A Bootstrap Approach for Generalized Autocontour Testing. Implications for VIX forecast densities

João Henrique G. Mazzeu*
Department of Statistics, Universidad Carlos III de Madrid

Gloria González-Rivera
Department of Economics, University of California, Riverside

Esther Ruiz
Department of Statistics, Universidad Carlos III de Madrid

Helena Veiga
Department of Statistics, Universidad Carlos III de Madrid
BRU-IUL, Instituto Universitário de Lisboa

July 27, 2017

Abstract

We propose an extension of the Generalized Autocontour (G-ACR) tests for dynamic specification of *in-sample* conditional densities and for evaluation of *out-of-sample* forecast densities. The new tests are based on probability integral transforms (PITs) computed from bootstrap conditional densities that incorporate parameter uncertainty. Then, the parametric specification of the conditional moments

*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 fourth author also acknowledges research project ECO2012-3240 and FCT grant UID/GES/00315/2013. 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. 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.

can be tested without relying on any parametric error distribution yet exploiting distributional properties of the variable of interest. We show that the finite sample distribution of the bootstrapped G-ACR (BG-ACR) tests are well approximated using standard asymptotic distributions. Furthermore, the proposed tests are easy to implement and are accompanied by graphical tools that provide information about the potential sources of misspecification. We apply the BG-ACR tests to the Heterogeneous Autoregressive (HAR) model and the Multiplicative Error Model (MEM) of the U.S. volatility index VIX. We find strong evidence against the parametric assumptions of the conditional densities, i.e. normality in the HAR model and semi non-parametric Gamma (GSNP) in the MEM. 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, variance and skewness changing over time. The preferred specification is the heteroscedastic HAR model with bootstrap conditional densities of the log-VIX. Supplementary materials for this article are available online.

Keywords: Distribution Uncertainty; Model Evaluation; Parameter Uncertainty; PIT; HAR model; Multiplicative Error Model

1 Introduction

Density forecasting is a very active and important area of research in the analysis of economic and financial time series. The need to consider the full predictive density has long been recognized in the related literature; see Tay and Wallis (2000) for a survey. A problem often faced by forecasters is testing the correct specification of a conditional forecast density; see, for example, Mitchell and Wallis (2011). Appropriate tests should take into account that the forecast conditional distribution is often unknown, the specification of conditional moments is also unknown, and the parameters in the conditional moments have to be estimated. Furthermore, a useful test would indicate the source of rejection of a given forecasting model, that is, whether it is rejected because of the specification of the functional form of the conditional distribution or because of the specification of the conditional moments.

Many tests for conditional forecast densities available in the literature are based on testing a joint hypothesis of uniformity and independence (i.i.d. U(0,1)) of the probability integral transforms (PITs) (Rosenblatt, 1952) which are applicable regardless of the particular user's loss function. Diebold et al. (1998) and Diebold et al. (1999) introduce these tests in the econometric literature. Intuitively, the i.i.d. assumption of the PITs is related with the correct specification of the conditional moments, while the U(0,1) property characterizes the correct specification of the error distribution. The PITs contain rich information on model misspecification that can be revealed by using their histograms and autocorrelograms as suggested by Diebold et al. (1998). However, it is nontrivial to develop a formal test for the joint hypothesis of independence and uniformity of the PITs. The well-known Kolmogorov-Smirnov test checks uniformity under the independence assumption rather than testing both properties jointly. Consequently, it would easily miss the non-independent alternatives when PITs have a marginal uniform distribution. Moreover, the Kolmogorov-Smirnov test does not take into account the impact of parameter estimation uncertainty on the asymptotic distribution of the statistic. To solve this problem, Bai (2003) proposes a Kolmogorov-Smirnov-type test based on a martingale transformation of the PITs whose asymptotic distribution is free from the impact of parameter estimation. Yet, the test proposed by Bai (2003) only checks uniformity and,

consequently, it has no asymptotic unit power if the transformed PITs are uniform but not independent; see Corradi and Swanson (2006) and Rossi and Sekhposyan (2013). Alternatively, Hong and Li (2005) propose a nonparametric-kernel-based test with power against violations of both independence and density functional form, but it depends on the choice of a bandwidth, which could be problematic to choose in an empirical context.

Instead of testing for independence and uniformity of PITs, González-Rivera et al. (2011) and González-Rivera and Yoldas (2012) propose autocontour (ACR) tests to evaluate the adequacy of the conditional density model based on the generalized errors of the model. They propose the "autocontour" device as a graphical tool that can be very helpful for guiding the modelling. Moreover, it permits to focus on different areas of the conditional density in order to assess those regions of interest. The ACR tests, which can be applied to both original series and model residuals, have several advantages: i) they have standard convergence rates and standard limiting distributions that deliver superior power; ii) they are computationally easy to implement as they are based on a counting process; iii) they do not require either a transformation of the original data or an assessment of the Kolmogorov goodness of fit; and iv) they explicitly account for parameter uncertainty. Yet, they assume a parametric time-invariant functional form of the conditional density and, once we depart from standard densities, e.g. normal, Student-t, etc., the analytical expressions of the autocontours may be mathematically cumbersome to derive. This problem becomes more severe when we deal with conditional multivariate densities. To overcome these problems, González-Rivera and Sun (2015) propose the generalized autocontour (G-ACR) tests that are based on PITs instead of original observations or residuals. In this way, the G-ACR tests inherit both the advantages of using PITs and those of using autocontours. However, the tests are still based on assuming a particular specification of the conditional density in order to compute the PITs. 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 (often normality) or to the specification of the conditional moments. However, it is important to understand whether the rejection of a forecast density is due to the often assumed normal distribution or to the specification of the conditional moments. Furthermore, there are applications in which the conditional density does not

have a known closed-form expression. For instance, when the errors are non-Gaussian or when the model is non-linear, it is difficult to obtain the functional form of the multi-step predictive densities.

In this paper, we propose an extension of the G-ACR tests for (in-sample) dynamic specification of a density model and for (out-of-sample) 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. 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 distributed. The bootstrap procedure 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, using graphical devices, the procedure allows the identification of the source of misspecification, namely, whether it is the error distribution or linear/non-linear dynamics.

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). The VIX is important because it is a barometer of the overall market sentiment; see Whaley (2000, 2009) and 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. The number of VIX futures contracts traded increased dramatically from about 1 million in 2007 to about 24 million in 2012 with the largest growth occurring after 2009, likely caused by the recent financial crisis; see, for example, Park (2016), Song and Xiu (2016) and Martin (2017) for some recent references on pricing VIX derivatives and Mencía and Sentana (2016) 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. After implementing the BG-ACR tests, we show that the HAR and MEM specifications are both rejected if they do not incorporate conditional heterocedasticity. Furthermore, we also show that normality of the errors of the HAR model and the semiparametric Gamma distribution of the errors of the MEM model are also rejected.

The rest of the paper is organized as follows. In section 2, we briefly describe the G-ACR test to make the exposition self-contained. In section 3, we provide the main contribution with the description of the new proposed BG-ACR tests. In section 4, we analyze their in-sample finite sample performance and, in section 5, their out-of-sample performance. In section 6, we offer an empirical application to illustrate the advantages of the BG-ACR tests; we test for the adequacy of the HAR and MEM models to obtain forecast densities of the VIX index. Finally, we conclude in section 7.

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.

Let $\{y_t\}_{t=1}^T$ denote the random process of interest with 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. The random process y_t may enjoy very general statistical properties, e.g. heterogeneity, dependence, etc. 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 follows

$$u_t = \int_{-\infty}^{y_t} g_t(v_t|Y_{t-1})\, dv_t. \tag{1}$$

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) and Diebold et al. (1998).

Therefore, the null hypothesis $H_0 : g_t(y_t|Y_{t-1}) = f_t(y_t|Y_{t-1})$ is equivalent to the null hypothesis

$$H_0' : \{u_t\}_{t=1}^{T} \quad \text{is i.i.d.} \quad U(0,1). \tag{2}$$

Note that, if the forecast density coincides with the true data generating process (DGP), then it is preferred by all forecasters regardless of their particular loss function; see Diebold et al. (1998) and Granger and Pesaran (2000a,b). In order to compute the PITs in equation (1), one needs to assume a particular distribution function for $g_t(y_t|Y_{t-1})$. Simple tests of independence and uniformity, such as, the Kolmogorov-Smirnov test suffer from the problems described in the introduction. Alternatively, González-Rivera and Sun (2015) propose an extension of the autocontour concepts in González-Rivera et al. (2011) to evaluate the properties of the PITs under the null hypothesis (2).

Define G-ACR$_{k,\alpha_i}$ as the set of points in the plane (u_t, u_{t-k}) such that the square with $\sqrt{\alpha_i}$-side and origin at (0,0) contains α_i% of observations[1], i.e.

$$\text{G-ACR}_{k,\alpha_i} = \{B(u_t, u_{t-k}) \subset \Re^2 | 0 \leq u_t \leq \sqrt{\alpha_i} \quad \text{and} \quad 0 \leq u_{t-k} \leq \sqrt{\alpha_i}, s.t. : u_t \times u_{t-k} \leq \alpha_i\}, \tag{3}$$

and the indicator series I_t^{k,α_i}, which takes value one if $(u_t, u_{t-k}) \in$ G-ACR$_{k,\alpha_i}$ and zero otherwise.

