte Carlo average and standard deviation of $\hat{\alpha}_{k,i}$ (first two rows) together with the Monte Carlo average of the bootstrap estimated standard deviation, $\bar{\sigma}^*_{\alpha_i}$ (third row), and the size of the t^*_{1,α_i} test (fourth row).

T	α_i	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
50														
	$\hat{\alpha}_{k,i}$	0.015	0.058	0.109	0.209	0.307	0.407	0.504	0.603	0.705	0.804	0.901	0.950	0.984
	Std	(0.022)	(0.048)	(0.067)	(0.089)	(0.098)	(0.102)	(0.097)	(0.094)	(0.081)	(0.064)	(0.043)	(0.034)	(0.023)
	$\bar{\sigma}^*_{\alpha_i}$	0.022	0.049	0.068	0.090	0.100	0.103	0.101	0.095	0.086	0.072	0.054	0.045	0.032
	Size	0.061	0.046	0.026	0.022	0.015	0.023	0.016	0.025	0.013	0.012	0.001	0.004	0.009
100														
	$\hat{\alpha}_{k,i}$	0.012	0.055	0.106	0.205	0.305	0.406	0.503	0.602	0.702	0.803	0.900	0.949	0.989
	Std	(0.014)	(0.032)	(0.045)	(0.060)	(0.064)	(0.067)	(0.062)	(0.057)	(0.049)	(0.038)	(0.027)	(0.020)	(0.012)
	$\bar{\sigma}^*_{\alpha_i}$	0.015	0.033	0.046	0.060	0.066	0.067	0.065	0.060	0.053	0.043	0.032	0.025	0.018
	Size	0.060	0.037	0.029	0.025	0.021	0.025	0.014	0.012	0.014	0.014	0.008	0.005	0.000
300														
	$\hat{\alpha}_{k,i}$	0.011	0.052	0.102	0.202	0.303	0.402	0.502	0.601	0.701	0.800	0.899	0.949	0.988
	Std	(0.007)	(0.017)	(0.024)	(0.030)	(0.033)	(0.032)	(0.032)	(0.028)	(0.024)	(0.018)	(0.013)	(0.009)	(0.006)
	$\bar{\sigma}^*_{\alpha_i}$	0.008	0.017	0.024	0.031	0.034	0.034	0.032	0.030	0.026	0.020	0.014	0.011	0.007
	Size	0.044	0.036	0.033	0.039	0.034	0.022	0.032	0.024	0.031	0.017	0.026	0.018	0.011
1000														
	$\hat{\alpha}_{k,i}$	0.011	0.051	0.101	0.201	0.301	0.401	0.501	0.600	0.700	0.800	0.899	0.949	0.988
	Std	(0.004)	(0.009)	(0.012)	(0.016)	(0.017)	(0.017)	(0.016)	(0.015)	(0.012)	(0.009)	(0.006)	(0.004)	(0.003)
	$\bar{\sigma}^*_{\alpha_i}$	0.004	0.009	0.012	0.016	0.017	0.017	0.016	0.015	0.012	0.010	0.007	0.005	0.003
	Size	0.054	0.048	0.046	0.051	0.039	0.040	0.042	0.048	0.045	0.034	0.037	0.043	0.101
5000														
	$\hat{\alpha}_{k,i}$	0.010	0.050	0.101	0.200	0.300	0.400	0.500	0.600	0.700	0.799	0.900	0.950	0.989
	Std	(0.002)	(0.004)	(0.005)	(0.007)	(0.008)	(0.007)	(0.007)	(0.006)	(0.005)	(0.004)	(0.002)	(0.002)	(0.001)
	$\bar{\sigma}^*_{\alpha_i}$	0.002	0.004	0.005	0.007	0.007	0.007	0.007	0.006	0.005	0.004	0.002	0.002	0.001
	Size	0.049	0.063	0.055	0.046	0.054	0.039	0.049	0.046	0.051	0.044	0.051	0.056	0.162

The DGP is $y_t = 0.95y_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \chi^2_{(5)}$ and the nominal size is 5%.

approximation to the finite sample distribution of the proposed BG-ACR test under the null of correct specification as far as we do not consider extreme autocontours. This conclusion is valid regardless of the particular error distribution and the persistence properties of the conditional mean.[2]

To study the finite sample power of the tests, we generate replicates using the models in Eqs. (12) and (13). In both cases, we fit an AR(1) model. Under the null hypothesis, we test the correct specification of the AR(1) model without drift. For the DGP in (12), we analyze their power against breaks in the conditional mean while for the DGP in (13), we study their power against misspecification in the conditional variance. In Table 2, we report the power results corresponding to the portmanteau tests. Both $L^{5*}_{\alpha_i}$ and C^*_1 are very powerful for detecting breaks in the conditional mean when the sample size is 300 and above. Detecting misspecification in the conditional variance is more difficult in small samples and we need sample sizes beyond 1000 observations to obtain high power.[3] As with the t^*_{1,α_i}, the power of $L^{5*}_{\alpha_i}$ is higher in the extreme autocontours.[4]

4. Out-of-sample *h*-step-ahead BG-ACR tests

We extend the procedures and tests described in the previous section to obtain out-of-sample h-step-ahead densities. In order to compute the proportion $\hat{\alpha}_{k,i}$, it is necessary to obtain

[2]Results for the AR(1) model with $\phi_1 = 0.5$ and Gaussian errors are reported in Tables A and B of the supplementary material.
[3]Note that this result is expected as inference in nonlinear GARCH models requires large samples.
[4]Results on the power when the DGP is the AR(2) model in (11) are reported in Tables C and D of the supplementary material. The proposed tests are very powerful even for small sample sizes.