If $g_t(y_t|Y_{t-1})$ is a consistent estimator of $f_t(y_t|Y_{t-1})$, then I_t^{k,α_i} is asymptotically a Bernoulli MA process whose order depends on k. The sample proportion of PIT pairs (u_t, u_{t-k}) within the G-ACR$_{k,\alpha_i}$ cube is given by

$$\widehat{\alpha}_{k,i} = \frac{\sum_{t=k+1}^{T} I_t^{k,\alpha_i}}{T-k}. \tag{4}$$

Consider the following statistic

$$t_{k,\alpha_i} = \frac{\sqrt{T-k}(\widehat{\alpha}_{k,i} - \alpha_i)}{\sigma_{\alpha_i}}, \tag{5}$$

where $\sigma_{\alpha_i}^2 = \alpha_i(1-\alpha_i) + 2\alpha_i^{3/2}(1-\alpha_i^{1/2})$. González-Rivera and Sun (2015) show that, under

[1]It is possible to define other regions in the unit cube depending on the interest of the researcher.

the null hypothesis in (2), t_{k,α_i} is asymptotically standard normal distributed.

The t-statistic in (5) 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 $\ell_{k,\alpha_i} = \sqrt{T-k}(\widehat{\alpha}_{k,i} - \alpha_i)$. Under H_0' in (2), $L_{\alpha_i}' \Lambda_{\alpha_i}^{-1} L_{\alpha_i}$ 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 $c_{k,i} = \sqrt{T-k}(\widehat{\alpha}_{k,i} - \alpha_i)$. Once more, under H_0' in (2), $C_k' \Omega_k^{-1} C_k$ has asymptotically a χ_C^2 distribution, where a typical element of the asymptotic covariance matrix, Ω_k, is given by:

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

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 a given known 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 $T \longrightarrow \infty$, $H \longrightarrow \infty$ and $H/T \longrightarrow 0$, parameter uncertainty is asymptotically negligible and no adjustment to the test is needed.

Finally, note that, if any of the G-ACR tests described above rejects the null hypothesis, there is not any 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 also justify heuristically the asymptotic distribution of the corresponding statistics and carry out Monte Carlo experiments to establish the finite sample performance of the new proposed tests.

Consider the following parametric location-scale model for the series of interest, y_t, $t = 1, ..., T$,

$$y_t = \mu_t + \sigma_t \varepsilon_t, \tag{6}$$

where μ_t and σ_t^2 are the conditional mean and variance of y_t, which are specified as parametric functions of the information set Y_{t-1}, and ε_t is a strict 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. Asymptotic normality of the parameter estimator is a requirement for the bootstrap to be asymptotically valid for the estimation of its sample distribution; see, for example, Hall and Yao (2003).

We consider a particular specification of (6) to illustrate the proposed tests, namely the following popular AR(1)-GARCH(1,1) model,

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + a_t, \\ a_t &= \varepsilon_t \sigma_t, \\ \sigma_t^2 &= \omega_0 + \omega_1 a_{t-1}^2 + \omega_2 \sigma_{t-1}^2, \end{aligned} \tag{7}$$

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. We consider the Quasi-Maximum Likelihood (QML) estimators of the parameters of the AR(1)-GARCH(1,1) model in (7) obtained by maximizing the Gaussian log-likelihood function. Francq and Zakoïan (2004) prove the strong consistency and asymptotic normality of the QML estimator of the ARMA-GARCH model under finite fourth order moments of the observed series.

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 (7). The algorithm is based on the residual bootstrap algorithms of Pascual et al. (2004, 2006) for the construction of forecast densities in linear ARMA models and GARCH models, respectively.

In-sample bootstrap algorithm

Step 1 Obtain the residuals

Estimate the parameters of model (7) by a two-step QML estimator: $\widehat{\phi}_0$, $\widehat{\phi}_1$, $\widehat{\omega}_0$, $\widehat{\omega}_1$ and $\widehat{\omega}_2$. Obtain the standardized residuals $\widehat{\varepsilon}_t = \frac{\widehat{a}_t}{\widehat{\sigma}_t}$, $t = 2, ..., T$, where $\widehat{a}_t = y_t - \widehat{\phi}_0 - \widehat{\phi}_1 y_{t-1}$, $\widehat{\sigma}_2^2 = \widehat{\omega}_0/(1 - \widehat{\omega}_1 - \widehat{\omega}_2)$ and $\widehat{\sigma}_t^2 = \widehat{\omega}_0 + \widehat{\omega}_1 \widehat{a}_{t-1}^2 + \widehat{\omega}_2 \widehat{\sigma}_{t-1}^2$, for $t = 3, ..., T$. Denote by $\widehat{F}_{\widehat{\varepsilon}}$ 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

$$\begin{aligned} \sigma_t^{*2(b)} &= \widehat{\omega}_0 + \widehat{\omega}_1 a_{t-1}^{*2(b)} + \widehat{\omega}_2 \sigma_{t-1}^{*2(b)}, \quad (8) \\ a_t^{*(b)} &= \varepsilon_t^{*(b)} \sigma_t^{*(b)}, \\ y_t^{*(b)} &= \widehat{\phi}_0 + \widehat{\phi}_1 y_{t-1}^{*(b)} + a_t^{*(b)}, \quad (9) \end{aligned}$$

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

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 as follows:

$$\begin{aligned} \sigma_t^{**2(b)} &= \widehat{\omega}_0^{*(b)} + \widehat{\omega}_1^{*(b)} (y_{t-1} - \widehat{\phi}_0^{*(b)} - \widehat{\phi}_1^{*(b)} y_{t-2})^2 + \widehat{\omega}_2^{*(b)} \sigma_{t-1}^{**2(b)}, \quad (10) \\ y_t^{**(b)} &= \widehat{\phi}_0^{*(b)} + \widehat{\phi}_1^{*(b)} y_{t-1} + \sigma_t^{**(b)} \varepsilon_t^{*(b)}, \quad (11) \end{aligned}$$

where $\sigma_2^{**2(b)} = \widehat{\omega}_0^{*(b)} / (1 - \widehat{\omega}_1^{*(b)} - \widehat{\omega}_2^{*(b)})$ and $\varepsilon_t^{*(b)}$ are random extractions with replacement from $\widehat{F}_{\widehat{\varepsilon}}$.

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

Note that, in step 2, we obtain replicates of y_t^* that are not conditional on $\{y_1, ..., y_{t-1}\}$. In (8), σ_t^{*2} depends on a_{t-1}^{*2} and in (9), y_t^* depends on y_{t-1}^*. Therefore, independent replicates of the process are generated to estimate the parameters and to obtain an estimate of their sample distribution. However, in step 3, the bootstrap replicates, σ_t^{**2} and y_t^{**}, in (10) and (11) 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 way, at each moment of time, $t = 3, ..., T$, the above algorithm

generates $B^{(1)}$ 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). Note that a non-linear GARCH model will require a larger number of bootstrap replicates than a linear model.

In-sample PITs can be easily computed as follows

$$u_t = \frac{1}{B^{(1)}} \sum_{b=1}^{B^{(1)}} \mathbf{1}(y_t^{**(b)} < y_t), \tag{12}$$

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}, the sample proportions, $\widehat{\alpha}_{k,i}$, can be calculated as in (4). Finally, the t_{k,α_i}, L_{α_i} and C_k statistics can be calculated as explained in the previous section.

It is worth noting 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; see, for example, Mika and Saikkonen (2011) who prove the strong consistency and asymptotic normality of the Gaussian QML estimator allowing both the conditional mean and the conditional variance to be nonlinear.

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:

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \tag{13}$$

with $\phi_1 = (0.5, 0.95)$ and i.i.d. ε_t either N(0,1), or centered and standardized Student-5, or $\chi^2_{(5)}$. In each case, an AR(1) model is fitted to the artificial series with the parameters estimated by QML. Then, the 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 $B^{(1)} = 999$ replicates; see Pascual et al. (2004, 2006) and Pan and Politis

(2016) for the same number of replicates. In Figure 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 not indication about whether the rejection is coming from the misspecification of the conditional mean or from a misspecified functional form of the error distribution.

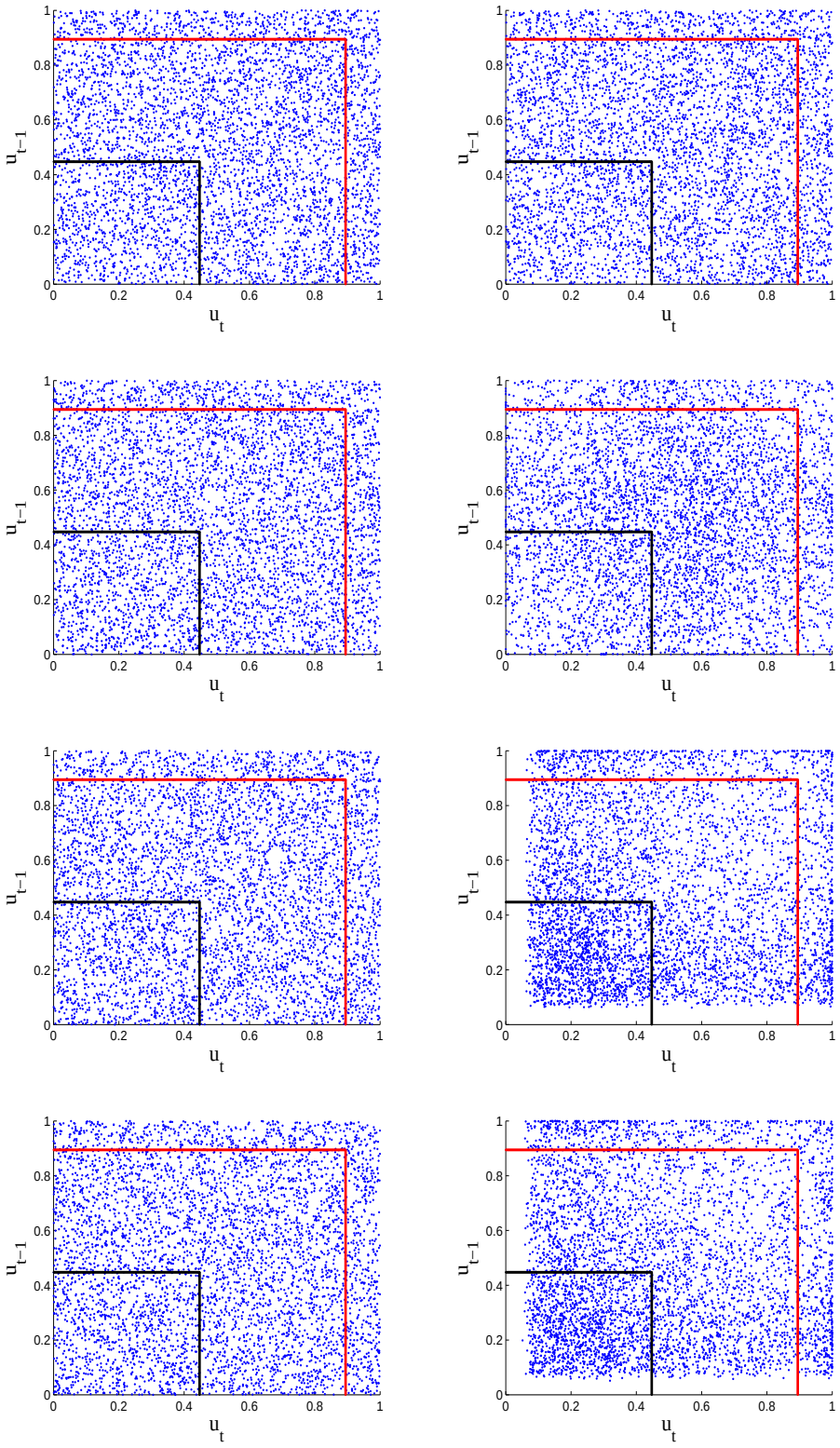

Figure 1: Pairs (u_t, u_{t-1}) and autocontours for the estimated AR(1) model with $T = 5000$. $ACR_{20\%,1}$ corresponds to the black box and the $ACR_{80\%,1}$ to the red box. The DGPs are the AR(1) model with: $\phi_1 = 0.5$ and $\varepsilon_t \sim N(0,1)$ (first row); $\phi_1 = 0.5$ and $\varepsilon_t \sim$ 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 were computed using the bootstrap algorithm with $B^{(1)}$=1000 (first column), or assuming Gaussian errors (second column).

Consider now the following three DGPs, from which we generate three time series of size $T = 5000$

$$y_t = 0.3y_{t-1} + 0.6y_{t-2} + \varepsilon_t, \tag{14}$$

$$y_t = \begin{cases} 0.5y_{t-1} + \varepsilon_t, & \text{for } t < T/2, \\ 1 + 0.5y_{t-1} + \varepsilon_t, & \text{for } t \geq T/2, \end{cases} \tag{15}$$

$$\begin{aligned} y_t &= 0.5y_{t-1} + \varepsilon_t\sigma_t, \\ \sigma_t^2 &= 0.05 + 0.5\varepsilon_{t-1}^2\sigma_{t-1}^2 + 0.45\sigma_{t-1}^2, \end{aligned} \tag{16}$$

with ε_t defined as above. We fit an AR(1) model to each of the simulated series and estimate its parameters by QML. As above, we compute the PITs both assuming normal errors and using the proposed bootstrap procedure. In Figure 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 (14) with $\chi^2_{(5)}$ errors (first row); the AR(1) model with structural break in the mean in (15) with $\varepsilon_t \sim \chi^2_{(5)}$ (second row); the GARCH model in (16) with normal errors (third row); and the GARCH model in (16) 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 towards the four corners of the unit-square. Furthermore, in this last case, the autocontour plots are very similar regardless of the error distribution of the DGP. 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.

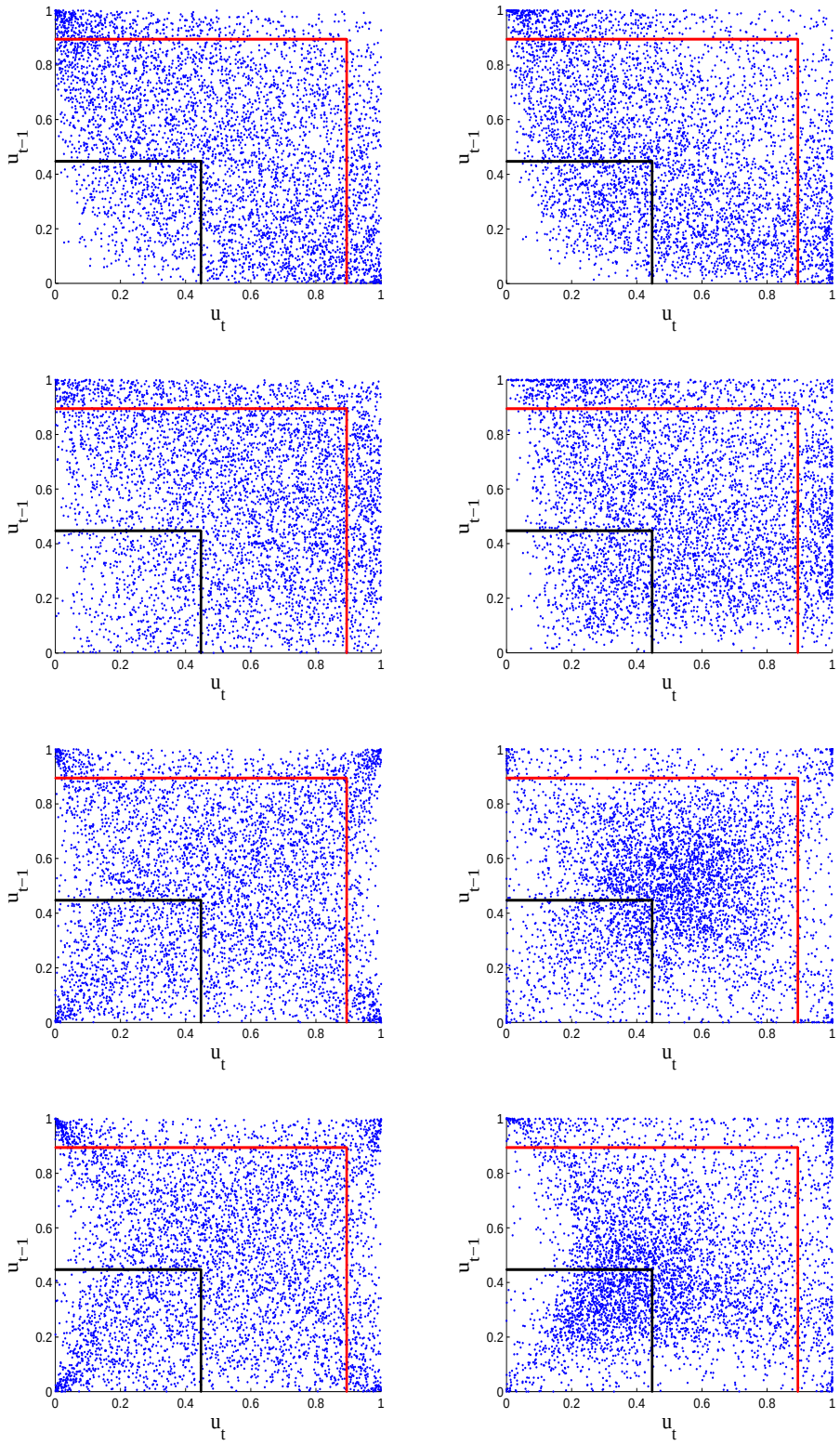

Figure 2: Pairs (u_t, u_{t-1}) and autocontours for estimated AR(1) model with $T = 5000$. $ACR_{20\%,1}$ corresponds to the black box and the $ACR_{80\%,1}$ to the red box. The DGPs are: AR(2) with $\varepsilon_t \sim \chi^2_{(5)}$ (first row); AR(1) model with break in the mean with $\varepsilon_t \sim \chi^2_{(5)}$ (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 were computed using the bootstrap algorithm with $B^{(1)} = 1000$ (first column), or assuming Gaussian errors (second column).