Table 2. Monte Carlo power results for $L_{\alpha_i}^{5*}$ and C_1^* statistics.

	$L_{0.01}^{5*}$	$L_{0.05}^{5*}$	$L_{0.1}^{5*}$	$L_{0.2}^{5*}$	$L_{0.3}^{5*}$	$L_{0.4}^{5*}$	$L_{0.5}^{5*}$	$L_{0.6}^{5*}$	$L_{0.7}^{5*}$	$L_{0.8}^{5*}$	$L_{0.9}^{5*}$	$L_{0.95}^{5*}$	$L_{0.99}^{5*}$	C_1^{13*}
Panel A														
50	0.000	0.006	0.019	0.063	0.121	0.155	0.205	0.257	0.269	0.280	0.257	0.300	0.240	0.054
100	0.002	0.014	0.047	0.131	0.256	0.292	0.326	0.362	0.379	0.375	0.391	0.404	0.186	0.166
300	0.004	0.339	0.586	0.789	0.869	0.891	0.900	0.891	0.879	0.839	0.726	0.591	0.276	0.855
1000	0.743	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.984	0.676	1.000
5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Panel B														
50	0.177	0.093	0.064	0.054	0.041	0.040	0.045	0.031	0.055	0.085	0.122	0.180	0.052	0.059
100	0.301	0.104	0.076	0.065	0.051	0.050	0.055	0.057	0.056	0.095	0.207	0.279	0.061	0.107
300	0.589	0.175	0.071	0.064	0.061	0.063	0.074	0.084	0.088	0.143	0.282	0.473	0.314	0.381
1000	0.935	0.366	0.144	0.090	0.088	0.091	0.106	0.161	0.238	0.331	0.520	0.653	0.886	0.907
5000	0.999	0.875	0.345	0.166	0.154	0.187	0.332	0.557	0.770	0.890	0.940	0.941	0.972	1.000

The DGPs are: AR(1) model with break in the mean (Panel A) and AR(1)-GARCH(1,1) (Panel B). The nominal size is 5%.

$(H-h+1)$ h-step-ahead bootstrap forecast densities. In this section, we are using a fixed scheme, i.e. the parameters are estimated only once.[5] If the parameters are not re-estimated each time a new observation is available, then the in-sample algorithm can be implemented as described in Section 3 with step 3 modified as follows:

Step 3′. Obtain out-of-sample *h*-step-ahead bootstrap forecast densities. For $h = 1, 2, \ldots$ and $j = 0, \ldots, H-h$ obtain out-of-sample h-step-ahead conditional estimates of volatilities and observations as follows:

$$\begin{aligned}\sigma_{T+h+j|T+j}^{**2(b)} &= \hat{\omega}_0^{*(b)} + \hat{\omega}_1^{*(b)}\left(y_{T+h-1+j|T+j}^{**(b)} - \hat{\phi}_0^{*(b)} - \hat{\phi}_1^{*(b)} y_{T+h-2+j|T+j}\right)^2 + \hat{\omega}_2^{*(b)} \sigma_{T+h-1+j|T+j}^{**2(b)}, \\ y_{T+h+j|T+j}^{**(b)} &= \hat{\phi}_0^{*(b)} + \hat{\phi}_1^{*(b)} y_{T+h-1+j|T+j}^{**(b)} + \sigma_{T+h+j|T+j}^{**(b)} \varepsilon_{T+h}^{*(b)},\end{aligned} \tag{14}$$

where $y_{T+i|T}^{**(b)} = y_{T+i}$ when $i \leq 0$ and

$$\sigma_{i|i}^{**2(b)} = \frac{\hat{\omega}_0^{*(b)}}{1-\hat{\omega}_1^{*(b)} - \hat{\omega}_2^{*(b)}} + \hat{\omega}_1^{*(b)} \sum_{j=0}^{i-3} \hat{\omega}_2^{*(b)j} \left[\left(y_{i-j-1} - \hat{\phi}_0^{*(b)} - \hat{\phi}_1^{*(b)} y_{i-j-2}\right)^2 - \frac{\hat{\omega}_0^{*(b)}}{1-\hat{\omega}_1^{*(b)} - \hat{\omega}_2^{*(b)}} \right]$$

for $i = T, T+1, \ldots, T+H-1$.
At each moment $T+j$, $j = h, \ldots, H$, the out-of-sample multi-period PITs are

$$u_{T+j|T+j-h} = \frac{1}{B^{(1)}} \sum_{b=1}^{B^{(1)}} \mathbf{1}\left(y_{T+j|T+j-h}^{**(b)} < y_{T+j}\right).$$

Note that, when $h > 1$, under the null that the predictive density coincides with the true density, the PITs are still uniformly distributed but they are expected to be dependent. As a result, it is common in the literature to test the null of a well behaved density forecast by choosing PITs separated by h periods to ensure an independent sequence of PITs. This procedure may significantly reduce the evaluation sample when h is relatively large. In this case, the procedure can be implemented in several uncorrelated sub-samples of forecasts that are h periods apart and then use Bonferroni methods to obtain a joint test without discarding observations; see, for example, Diebold et al. (1998), Clements and Smith (2000), Manzan and Zerom (2008) and Rossi and Sekhposyan (2014), among others. Alternatively, Rossi and Sekhposyan (2019) propose bootstrapping the h-step-ahead PITs.

[5]The effect of the forecasting scheme is an interesting question to be developed in further research. In the fully parametric autocontour context, González-Rivera and Sun (2017) provide an analysis of the effects of the forecasting schemes (fixed, rolling, and recursive) on the size and power of autocontour-based tests.

Table 3. Monte Carlo size results for out-of-sample $L^{5*}_{\alpha_i}$ and C^*_1 statistics.

	$L^{5*}_{0.01}$	$L^{5*}_{0.05}$	$L^{5*}_{0.1}$	$L^{5*}_{0.2}$	$L^{5*}_{0.3}$	$L^{5*}_{0.4}$	$L^{5*}_{0.5}$	$L^{5*}_{0.6}$	$L^{5*}_{0.7}$	$L^{5*}_{0.8}$	$L^{5*}_{0.9}$	$L^{5*}_{0.95}$	$L^{5*}_{0.99}$	C^{13*}_1
T							**Panel A**							
50	0.116	0.113	0.096	0.083	0.070	0.072	0.090	0.091	0.108	0.108	0.121	0.117	0.147	0.086
100	0.100	0.075	0.084	0.050	0.039	0.051	0.062	0.080	0.107	0.105	0.116	0.136	0.094	0.078
300	0.120	0.080	0.068	0.059	0.066	0.073	0.069	0.070	0.077	0.091	0.118	0.139	0.087	0.071
1000	0.124	0.081	0.075	0.072	0.067	0.063	0.079	0.075	0.090	0.103	0.126	0.151	0.074	0.079
5000	0.093	0.077	0.054	0.058	0.053	0.062	0.063	0.062	0.073	0.079	0.116	0.161	0.090	0.064
							Panel B							
50	0.119	0.092	0.088	0.087	0.082	0.087	0.085	0.070	0.076	0.084	0.093	0.107	0.107	0.057
100	0.100	0.076	0.079	0.066	0.078	0.067	0.065	0.060	0.073	0.074	0.102	0.111	0.115	0.047
300	0.094	0.069	0.070	0.057	0.056	0.052	0.066	0.059	0.066	0.059	0.081	0.102	0.197	0.062
1000	0.074	0.063	0.059	0.055	0.060	0.059	0.065	0.056	0.075	0.068	0.081	0.090	0.164	0.065
5000	0.056	0.054	0.049	0.058	0.047	0.058	0.057	0.053	0.053	0.051	0.077	0.113	0.127	0.050

The DGP is $y_t = 0.95y_{t-1} + \varepsilon_t$ and $\varepsilon_t \sim N(0,1)$. The nominal size is 5%, $H = 50$ (Panel A) and $H = 500$ (Panel B).

Using the independent PITs $\{u_{T+ht|T+h(t-1)}\}_{t=1}^{[H/h]}$, we compute the corresponding indicators, I^{k,α_i}_{T+ht}. The t-statistic is given by

$$t_{k,\alpha_i} = \frac{\sqrt{H/h-k}(\hat{\alpha}_{k,i} - \alpha_i)}{\sigma_{\alpha_i}},$$

where $\hat{\alpha}_{k,i} = \frac{\sum_{t=k+1}^{[H/h]} I^{k,\alpha_i}_{T+ht}}{H/h-k}$ and $\sigma^2_{\alpha_i}$ is defined as in (2). Note that $\sigma^2_{\alpha_i}$ can be estimated either as in expression (2) or by bootstrapping. When testing the out-of-sample specification, the importance of parameter uncertainty decreases as far $H/T \to 0$ when $T \to \infty$ and $H \to \infty$. Therefore, if H is small relative to T, one can compute the variance $\sigma^2_{\alpha_i}$ by using the asymptotic expression.

As an illustration of the out-of-sample one-step-ahead performance of the tests, we generate $R = 1000$ replicates from the AR(1) model in expression (10) with $\phi_1 = 0.95$ and $\varepsilon_t \sim N(0,1)$. The model is estimated once by OLS using $T = 50$, 100, 300, 1000 and 5000 observations and $H = 50$ and 500 out-of-sample one-step-ahead densities obtained using a fixed scheme. Their corresponding PITs are obtained using the bootstrap procedure. The variance of $\hat{\alpha}_{k,i}$ and the covariances in Λ_{α_i} and Ω_k are computed by bootstrapping.[6] In Table 3, we report the size of the corresponding $L^5_{\alpha_i}$ and C^{13}_1 test statistics for $H = 50$ and $H = 500$. Increasing H improves the size properties of the tests as far as the ratio H/T is still small. For small estimation samples, the tests tend to be oversized but the size is corrected when the estimation and evaluation samples are larger.[7]

Finally, we study the finite sample power of the out-of-sample one-step-ahead tests. With this purpose, we generate $R = 1000$ replicates from the AR(1)-GARCH(1,1) model in (13). Under the null hypothesis, we estimate an AR(1) process without drift. We report the power results of the t_{1,α_i} tests in Table 4 with $H = 500$. We observe a similar behavior as in the in-sample tests. The information on heteroscedasticity is contained in the lower 1% and 5% autocontours and large estimation samples are required. In any case, it is important to note that in-sample tests are expected to be more powerful than out-of-sample tests. Inoue and Kilian (2005) conclude that results of in-sample tests will typically be more credible than results of out-of-sample tests.

Although in this section, we have analyzed the performance of the tests for out-of-sample one-step-ahead densities, we expect the same behavior for out-of-sample h-step-ahead forecasts as far

[6]Results based on the asymptotic expression of the variances and covariances are very similar when $H = 50$ and $T = 1000$ ($H/T = 0.05$) or $T = 5000$ ($H/T = 0.01$). When $H = 500$, the results are similar if $T = 5000$ ($H/T = 0.1$). As mentioned above, in these cases, the parameter uncertainty is irrelevant. These results are available upon request.

[7]Results for the t-tests are reported in Table E of the supplementary material. For small estimation sizes, the test tends to be oversized for the middle autocontours. When T is relatively large and H/T is small, the empirical size is about 5%.