The asymptotic distributions of the t_{k,α_i}, L_{α_i} and C_k statistics depend on the asymptotic validity of the residual bootstrap algorithm described above which has been established by Pascual et al. (2004) in the context of obtaining predictive densities of linear ARMA models. However, as far as we know, there is not a formal proof of the validity of the procedure to construct predictive densities of nonlinear GARCH models. In order to show that the algorithm is asymptotically valid, one needs first to show that the bootstrap procedure in step 2, generates asymptotically valid estimates of the model parameters. When implemented in GARCH models, Hidalgo and Zaffaroni (2007) show the first order validity of the bootstrap QML estimator of the parameters of an ARCH(∞) process characterized by a particular decay in the ARCH parameters.[2] 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 and, consequently, the distribution of $\widehat{\alpha}_{k,i}$ should be as in (5) with the asymptotic variance corrected to take into account parameter uncertainty.[3]

Following the suggestion of González-Rivera and Sun (2015), the variance of $\widehat{\alpha}_{k,i}$ is approximated using a bootstrap procedure to take into account parameter uncertainty. $B^{(2)}$ bootstrap replicates, $\{y_t^{*(b)}\}_{t=1}^T$ are generated as in (9) and $\widehat{\alpha}_{k,i}^{*(b)}$ is obtained using the bootstrap series as if they were the original series. The bootstrap variance of $\widehat{\alpha}_{k,i}$ is given by

$$\sigma_{\alpha_i}^{*2} = \frac{1}{B^{(2)}-1}\sum_{b=1}^{B^{(2)}}\left(\widehat{\alpha}_{k,i}^{*(b)} - \frac{1}{B^{(2)}}\sum_{b=1}^{B^{(2)}}\widehat{\alpha}_{k,i}^{*(b)}\right)^2, \tag{17}$$

and the corresponding corrected t-statistic is

$$t_{k,\alpha_i}^{*} = \frac{(\widehat{\alpha}_{k,i} - \alpha_i)}{\sigma_{\alpha_i}^{*}}, \tag{18}$$

[2]Shimizu (2010, 2013, 2014) prove the consistency of the bootstrap QML estimator in the context of an AR(1)-ARCH(1) model. However, the residual bootstrap considered by Shimizu (2010, 2013, 2014) is not exactly the same as that considered in this paper. All the trajectories share the same estimated conditional mean and variance when generating bootstrap replicates to estimate the parameters. It is important to point out that Corradi and Iglesias (2008) cast some doubts on the asymptotic validity of the residual bootstrap described in step 2. Alternatively, they show that a block bootstrap based on resampling the likelihood as proposed by Gonçalves and White (2004) is asymptotically valid. Therefore, in step 2 of the algorithm described above, one can consider using this block bootstrap instead of the residual bootstrap.

[3]Monte Carlo results on the size distortions of the t-statistic when the asymptotic variance is computed as in (5) are available upon request.

which asymptotically has a N(0,1) distribution. In this paper, results are based on $B^{(2)}$=500 bootstrap replicates to compute $\sigma^*_{\alpha_i}$; see González-Rivera and Sun (2015). Note that the number of replicates needed to estimate standard errors is smaller than that required to estimate intervals; see Efron (1987).

Obviously, the variances and covariances of the portmanteau statistics can also be computed using the same arguments. In particular, a typical element of the covariance matrix of L_{α_i}, say $\lambda^*_{j,k}$, is obtained as follows:

$$\lambda^*_{j,k} = \begin{cases} \sigma^{2*}_{\alpha_i}, & \text{if } j = k, \\ \frac{1}{B^{(2)}-1} \sum_{b=1}^{B^{(2)}} \left(\widehat{\alpha}^{*(b)}_{j,i} - \frac{1}{B^{(2)}} \sum_{b=1}^{B^{(2)}} \widehat{\alpha}^{*(b)}_{j,i} \right) \left(\widehat{\alpha}^{*(b)}_{k,i} - \frac{1}{B^{(2)}} \sum_{b=1}^{B^{(2)}} \widehat{\alpha}^{*(b)}_{k,i} \right), & \text{if } j \neq k. \end{cases} \tag{19}$$

Similarly, a typical element of the covariance matrix of C_k, say $\omega^*_{i,j}$, is obtained analogously to (19) as follows:

$$\omega^*_{i,j} = \begin{cases} \sigma^{2*}_{\alpha_i}, & \text{if } i = j, \\ \frac{1}{B^{(2)}-1} \sum_{b=1}^{B^{(2)}} \left(\widehat{\alpha}^{*(b)}_{k,i} - \frac{1}{B^{(2)}} \sum_{b=1}^{B^{(2)}} \widehat{\alpha}^{*(b)}_{k,i} \right) \left(\widehat{\alpha}^{*(b)}_{k,j} - \frac{1}{B^{(2)}} \sum_{b=1}^{B^{(2)}} \widehat{\alpha}^{*(b)}_{k,j} \right), & \text{if } i \neq j. \end{cases} \tag{20}$$

4 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 $\widehat{\alpha}_{k,i}$, L_{α_i} and C_k is $B^{(2)} = 500$.

4.1 Studying the size

To investigate the size properties of the tests, we consider as DGP the AR(1) in (13). For each Monte Carlo replicate, we compute the proportions $\widehat{\alpha}_{k,i}$, for $k = 1, ..., 5$, and their bootstrap variances. Then, we compute the Monte Carlo averages and standard deviations of $\widehat{\alpha}_{k,i}$, 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 the smallest sample size of $T = 50$, the Monte Carlo averages of $\widehat{\alpha}_{k,i}$ 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 $\widehat{\alpha}_{k,i}$. For relatively small sample sizes, the bootstrap standard deviations tend to overestimate the empirical standard deviations of $\widehat{\alpha}_{k,i}$, 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.

We also analyze the finite sample performance of the two portmanteau tests. Table 2 reports the Monte Carlo percentage of rejections of $L^5_{\alpha_i}$ (adding up the information over the first five lags) and of C_1 (adding information of the thirteen quantiles considered) for the same model considered in Table 1. Regarding $L^5_{\alpha_i}$, we observe that, the Monte Carlo percentage of rejections is very close to the nominal size with a tendency to over-reject for the largest quantiles. Regarding C_1, we observe that it under-rejects when the sample size is not large enough, and over-rejects for very large samples.

Summarizing, asymptotic normality is a good approximation to the finite sample performance 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.[4]

[4]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. The conclusions are the same.

Table 1

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

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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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

Table 2

Monte Carlo size results for $L^5_{\alpha_i}$ and C_1 statistics. The DGP is $y_t = 0.95y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim \chi^2_{(5)}$. The nominal size is 5%.

	$L^5_{0.01}$	$L^5_{0.05}$	$L^5_{0.1}$	$L^5_{0.2}$	$L^5_{0.3}$	$L^5_{0.4}$	$L^5_{0.5}$	$L^5_{0.6}$	$L^5_{0.7}$	$L^5_{0.8}$	$L^5_{0.9}$	$L^5_{0.95}$	$L^5_{0.99}$	C^{13}_1
50	0.121	0.091	0.073	0.043	0.036	0.033	0.048	0.059	0.064	0.063	0.081	0.107	0.058	0.023
100	0.098	0.065	0.053	0.045	0.038	0.041	0.054	0.048	0.051	0.038	0.091	0.115	0.027	0.008
300	0.083	0.051	0.057	0.032	0.047	0.045	0.041	0.040	0.036	0.050	0.060	0.095	0.075	0.023
1000	0.070	0.053	0.053	0.058	0.056	0.055	0.047	0.051	0.053	0.051	0.058	0.070	0.193	0.060
5000	0.063	0.051	0.051	0.048	0.043	0.044	0.044	0.034	0.050	0.062	0.051	0.048	0.120	0.089

4.1.1 Studying the power

To study the finite sample power of the tests, we generate replicates using the models in equations (15) and (16). 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 (15), we analyze their power against breaks in the conditional mean while for the DGP in (16), we study their power against misspecification in the conditional variance.