Table 4. For each sample size, T, the table reports the Monte Carlo average and standard deviation of $\hat{\alpha}_{k,i}$ (first two rows) together with the Monte Carlo average of the bootstrap estimated standard deviation, $\bar{\sigma}^*_{\alpha_i}$ (third row), and the power of the t^*_{1,α_i} test (fourth row).

T	α_i	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
50														
	$\hat{\alpha}_{k,i}$	0.030	0.070	0.115	0.213	0.323	0.432	0.537	0.630	0.718	0.800	0.875	0.912	0.943
	Std	(0.022)	(0.036)	(0.047)	(0.063)	(0.082)	(0.098)	(0.109)	(0.114)	(0.110)	(0.100)	(0.083)	(0.069)	(0.054)
	$\bar{\sigma}^*_{\alpha_i}$	0.011	0.027	0.038	0.053	0.062	0.068	0.070	0.071	0.069	0.064	0.053	0.044	0.033
	Power	0.388	0.183	0.119	0.090	0.118	0.148	0.198	0.219	0.243	0.248	0.193	0.228	0.289
100														
	$\hat{\alpha}_{k,i}$	0.028	0.069	0.114	0.210	0.314	0.421	0.525	0.625	0.719	0.808	0.889	0.929	0.963
	Std	(0.019)	(0.030)	(0.039)	(0.053)	(0.064)	(0.079)	(0.087)	(0.091)	(0.090)	(0.084)	(0.068)	(0.054)	(0.037)
	$\bar{\sigma}^*_{\alpha_i}$	0.009	0.021	0.032	0.045	0.052	0.057	0.058	0.057	0.054	0.049	0.040	0.031	0.020
	Power	0.464	0.228	0.128	0.059	0.078	0.143	0.180	0.213	0.246	0.280	0.263	0.209	0.290
300														
	$\hat{\alpha}_{k,i}$	0.026	0.067	0.113	0.207	0.308	0.413	0.519	0.619	0.718	0.811	0.899	0.943	0.977
	Std	(0.015)	(0.025)	(0.032)	(0.040)	(0.048)	(0.058)	(0.066)	(0.069)	(0.069)	(0.065)	(0.052)	(0.040)	(0.024)
	$\bar{\sigma}^*_{\alpha_i}$	0.006	0.016	0.024	0.034	0.040	0.044	0.045	0.043	0.040	0.035	0.027	0.021	0.011
	Power	0.551	0.281	0.135	0.066	0.065	0.120	0.169	0.214	0.265	0.305	0.314	0.300	0.271
1000														
	$\hat{\alpha}_{k,i}$	0.025	0.067	0.112	0.205	0.305	0.410	0.515	0.617	0.716	0.813	0.905	0.950	0.985
	Std	(0.012)	(0.019)	(0.024)	(0.030)	(0.035)	(0.041)	(0.046)	(0.049)	(0.051)	(0.049)	(0.040)	(0.031)	(0.017)
	$\bar{\sigma}^*_{\alpha_i}$	0.005	0.013	0.019	0.028	0.033	0.035	0.036	0.035	0.033	0.029	0.022	0.016	0.008
	Power	0.587	0.317	0.151	0.059	0.073	0.097	0.127	0.198	0.238	0.286	0.298	0.304	0.206
5000														
	$\hat{\alpha}_{k,i}$	0.024	0.066	0.110	0.202	0.303	0.408	0.513	0.617	0.716	0.814	0.908	0.954	0.989
	Std	(0.012)	(0.018)	(0.022)	(0.026)	(0.031)	(0.036)	(0.041)	(0.045)	(0.047)	(0.045)	(0.036)	(0.027)	(0.013)
	$\bar{\sigma}^*_{\alpha_i}$	0.005	0.012	0.017	0.024	0.029	0.031	0.032	0.031	0.030	0.026	0.020	0.014	0.007
	Power	0.632	0.330	0.148	0.065	0.066	0.087	0.133	0.189	0.244	0.302	0.320	0.316	0.124

The DGP is the AR(1)-GARCH(1,1) model with $\varepsilon_t \sim N(0,1)$. The nominal size is 5% and $H = 500$.

the number of independent PITs is large enough as to have enough information. However, it is important to note that, as the forecast horizon increases, this implies that we need to have a very large number of out-of-sample forecasts and, in empirical applications, this is not always realistic.

5. Empirical application: Modeling VIX

The VIX is important because it is a barometer of the overall market sentiment; see Diebold and Yilmaz (2015) who define it as a fear index. Furthermore, it reflects both the stock market uncertainty and the expected premium from selling stock market variance in a swap contract. Finally, there is an active market on VIX derivatives; see Mencia and Sentana (2018) for dynamic portfolio allocation for Exchange Traded Notes (ETNs) tracking short and mid-term VIX futures indices. The recent development of volatility-based derivative products generates an interest on predictive densities of volatility. Intuitively, risk averse investors must take into account not only the expected value of the payoffs, obtained from the conditional mean forecasts, but also the risk involved, which necessarily depends on features of the conditional density.

It is commonly accepted that VIX display long-memory; see, for example, Fernandes et al. (2014). Consequently, several authors propose variants of the simple and easy-to-estimate long-memory Heterogeneous Autoregressive (HAR) model of Corsi (2008) to predict the VIX; see Fernandes et al. (2014) and Psaradelli and Sermpinis (2016). Alternatively, Mencia and Sentana (2018) propose modeling the persistence of the VIX using the Multiplicative Error Model (MEM) of Engle and Gallo (2006).

In this section, we implement the BG-ACR tests to one-step-ahead in-sample conditional densities obtained after fitting the HAR and MEM models to V_t, the daily VIX index, observed from January 2, 1990 to January 15, 2013 with a total of 5807 observations. Fernandes et al. (2014),

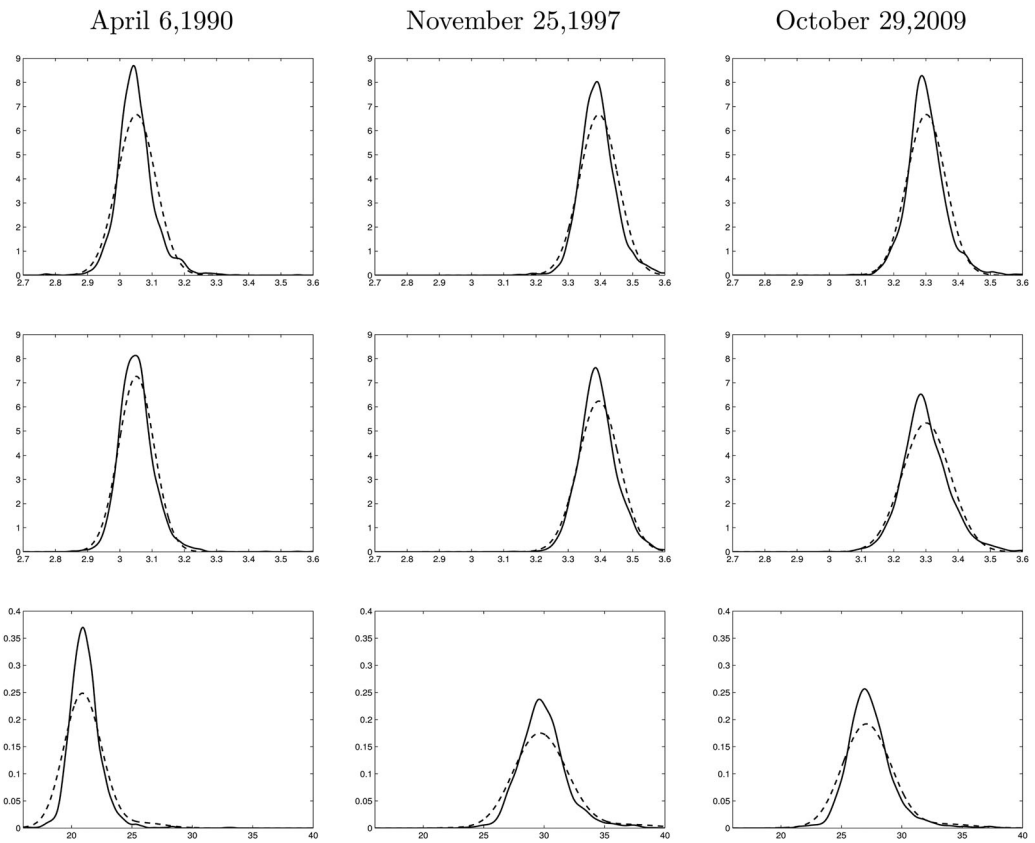

Figure 3. In-sample one-step-ahead densities obtained after fitting the HAR model (first row), the HAR-GJR model (second row) and the MEM model (third row) at three moments of time: April 6, 1990 (first column), November 25,1997 (second column) and October 29,2009 (third column). The solid lines represent the bootstrap densities and the dashed lines represent the normal density for the HAR and HAR-GJR models and the GSNP density for the MEM model.

who analyze the same series, show that the null hypothesis of a unit-root is clearly rejected and find strong evidence of long-memory. Consequently, the following HAR model is fitted to $y_t = \log V_t$[8]

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_5 \bar{y}_{t-1:5} + \phi_{10} \bar{y}_{t-1:10} + \phi_{22} \bar{y}_{t-1:22} + \phi_{66} \bar{y}_{t-1:66} + \varepsilon_t, \tag{15}$$

where $\bar{y}_{t:i} = i^{-1} \sum_{j=0}^{i-1} y_{t-j}$ and ε_t is an independent white noise sequence. Note that the HAR model in Eq. (15) is an AR(66) model reparameterized in a parsimonious way by imposing economically meaningful restrictions. As in Corsi (2008), the parameters in Eq. (15) are estimated by OLS. Standard OLS regression estimators are consistent and normally distributed. In order to account for the possible presence of serial correlation in the data, the Newey-West covariance correction for serial correlation can be employed.[9]

We compute the in-sample bootstrap conditional densities as described in Section 3. In Fig. 3, we plot kernel estimates of the bootstrap densities (solid lines) at different moments of time together with the corresponding normal density (dashed lines). We observe that not only the location but also the variance of the densities of the log-VIX change over time. When compared

[8]Fernandes et al. (2014) include explanatory variables in Eq. (15). However, we stick to a univariate model to simplify the implementation of the proposed testing procedure.

[9]Estimated parameters and residual diagnostics are reported in the supplementary material.

Table 5. In-sample G-ACR and BG-ACR tests for HAR and HAR-GJR models fitted to log-VIX and MEM model fitted to VIX. *, **, *** indicate that H_0 is rejected at 10%, 5% and 1% levels of significance, respectively.