In Tables 3 and 4, we report the power results of the test t_{1,α_i} for each of these two DGPs, respectively, and in Table 5, we report the power results corresponding to the portmanteau tests. When the DGP has a break in the conditional mean (Table 3), the tests have high power (70-100%) for sample sizes of 300 and above. When the DGP is an AR(1)-GARCH(1,1) model (Table 4), we observe that the power of t_{1,α_i} is the highest in the extreme 1% and 5% autocontours. This is consistent with the results that we have explained in Figure 2, where the pairs of PITs tend to concentrate into the corners of the unit-square. These results suggest that we need larger sample sizes for the tests to have power against misspecification in the conditional variance.[5]

Regarding the portmanteau tests (Table 5), 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. As with the t_{1,α_i}, the power of $L^5_{\alpha_i}$ is higher in the extreme autocontours.[6]

[5]Note that this result could be expected as inference in nonlinear GARCH models require large samples.

[6]Results on the power when the DGP is the AR(2) model in (14) are reported in Tables C and D of the supplementary material. The proposed tests are very powerful even for small sample sizes.

Table 3

Monte Carlo power results for t_{1,α_i}. The DGP is the AR(1) model with break in the mean and $\varepsilon_t \sim N(0,1)$. The nominal size is 5%.

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														
	$\widehat{\alpha}_{k,i}$	0.001	0.014	0.040	0.103	0.176	0.264	0.359	0.462	0.577	0.701	0.830	0.897	0.948
	std	(0.005)	(0.016)	(0.026)	(0.041)	(0.049)	(0.055)	(0.059)	(0.059)	(0.060)	(0.052)	(0.045)	(0.038)	(0.027)
	$\bar{\sigma}^*_{\alpha_i}$	0.016	0.037	0.053	0.076	0.090	0.099	0.102	0.099	0.092	0.079	0.061	0.049	0.033
	power	0.000	0.000	0.041	0.105	0.131	0.144	0.137	0.145	0.142	0.090	0.098	0.080	0.163
100														
	$\widehat{\alpha}_{k,i}$	0.002	0.017	0.043	0.109	0.184	0.273	0.372	0.477	0.589	0.712	0.840	0.909	0.967
	std	(0.004)	(0.013)	(0.019)	(0.027)	(0.033)	(0.039)	(0.042)	(0.041)	(0.039)	(0.036)	(0.029)	(0.026)	(0.017)
	$\bar{\sigma}^*_{\alpha_i}$	0.010	0.024	0.035	0.050	0.059	0.064	0.065	0.063	0.058	0.050	0.037	0.029	0.018
	power	0.001	0.107	0.284	0.417	0.488	0.516	0.486	0.460	0.441	0.374	0.258	0.212	0.136
300														
	$\widehat{\alpha}_{k,i}$	0.002	0.018	0.046	0.114	0.193	0.283	0.381	0.485	0.599	0.719	0.849	0.917	0.976
	std	(0.003)	(0.007)	(0.011)	(0.016)	(0.020)	(0.022)	(0.024)	(0.025)	(0.024)	(0.021)	(0.018)	(0.014)	(0.008)
	$\bar{\sigma}^*_{\alpha_i}$	0.006	0.013	0.019	0.027	0.032	0.034	0.035	0.033	0.030	0.025	0.018	0.013	0.008
	power	0.000	0.785	0.927	0.985	0.989	0.994	0.995	0.994	0.986	0.969	0.874	0.716	0.421
1000														
	$\widehat{\alpha}_{k,i}$	0.002	0.019	0.047	0.116	0.196	0.287	0.385	0.490	0.604	0.724	0.851	0.920	0.979
	std	(0.002)	(0.004)	(0.006)	(0.009)	(0.011)	(0.012)	(0.013)	(0.013)	(0.013)	(0.012)	(0.009)	(0.008)	(0.004)
	$\bar{\sigma}^*_{\alpha_i}$	0.003	0.007	0.010	0.014	0.017	0.018	0.018	0.017	0.015	0.013	0.009	0.006	0.003
	power	0.824	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.911
5000														
	$\widehat{\alpha}_{k,i}$	0.003	0.019	0.047	0.116	0.197	0.287	0.387	0.491	0.604	0.724	0.853	0.921	0.980
	std	(0.001)	(0.002)	(0.003)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.004)	(0.003)	(0.002)
	$\bar{\sigma}^*_{\alpha_i}$	0.001	0.003	0.004	0.006	0.007	0.008	0.008	0.007	0.007	0.005	0.004	0.002	0.001
	power	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

Table 4

Monte Carlo power results for t_{1,α_i}. The DGP is the AR(1)-GARCH(1,1) model in (16) and $\varepsilon_t \sim N(0,1)$. The nominal size is 5%.

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														
	$\widehat{\alpha}_{k,i}$	0.018	0.063	0.111	0.209	0.312	0.418	0.522	0.622	0.719	0.815	0.908	0.953	0.991
	std	(0.019)	(0.039)	(0.056)	(0.078)	(0.089)	(0.096)	(0.093)	(0.086)	(0.077)	(0.061)	(0.041)	(0.031)	(0.019)
	$\bar{\sigma}^*_{\alpha_i}$	0.014	0.035	0.054	0.079	0.094	0.103	0.104	0.100	0.090	0.075	0.057	0.047	0.033
	power	0.204	0.073	0.048	0.034	0.022	0.025	0.019	0.012	0.014	0.010	0.006	0.006	0.009
100														
	$\widehat{\alpha}_{k,i}$	0.021	0.066	0.112	0.207	0.308	0.413	0.517	0.620	0.718	0.816	0.909	0.953	0.990
	std	(0.014)	(0.029)	(0.043)	(0.061)	(0.071)	(0.075)	(0.072)	(0.064)	(0.050)	(0.038)	(0.026)	(0.018)	(0.011)
	$\bar{\sigma}^*_{\alpha_i}$	0.010	0.025	0.039	0.058	0.069	0.074	0.074	0.069	0.060	0.047	0.033	0.025	0.018
	power	0.321	0.132	0.067	0.043	0.051	0.035	0.028	0.027	0.017	0.015	0.009	0.004	0.002
300														
	$\widehat{\alpha}_{k,i}$	0.023	0.066	0.112	0.206	0.306	0.410	0.516	0.618	0.718	0.816	0.908	0.953	0.989
	std	(0.009)	(0.022)	(0.031)	(0.041)	(0.046)	(0.048)	(0.045)	(0.039)	(0.031)	(0.021)	(0.012)	(0.008)	(0.005)
	$\bar{\sigma}^*_{\alpha_i}$	0.006	0.015	0.024	0.035	0.042	0.045	0.044	0.041	0.034	0.025	0.015	0.011	0.007
	power	0.542	0.232	0.094	0.052	0.045	0.044	0.038	0.043	0.040	0.051	0.031	0.006	0.006
1000														
	$\widehat{\alpha}_{k,i}$	0.024	0.066	0.111	0.204	0.303	0.408	0.514	0.616	0.717	0.814	0.908	0.953	0.989
	std	(0.006)	(0.011)	(0.016)	(0.023)	(0.026)	(0.027)	(0.026)	(0.021)	(0.017)	(0.012)	(0.006)	(0.004)	(0.002)
	$\bar{\sigma}^*_{\alpha_i}$	0.003	0.008	0.013	0.020	0.024	0.026	0.025	0.023	0.019	0.013	0.007	0.005	0.003
	power	0.929	0.494	0.152	0.061	0.048	0.051	0.066	0.070	0.105	0.159	0.145	0.070	0.031
5000														
	$\widehat{\alpha}_{k,i}$	0.024	0.066	0.110	0.202	0.302	0.406	0.512	0.615	0.716	0.814	0.908	0.954	0.989
	std	(0.003)	(0.007)	(0.010)	(0.014)	(0.016)	(0.017)	(0.016)	(0.015)	(0.012)	(0.007)	(0.003)	(0.002)	(0.001)
	$\bar{\sigma}^*_{\alpha_i}$	0.001	0.004	0.006	0.010	0.011	0.012	0.012	0.011	0.009	0.006	0.003	0.002	0.001
	power	0.999	0.925	0.419	0.119	0.098	0.113	0.197	0.330	0.482	0.692	0.724	0.476	0.110

Table 5

Monte Carlo power results for $L^5_{\alpha_i}$ and C_1 statistics. 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%.