						HAR							
α_i	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
	BG-ACR												
$\hat{\alpha}_{k,i}$	0.010	0.053	0.103	0.205	0.308	0.409	0.514	0.610	0.707	0.804	0.899	0.950	0.989
$\lvert t^*_{1,\alpha_i}\rvert$	|0.377|	|1.204|	|1.147|	|1.489	|2.190|**	|2.721|***	|4.720|***	|4.232|***	|3.018|***	|2.039|**	|0.333|	|0.145|	|1.021|
$L^{5*}_{\alpha_i}$	8.654	2.595	6.522	7.004	9.367*	9.033	27.702***	22.698***	15.490***	10.001*	5.468	18.588***	21.017***
C_1^{13*}	36.747***												
						G-ACR							
$\hat{\alpha}_{k,i}$	0.005	0.037	0.093	0.228	0.367	0.489	0.596	0.684	0.764	0.837	0.895	0.927	0.969
$\lvert t_{1,\alpha_i}\rvert$	|3.93|***	|5.02|***	|2.08|**	|6.30|***	|12.69|***	|16.98|***	|17.85||***	|16.02|***	|13.51|***	|7.87|***	|1.21|	|7.56|***	|12.54|***
$L^{5}_{\alpha_i}$	30.30***	73.09***	23.12***	51.01***	193.75***	321.82***	343.19***	284.29***	201.29***	71.28***	7.22	66.05***	198.36***
C_1^{13}	881.66***												
						HAR-GJR							
						BG-ACR							
$\hat{\alpha}_{k,i}$	0.008	0.052	0.105	0.205	0.306	0.405	0.508	0.602	0.698	0.800	0.901	0.947	0.990
$\lvert t^*_{1,\alpha_i}\rvert$	|1.699|*	|0.798|	|1.453|	|1.246|	|1.530|	|1.335|	|2.102|**	|0.699|	|1.033|	|0.070|	|0.463|	|2.291|**	|0.137|
$L^{5*}_{\alpha_i}$	5.518	3.384	5.964	5.150	7.991	5.920	6.888	1.903	3.480	3.530	7.249	7.468	5.558
C_1^{13*}	21.092*												
						G-ACR							
$\hat{\alpha}_{k,i}$	0.005	0.042	0.102	0.236	0.363	0.475	0.576	0.666	0.742	0.812	0.892	0.930	0.970
$\lvert t_{1,\alpha_i}\rvert$	|4.589|***	|2.991|***	|0.594|	|7.265|***	|11.343|***	|12.749|***	|12.518|***	|11.466|***	|7.590|***	|2.415|**	|1.974|**	|6.464|***	|10.775|***
$L^{5}_{\alpha_i}$	56.361***	28.913***	4.768	63.059***	144.393***	176.729***	179.438***	143.983***	69.193***	12.108**	9.264*	54.380***	131.270***
C_1^{13}	318.896***												
						MEM							
						BG-ACR							
$\hat{\alpha}_{k,i}$	0.013	0.057	0.109	0.205	0.304	0.401	0.497	0.590	0.689	0.787	0.894	0.948	0.988
$\lvert t^*_{1,\alpha_i}\rvert$	|2.450|**	|2.410|**	|2.492|**	|1.444|	|1.039|	|0.188|	|0.667|	|2.086|**	|2.383|**	|3.179|***	|1.852|*	|1.252|	|1.616|
$L^{5*}_{\alpha_i}$	16.861***	8.985	10.426*	6.305	11.192**	9.639*	6.895	8.528	9.276**	16.713***	6.479	17.387***	22.263***
C_1^{*13}	18.511												
						G-ACR							
$\hat{\alpha}_{k,i}$	0.007	0.046	0.105	0.225	0.342	0.455	0.546	0.636	0.721	0.807	0.898	0.957	0.988
$\lvert t_{1,\alpha_i}\rvert$	|2.578|***	|1.397|	|1.207|	|4.187|***	|6.252|***	|7.549|***	|6.004|***	|4.881|***	|3.081|***	|1.076|	|0.452|	|3.488|***	|1.298|
$L^{5}_{\alpha_i}$	7.671	8.887	6.391	20.978***	40.934***	59.412***	36.800***	25.932***	12.606**	4.555	3.978	16.503***	5.035
C_1^{13}	103.417***												

with the normal densities, we also observe large distortions. The boostrap densities are more peaked than the normal densities and they are rightly skewed.

We formally test the null hypothesis of correct specification of the HAR model for the log-VIX. Table 5 reports the sample proportions, $\hat{\alpha}_{k,i}$, and the in-sample BG-ACR statistics $t_{1,\alpha_i}, L^5_{\alpha_i}$ and C^{13}_1. We observe that the specification is strongly rejected from the 30% to the 99% autocontours by the t_{1,α_i} and $L^5_{\alpha_i}$ statistics. The C^{13}_1 statistic, which is computed adding information of all autocontours, rejects H_0 at 1% significance level. Therefore, the basic HAR model is not adequate to model the conditional densities of the daily log-VIX. Table 5 also reports the corresponding tests assuming Gaussian errors. The null is strongly rejected for almost all autocontours, with the statistics being much larger than those of the BG-ACR tests. Therefore, the overall conditional density model provided by the HAR specification of the log-VIX is strongly rejected.

Based on the information provided by the BG-ACR test, we incorporate asymmetric conditional heteroscedasticity, and fit the HAR-GJR model; see, for example, Corsi et al. (2008) for HAR-GARCH specifications in the context of realized volatility.[10] The HAR-GJR model is estimated by a two-step QML estimation, in which the HAR equation is estimated by OLS and the GJR equation by G-QML. In Fig. 3 (second row), we plot the one-step-ahead in-sample bootstrap conditional densities for three different dates. We observe that the locations of these densities are similar to those obtained with the homoscedastic HAR model. The shapes of the bootstrap densities, although still mildly asymmetric and slightly more peaked than the normal, are closer to normality. We also observe changes in the variance of the log-VIX. These differences may have important implications for developing volatility-based derivative products. In Table 5, we report the statistics corresponding to the HAR-GJR model based on conditional normality and on bootstrapping. The HAR-GJR model with bootstrap conditional densities is not rejected while the HAR-GJR with normal conditional densities is strongly rejected. This is a prime example of the power of the proposed tests because they are able to use distributional properties of the error to enhance the testing of the dynamics of the moments of interest, which in our case involves not only the specification of the conditional mean but also the conditional variance of the log-VIX.