	$L^5_{0.01}$	$L^5_{0.05}$	$L^5_{0.1}$	$L^5_{0.2}$	$L^5_{0.3}$	$L^5_{0.4}$	$L^5_{0.5}$	$L^5_{0.6}$	$L^5_{0.7}$	$L^5_{0.8}$	$L^5_{0.9}$	$L^5_{0.95}$	$L^5_{0.99}$	C^{13}_1
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

5 Out-of-sample h-step-ahead Bootstrap G-ACR (BG-ACR) tests

We extend the procedures and tests described in the previous section to obtain out-of-sample h-step-ahead densities. The in-sample bootstrap algorithm can also be applied to obtain bootstrap replicates of the out-of-sample multi-step-ahead observations. Then, the corresponding PITs and indicators can be computed. In order to compute the proportion $\widehat{\alpha}_{k,i}$, it is necessary to obtain $(H-h+1)$ h-step-ahead bootstrap forecast densities. 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 $h=1,2,...$ and $j=0,...,H-h$ obtain out-of-sample h-step-ahead conditional estimates of volatilities and observations as follows:

$$\begin{aligned}\sigma^{**2(b)}_{T+h+j|T+j} &= \widehat{\omega}^{*(b)}_0+\widehat{\omega}^{*(b)}_1(y^{**(b)}_{T+h-1+j|T+j}-\widehat{\mu}^{*(b)}-\widehat{\phi}^{*(b)}y_{T+h-2+j|T+j})^2+\widehat{\omega}^{*(b)}_2\sigma^{**2(b)}_{T+h-1+j|T+j},\\ y^{**(b)}_{T+h+j|T+j} &= \widehat{\mu}^{*(b)}+\widehat{\phi}^{*(b)}y^{**(b)}_{T+h-1+j|T+j}+\sigma^{**(b)}_{T+h+j|T+j}\varepsilon^{*(b)}_{T+h},\end{aligned} \tag{21}$$

where $y^{**(b)}_{T+i|T}=y_{T+i}$ when $i\leq 0$ and
$\sigma^{**2(b)}_{i|i}=\frac{\widehat{\omega}^{*(b)}_0}{1-\widehat{\omega}^{*(b)}_1-\widehat{\omega}^{*(b)}_2}+\widehat{\omega}^{*(b)}_1\sum_{j=0}^{i-3}\widehat{\omega}^{*(b)j}_2\left[(y_{i-j-1}-\widehat{\mu}^{*(b)}-\widehat{\phi}^{*(b)}y_{i-j-2})^2-\frac{\widehat{\omega}^{*(b)}_0}{1-\widehat{\omega}^{*(b)}_1-\widehat{\omega}^{*(b)}_2}\right]$ for $i=T,T+1,...,T+H-1$.

At each moment $T+j$, $j=h,...,H$, we compute out-of-sample multi-period PITs as follows

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

Note that the PITs based on h-step-ahead density forecasts will generally follow a moving average process of order $h-1$. When $h>1$, under the null that the predictive density coincides with the true density, the distributional features of the PITs are not well defined. As a result, it is only possible to test the null of a well behaved density forecast jointly with an assumed model of the process driving the associated h-step-ahead PITs. Alternatively,

one can choose PITs separated by h periods to ensure an uncorrelated sample. 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), Manzan and Zerom (2008) and Rossi and Sekhposyan (2016) among others.

Using the uncorrelated PITs $\{u_{T+ht|T+h(t-1)}\}_{t=1}^{[H/h]}$, we compute the corresponding indicators, I_{T+ht}^{k,α_i}, and the proportions

$$\widehat{\alpha}_{k,i} = \frac{\sum_{t=k+1}^{[H/h]} I_{T+ht}^{k,\alpha_i}}{H/h - k}.$$

Finally, the t-statistic is given by

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

where $\sigma_{\alpha_i}^2$ is defined as in (5). Note that $\sigma_{\alpha_i}^2$ can be estimated either as in expression (5) or by bootstrapping. As mentioned above, when testing the in-sample specification, ignoring parameter uncertainty may cause severe distortions in the size of the tests. However, 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 $\sigma_{\alpha_i}^2$ 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 (13) with $\phi_1 = 0.95$ and $\varepsilon_t \sim N(0,1)$. The model is estimated by OLS using $T = 50$, 100, 300, 1000 and 5000 observations and $H = 50$ and 500 out-of-sample one-step-ahead densities. Their corresponding PITs are obtained using the bootstrap procedure. The variance of $\widehat{\alpha}_{k,i}$ and the covariances in Λ_{α_i} and Ω_k are computed by bootstrapping.[7] In Table 6, we report the

[7]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.

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.[8]

[8]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 6

Monte Carlo size results for out-of-sample $L^5_{\alpha_i}$ and C_1 statistics. 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).

	$L^5_{0.01}$	$L^5_{0.05}$	$L^5_{0.1}$	$L^5_{0.2}$	$L^5_{0.3}$	$L^5_{0.4}$	$L^5_{0.5}$	$L^5_{0.6}$	$L^5_{0.7}$	$L^5_{0.8}$	$L^5_{0.9}$	$L^5_{0.95}$	$L^5_{0.99}$	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

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(2) model in (14) and from the AR(1)-GARCH(1,1) model in (16). Under the null hypothesis, we estimate an AR(1) process without drift. In Table 7, we report the power of the t_{1,α_i} test when the DGP is the AR(2) model and $H = 500$. For small estimation samples ($T = 50, 100$), the power is high in the middle 20%-70% autocontours. As the estimation sample grows, the power reaches 1 for most autocontours. When the information is accumulated either over several lags or over several quantiles, as in the $L^5_{\alpha_i}$ and C^{13}_1 tests (Table 9, Panel A), the power is very high even for small estimation samples. We report the power results of the t_{1,α_i} tests corresponding to the AR(1)-GARCH(1,1) in Table 8 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 Panel B of Table 9, we observe an acceptable power of around 60% for the C^{13}_1 statistic for estimation samples of $T = 300$ and above. The test $L^5_{\alpha_i}$ delivers similar power for the lowest 1% autocontour and for the highest 95-99% autocontours. It is important to note that in-sample tests are expected to be more powerful than out-of-sample tests; see, for example, Inoue and Kilian (2005).

Table 7

Monte Carlo power results for out-of-sample t_{1,α_i}. The DGP is the AR(2) model and $\varepsilon_t \sim N(0,1)$. The nominal size is 5% and $H = 500$.

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														
	$\widehat{\alpha}_{k,i}$	0.000	0.002	0.008	0.038	0.100	0.193	0.305	0.430	0.559	0.692	0.829	0.898	0.948
	std	(0.001)	(0.003)	(0.008)	(0.021)	(0.038)	(0.059)	(0.079)	(0.095)	(0.102)	(0.098)	(0.083)	(0.068)	(0.049)
	$\bar{\sigma}^*_{\alpha_i}$	0.024	0.043	0.056	0.069	0.077	0.083	0.086	0.086	0.084	0.078	0.064	0.051	0.036
	power	0.000	0.030	0.450	0.764	0.809	0.745	0.653	0.526	0.418	0.327	0.234	0.212	0.211
100														
	$\widehat{\alpha}_{k,i}$	0.000	0.002	0.010	0.047	0.115	0.211	0.324	0.449	0.579	0.711	0.845	0.914	0.967
	std	(0.000)	(0.003)	(0.007)	(0.019)	(0.032)	(0.047)	(0.059)	(0.069)	(0.076)	(0.073)	(0.060)	(0.047)	(0.030)
	$\bar{\sigma}^*_{\alpha_i}$	0.012	0.026	0.037	0.049	0.056	0.061	0.064	0.064	0.063	0.057	0.045	0.035	0.021
	power	0.000	0.415	0.874	0.956	0.952	0.898	0.795	0.644	0.482	0.353	0.261	0.226	0.196
300														
	$\widehat{\alpha}_{k,i}$	0.000	0.002	0.011	0.055	0.129	0.229	0.344	0.469	0.599	0.730	0.859	0.925	0.978
	std	(0.000)	(0.002)	(0.006)	(0.015)	(0.023)	(0.033)	(0.040)	(0.046)	(0.049)	(0.047)	(0.038)	(0.028)	(0.015)
	$\bar{\sigma}^*_{\alpha_i}$	0.007	0.016	0.023	0.032	0.038	0.042	0.044	0.044	0.043	0.038	0.029	0.022	0.012
	power	0.000	0.999	1.000	1.000	1.000	0.991	0.951	0.832	0.665	0.444	0.310	0.244	0.194
1000														
	$\widehat{\alpha}_{k,i}$	0.000	0.002	0.012	0.058	0.134	0.237	0.355	0.482	0.610	0.739	0.866	0.930	0.982
	std	(0.000)	(0.002)	(0.006)	(0.012)	(0.019)	(0.026)	(0.033)	(0.038)	(0.039)	(0.036)	(0.029)	(0.022)	(0.011)
	$\bar{\sigma}^*_{\alpha_i}$	0.005	0.013	0.018	0.026	0.031	0.034	0.035	0.035	0.033	0.029	0.023	0.017	0.008
	power	0.149	1.000	1.000	1.000	1.000	1.000	0.985	0.898	0.736	0.521	0.349	0.258	0.205
5000														
	$\widehat{\alpha}_{k,i}$	0.000	0.002	0.012	0.059	0.139	0.239	0.358	0.484	0.614	0.741	0.869	0.933	0.985
	std	(0.000)	(0.002)	(0.005)	(0.012)	(0.018)	(0.023)	(0.028)	(0.033)	(0.033)	(0.030)	(0.025)	(0.018)	(0.009)
	$\bar{\sigma}^*_{\alpha_i}$	0.005	0.012	0.017	0.023	0.028	0.030	0.031	0.031	0.029	0.026	0.020	0.014	0.007
	power	0.813	1.000	1.000	1.000	1.000	1.000	1.000	0.961	0.802	0.608	0.385	0.246	0.187

Table 8

Monte Carlo power results for out-of-sample t_{1,α_i}. The DGP is the AR(1)-GARCH(1,1) model with $\varepsilon_t \sim N(0,1)$. The nominal size is 5% and $H = 500$.