In addition to the HAR specification, we also consider the MEM model of Mencia and Sentana (2018) that deals directly with the untransformed VIX, i.e. V_t. Mencia and Sentana (2018) consider the following MEM model with a GSNP distribution for the innovations

$$\begin{aligned} V_t - \Delta &= \mu_t \varepsilon_t, \\ \mu_t &= \varsigma_t + s_t, \\ \varsigma_t &= \varphi_0 + \varphi_1 \varsigma_{t-1} + \varphi_2 (V_{t-1} - \Delta - \mu_{t-1}), \\ s_t &= (\beta_1 + \beta_2) s_{t-1} + \beta_1 (V_{t-1} - \Delta - \mu_{t-1}), \end{aligned} \tag{16}$$

where $\varphi_0 > 0, |\varphi_1|, |\varphi_2|, |\beta_1|, |\beta_2| < 1, \beta_1 + \beta_2 < 1$ and Δ is a constant shift introduced to improve the fit by assigning zero probability to those events in which $V_t < \Delta$. The parameters of the MEM model are estimated by maximum likelihood.[11] Fig. 3 (third row) plots the one-step-ahead bootstrap conditional densities (solid lines) together with the corresponding assumed GSNP densities (dashed lines) for three different dates. It is important to note that the densities from the MEM model are not directly comparable with those from the HAR models as the former are densities for VIX while the latter correspond to log-VIX. However, the locations implied by the MEM model are similar to those implied by the HAR models. We observe large differences among the densities. The bootstrap densities are more skewed to the right and more peaked than

[10]Note that these authors conclude that the error distribution is better represented by a normal inverse Gaussian (NIG) or a normal-mixture distributions.

[11]We use the values of the parameters estimated in Mencia and Sentana (2018) as initial conditions for our estimation. Estimation results are reported in the supplementary material.

the GSNP densities. It seems that GSNP densities assign more probability mass to the observations in the left tail. The G-ACR and BG-ACR statistics reported in Table 5 confirm these conclusions. The MEM-GSNP model is clearly rejected for almost all autocontours. In Table 5, the BG-ACR statistics t_{1,α_i} indicate a mild rejection of the MEM model but the portmanteau test C_1^{13} does not reject. The portmanteau test $L_{\alpha_i}^5$ tend to reject MEM only for the extreme autocontours.

In summary, we have found strong evidence against the standard parametric assumptions of the conditional densities of the HAR and MEM models for the VIX index. In both cases, the true conditional density seems to be more skewed to the right and more peaked than either normal or GSNP densities, with location and variance changing over time. We have shown that bootstrap densities deliver good results for the testing of the density model of the VIX index. The preferred specification is the heteroscedastic HAR-GJR model with bootstrap conditional densities of the log-VIX.

6. Conclusions

We propose an extension of the G-ACR tests of González-Rivera and Sun (2015) for dynamic specification of a density model (in-sample) and for evaluation of forecast densities (out-of-sample). Our contribution lies on computing the PITs from a bootstrapped conditional density so that no assumption on the functional form of the density is needed. Furthermore, the bootstrap procedure directly incorporates parameter uncertainty. Our proposed tests are easy to compute and have standard asymptotic distributions that approximate well the finite sample distribution under the null. The tests, which are powerful for detecting departures from the assumed conditional density, are accompanied by a graphical tool that provides information on the potential sources of misspecification.

The proposed approach is particularly useful to evaluate forecast densities when the error distribution is unknown as, for example, in the context of multi-step forecasts in nonlinear and/or non-Gaussian models. A very interesting application is the modeling of the VIX index where several parametric conditional densities have been proposed. We evaluate the adequacy of conditional densities of the daily VIX index derived from the HAR and MEM models and strongly reject the standard parametric assumptions of normality in the case of HAR model and of GSNP in the case of the MEM models. Our results suggest that conditional heteroscedasticity should be taken into account for an adequate construction of the conditional density regardless of the specification used for the conditional mean.

Acknowledgments

We are grateful to the participants at the New Developments in Econometrics and Time Series Workshop, Madrid, October 2016, and at the IMF/IIF Workshop on Forecasting Issues on Developing Economies, Washington DC, April 2017, and to seminar participants at the Management School of the University of Liverpool, for their very useful comments. We are also thankful to J. Mencía and E. Sentana for their help with the codes to estimate the MEM model.

Funding

Financial support from the Spanish Ministry of Education and Science, research project ECO2015-70331-C2-2-R (MINECO/FEDER) is acknowledged by the four authors. The first author and fourth authors also acknowledge financial support from Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES) grant 88882.305837/2018-01 and research projects PGC2018-096977-B-100 and FCT grant UID/GES/00315/2019, respectively. Gloria González-Rivera wishes to thank the Department of Statistics at UC3M for their hospitality and the financial support of the 2015 Chair of Excellence UC3M/Banco de Santander, and the UCR Academic Senate grant.