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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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														
	$\widehat{\alpha}_{k,i}$	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

Table 9

Monte Carlo power results for out-of-sample $L^5_{\alpha_i}$ and C_1 statistics with $H = 500$. The DGPs are: the AR(2) model with $\varepsilon_t \sim N(0,1)$ (Panel A); and the AR(1)-GARCH(1,1) model with $\varepsilon_t \sim N(0,1)$ (Panel B). The nominal size is 5%.

	$L^5_{0.01}$	$L^5_{0.05}$	$L^5_{0.1}$	$L^5_{0.2}$	$L^5_{0.3}$	$L^5_{0.4}$	$L^5_{0.5}$	$L^5_{0.6}$	$L^5_{0.7}$	$L^5_{0.8}$	$L^5_{0.9}$	$L^5_{0.95}$	$L^5_{0.99}$	C^{13}_1
Panel A														
50	0.288	0.773	0.978	1.000	1.000	1.000	1.000	1.000	0.987	0.919	0.713	0.521	0.348	0.596
100	0.418	0.928	0.999	1.000	1.000	1.000	1.000	1.000	0.996	0.964	0.751	0.526	0.297	0.728
300	0.565	0.999	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.980	0.827	0.573	0.415	0.989
1000	0.691	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.986	0.832	0.602	0.385	1.000
5000	0.721	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	0.845	0.624	0.327	1.000
Panel B														
50	0.371	0.150	0.111	0.096	0.122	0.128	0.157	0.189	0.223	0.259	0.377	0.448	0.528	0.385
100	0.445	0.185	0.122	0.087	0.092	0.118	0.153	0.186	0.242	0.297	0.378	0.500	0.558	0.495
300	0.525	0.242	0.110	0.088	0.095	0.108	0.132	0.164	0.252	0.341	0.441	0.491	0.599	0.569
1000	0.569	0.238	0.126	0.077	0.071	0.078	0.097	0.170	0.239	0.360	0.479	0.500	0.503	0.573
5000	0.602	0.277	0.115	0.064	0.071	0.079	0.113	0.161	0.231	0.322	0.484	0.551	0.412	0.600

6 Empirical application: Modeling VIX

There is an increasing interest in modeling and forecasting the volatility index VIX from the CBOE. The VIX was originally computed as the weighted average of the implied volatilities from eight at-the-money call and put options of the S&P100 index with an average time to maturity of 30 days. In 2003, the VIX was entirely revised by changing the reference index to the S&P500 index, taking into account a wide range of strike prices with the same time to maturity, and freeing its calculation from any specific option pricing model; see Whaley (2009) for a history of the VIX and Fernandes et al. (2014) for a detailed description of the VIX calculation. 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, which can be obtained from the conditional mean forecasts, but also the risk involved, which necessarily depends on features of the conditional density. In the context of VIX, Konstantinidi and Skiadopoulos (2011) implement the bootstrap procedure of Pascual et al. (2004) to obtain forecast intervals for the VIX that are then used in a trading strategy. Konstantinidi et al. (2008) and Psaradellis and Sermpinis (2016) also compare several specifications of the VIX for trading purposes.

It is commonly accepted that the VIX display long-memory; see, for example, Konstantinidi et al. (2008) and Fernandes et al. (2014). Consequently, several authors propose variants of the simple and easy-to-estimate long-memory Heterogeneous Autoregressive (HAR) model of Corsi (2009) to represent and predict the VIX; see Fernandes et al. (2014), Caporin et al. (2016) and Psaradellis and Sermpinis (2016). Alternatively, Mencía and Sentana (2016) propose modeling the persistence of the VIX using the Multiplicative Error Model (MEM) of Engle and Gallo (2006) with a semi-nonparametric expansion of the Gamma distribution.

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), 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 $y_t = \log V_t$[9]

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

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 equation (22) is an AR(66) model reparameterized in a parsimonious way by imposing economically meaningful restrictions. As in Corsi (2009), the parameters in equation (22) 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[10].

We compute the in-sample bootstrap conditional densities as described in Section 3. In Figure 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 with the normal densities, we also observe large distortions. The boostrap densities are more peaked than the normal densities and they are rightly skewed.

After computing the in-sample PITs, we plot the pairs (u_t,u_{t-1}) in Figure 4 (first row) together with 20% and 80% autocontours. We observe that they are not uniformly distributed on the unit square. There is a concentration of PITs in the left and right top corners, suggesting that conditional heteroscedasticity has not been modeled when computing the conditional densities for log-VIX. For comparison purposes, we also plot pairs of the PITs computed as in González-Rivera and Sun (2015) assuming that the errors are Gaussian (second column of Figure 4). We observe a concentration of pairs in the central section of the unit-square, suggesting that the HAR model is misspecified if Gaussian errors are assumed.

We formally test the null hypothesis of correct specification of the HAR model for the log-VIX. In Table 10, we report the sample proportions, $\widehat{\alpha}_{k,i}$, and the in-sample BG-ACR statistics t_{1,α_i}, $L^5_{\alpha_i}$ and C^{13}_1. We observe that the specification is strongly rejected from

[9]Fernandes et al. (2014) include explanatory variables in equation (22). However, we stick to an univariate model to simplify the implementation of the proposed testing procedure.

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

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, as suggested in Figure 4, the basic HAR model is not adequate to model the conditional densities of the daily log-VIX. In Table 10, we also report the corresponding tests assuming Gaussian errors. As expected from the information contained in Figure 4, 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, i.e. dynamics and conditional normality, provided by the HAR specification of the log-VIX is strongly rejected.

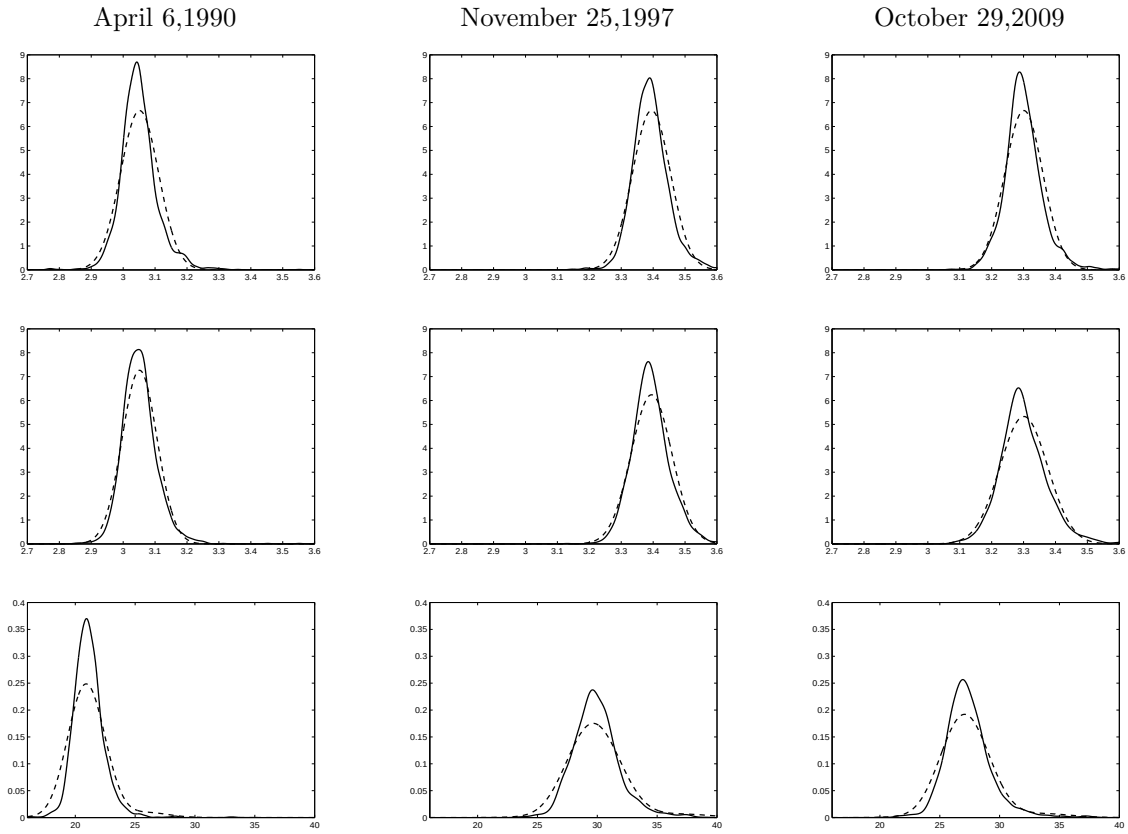

Figure 3: In-sample one-step-ahead densities obtained after fitting the HAR model (first row), the HAR-GJR model (second row) and the MEM model (third row) at three moments of time: April 6,1990 (first column), November 25,1997 (second column) and October 29,2009 (third column). The solid lines represent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