References

Ait-Sahalia, Y., Fan, J., Peng, H. (2009). Nonparametric transition-based tests for jump diffusions. *Journal of the American Statistical Association* 104(487):1102–1116. doi:10.1198/jasa.2009.tm08198

Altissimo, F., Mele, A. (2009). Simulated non-parametric estimation of dynamic models. *Review of Economic Studies* 76(2):413–450. doi:10.1111/j.1467-937X.2008.00527.x

Andrews, D. W. K., Buchinsky, M. (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica* 68(1):23–51. doi:10.1111/1468-0262.00092

Bhardwaj, G., Corradi, V., Swanson, N. R. (2008). A simulation-based specification test for diffusion processes. *Journal of Business & Economic Statistics* 26(2):176–193. doi:10.1198/073500107000000412

Bierens, H. J., Wang, L. (2017). Weighted simulated integrated conditional moment tests for parametric conditional distributions of stationary time series processes. *Econometric Reviews* 36(1–3):103–135. doi:10.1080/07474938.2015.1114275

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(4):255–276. doi:10.1002/1099-131X(200007)19:4<255::AID-FOR773>3.0.CO;2-G

Corradi, V., Swanson, N. R. (2006a). Predictive density evaluation. *Handbook of Economic Forecasting*, Amsterdam: Elsevier (North Holland Publishing Co.), Vol. 1, Chapter 5, pp. 197–284.

Corradi, V., Swanson, N. R. (2006b). Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics* 133(2):779–806. doi:10.1016/j.jeconom.2005.06.013

Corsi, F. (2008). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7(2):174–196. doi:10.1093/jjfinec/nbp001

Corsi, F., Mittnik, S., Pigorsch, C., Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews* 27(1–3):46–78. doi:10.1080/07474930701853616

Diebold, F. X., Gunther, T. A., Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39(4):863–882. doi:10.2307/2527342

Diebold, F. X., Yilmaz, K. (2015). Trans-Atlantic equity volatility connectedness: U.S. and European financial institutions. *Journal of Financial Econometrics* 14(1):81–127.

Engle, R. F., Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* 131(1–2):3–27. doi:10.1016/j.jeconom.2005.01.018

Fernandes, M., Medeiros, M. C., Scharth, M. (2014). Modeling and predicting the CBOE market volatility. *Journal of Banking & Finance* 40:1–10. doi:10.1016/j.jbankfin.2013.11.004

Francq, C., Zakoian, J.-M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* 10(4):605–637. doi:10.3150/bj/1093265632

González-Rivera, G., Sun, Y. (2015). Generalized autocontours: Evaluation of multivariate density models. *International Journal of Forecasting* 31(3):799–814. doi:10.1016/j.ijforecast.2014.03.019

González-Rivera, G., Sun, Y. (2017). Density forecast evaluation in unstable environments. *International Journal of Forecasting* 33(2):416–432. doi:10.1016/j.ijforecast.2016.10.003

Hall, P., Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* 71(1):285–317. doi:10.1111/1468-0262.00396

Hidalgo, J., Zaffaroni, P. (2007). A goodness-of-fit test for ARCH(∞) models. *Journal of Econometrics* 141(2):973–1013. doi:10.1016/j.jeconom.2006.11.008

Hong, Y., Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies* 18(1):37–84. doi:10.1093/rfs/hhh006

Inoue, A., Kilian, L. (2005). In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews* 23(4):371–402. doi:10.1081/ETC-200040785

Jeong, M. (2017). Residual-based GARCH bootstrap and second order asymptotic refinement. *Econometric Theory* 33(3):779–790. doi:10.1017/S0266466616000104

Kless, P. C. (2019). New bootstrap methods for financial and economic time series. PhD thesis. Copenhagen: University of Copenhagen.

Kreiss, J.-P., Lahiri, S. N. (2012). Bootstrap methods for time series. In: Rao, T. S., Rao, S. S., Rao, C. R., eds., *Handbook of Statistics, Time Series Analysis: Methods and Applications*, Vol. 30, pp. 3-26, North Holland.

Manzan, S., Zerom, D. (2008). A bootstrap-based non-parametric forecast density. *International Journal of Forecasting* 24(3):535–1550. doi:10.1016/j.ijforecast.2007.12.004

Mencia, J., Sentana, E. (2018). Volatility-related exchange traded assets: An econometric investigation. *Journal of Business & Economic Statistics* 36(4):599–614. doi:10.1080/07350015.2016.1216852

Mika, M., Saikkonen, P. (2011). Parameter estimation in non-linear AR-GARCH models. *Econometric Theory* 27(6):1236–1278. doi:10.1017/S0266466611000041

Pascual, L., Romo, J., Ruiz, E. (2004). Bootstrap predictive inference for ARIMA processes. *Journal of Time Series Analysis* 25(4):449–465. doi:10.1111/j.1467-9892.2004.01713.x

Pascual, L., Romo, J., Ruiz, E. (2006). Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics & Data Analysis* 50(9):2293–2312. doi:10.1016/j.csda.2004.12.008

Perera, I., Silvapulle, M. J. (2018). Specification tests for time series models with GARCH-type conditional variance. Discussion paper. Available at SSRN: https://ssrn.com/abstract=3141822. Last accessed May 2019.

Perera, I., Silvapulle, M. J. (in press). Bootstrap based probability forecasting in multiplicative error models. *Journal of Econometrics*.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science* 18(2):219–230. doi:10.1214/ss/1063994977

Psaradelli, I., Sermpinis, G. (2016). Modelling and trading the US implied volatility indices. Evidence from the VIX, VXN and VXD indices. *International Journal of Forecasting* 32(4):1268–1283. doi:10.1016/j.ijforecast.2016.05.004

Reeves, J. J. (2005). Bootstrap prediction intervals for ARCH models. *International Journal of Forecasting* 21(2):237–248. doi:10.1016/j.ijforecast.2004.09.005

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* 23(3):470–472. doi:10.1214/aoms/1177729394

Rossi, B., Sekhposyan, T. (2014). Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set. *International Journal of Forecasting* 30(3):662–682. doi:10.1016/j.ijforecast.2013.03.005

Rossi, B., Sekhposyan, T. (2019). Alternative tests for correct specification of conditional forecast densities. *Journal of Econometrics* 208(2):638–657. doi:10.1016/j.jeconom.2018.07.008