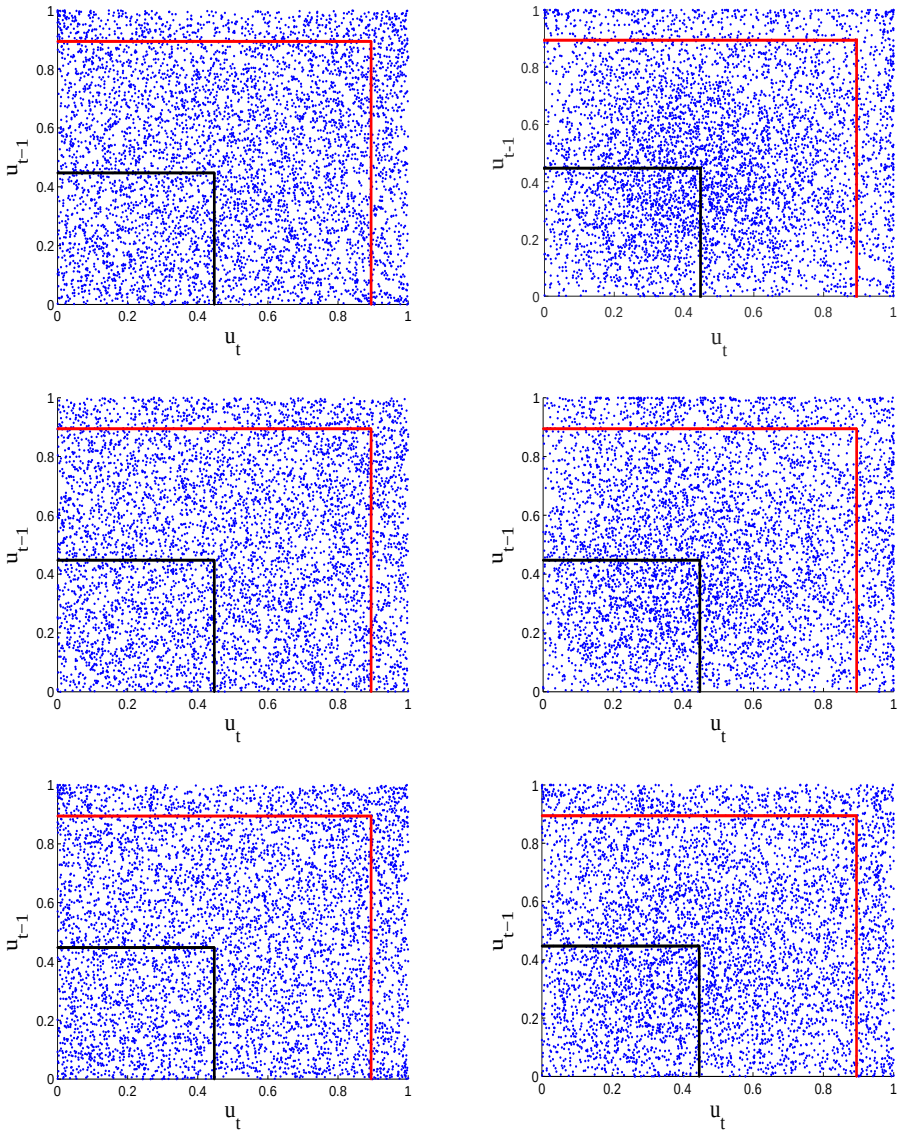

Figure 4: Pairs (u_t, u_{t-1}) and autocontours for the HAR model (first line), HAR-GJR model (second line) and MEM (third line). The PITS are obtained with the bootstrap procedure (first column) and assuming Gaussian errors for the HAR models and the GSNP distribution for the MEM (second column). $ACR_{20\%,1}$ corresponds to the black box and $ACR_{80\%,1}$ to the red box.

Table 10

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							
$\widehat{\alpha}_{k,i}$	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^{13}_1	36.747***												
						G-ACR							
$\widehat{\alpha}_{k,i}$	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^{13}_1	881.66***												
						HAR-GJR							
						BG-ACR							
$\widehat{\alpha}_{k,i}$	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^{13}_1	21.092*												
						G-ACR							
$\widehat{\alpha}_{k,i}$	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^{13}_1	318.896***												
						MEM							
						BG-ACR							
$\widehat{\alpha}_{k,i}$	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^{13}_1	18.511												
						G-ACR							
$\widehat{\alpha}_{k,i}$	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^{13}_1	103.417***												

Based on the information provided by the BG-ACR test and the autocontours plotted in Figure 4, we incorporate asymmetric conditional heteroscedasticity, and fit the HAR-GJR model in (22) with heteroscedastic ε_t as follows:

$$\begin{aligned} \varepsilon_t &= \sigma_t a_t, \\ \sigma_t^2 &= \omega_0 + \omega_1 \varepsilon_{t-1}^2 + \omega_2 \sigma_{t-1}^2 + \lambda \mathbf{1}(\varepsilon_{t-1} < 0)\varepsilon_{t-1}^2, \end{aligned} \tag{23}$$

where $\lambda < 2(1-\omega_1-\omega_2)$, $\omega_0 > 0$ and $\omega_1, \omega_2 \geq 0$ to guarantee the stationarity of ε_t and the positiveness of the conditional variance. a_t is an independent white noise sequence with variance 1; see Corsi et al. (2008), Bollerslev et al. (2009) and Huang et al. (2016) for HAR-GARCH specifications in the context of realized volatility.[11] 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 QML maximizing the Gaussian log-likelihood function; see McAleer et al. (2009) for the asymptotic properties of the QML estimator of the ARMA-GJR model.[12]

In Figure 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 becoming 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 Figure 4 (second row), we plot the PIT pairs (u_t,u_{t-1}) from the bootstrap conditional densities (first column) and assuming conditional normality (second column). Comparing both plots of PITs, we observe that, while they are uniformly distributed in the former case, the normal PITs are still concentrated in some areas of the unit-square. In Table 10, we report the corresponding statistics. The HAR-GJR model with bootstrap conditional densities is not rejected while the HAR-GJR with normal conditional densities is strongly

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

[12] Corsi and Renò (2012) and Todorova (2015) propose alternative specifications of the leverage in the context of HAR-GARCH models.

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 Mencía and Sentana (2016) that deals directly with the untransformed VIX, i.e. V_t. Their specification is the following, [13]

$$\begin{aligned} V_t - \Delta &= \mu_t \varepsilon_t, \\ \mu_t &= \varsigma_t + s_t, \\ \varsigma_t &= \varphi_0 + \varphi_1 \varsigma_{t-1} + \varphi_2 (V_{t-1} - \Delta - \mu_{t-1}), \\ s_t &= (\beta_1 + \beta_2) s_{t-1} + \beta_1 (V_{t-1} - \Delta - \mu_{t-1}), \end{aligned} \tag{24}$$

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 component ς_t reverts to $\varphi_0/(1-\varphi_1)$, the unconditional mean of V_{t-1}, while s_t reverts to zero. The noise ε_t is assumed to be i.i.d. whose density is given by the following semi-nonparametric expansion of order 2 of the Gamma density (GSNP)

$$f_{GSNP}(\varepsilon_t; \nu, \psi, \boldsymbol{\delta}) = \frac{1}{\Gamma(\nu)\psi^\nu} \varepsilon_t^{\nu-1} \exp(-\varepsilon_t/\psi) \left[\sum_{j=0}^{2} \delta_j \left(\frac{\varepsilon}{\psi} \right)^j \right]^2 \frac{1}{d}, \tag{25}$$

where $\Gamma(.)$ denotes the Gamma function, ν are the degrees of freedom, $\psi = d \left[\sum_{j=0}^{4} \gamma_j(\boldsymbol{\delta}) \frac{\Gamma(\nu+j+1)}{\Gamma(\nu)} \right]^{-1}$ is a scale parameter with $d = \sum_{j=0}^{4} \gamma_j(\boldsymbol{\delta}) \frac{\Gamma(\nu+j)}{\Gamma(\nu)}$ being a constant that ensures that the density integrates to 1, and $\gamma_j(\boldsymbol{\delta}) = \sum_{k=max(j-2,0)}^{min(j,2)} \delta_j \delta_{j-k}$, where δ_j are parameters to be estimated, such that $\delta_0 = 1$ and $\boldsymbol{\delta}'\boldsymbol{\delta} = 1$. The parameters of the MEM model are estimated by maximum likelihood.[14]

[13] Mencía and Sentana (2016) develop a theoretical framework to a dynamic portfolio allocation for Exchange Traded Notes tracking short- and mid-term VIX futures indices. They model the distribution of the future index returns conditional on past information and on the VIX, which in turn, is modeled by a MEM process with a GSNP distribution for the innovations. They conclude that the fit of the VIX seems to depend more on the assumed distribution whereas the fit of the futures prices seems to depend more on the dynamics of the conditional mean of the MEM model.

[14] We use the values of the parameters estimated in Mencía and Sentana (2016) as initial conditions for

In Figure 3 (third row), we plot 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 between the densities. The bootstrap densities are more skewed to the right and more peaked than the GSNP densities. It seems that GSNP densities assign more probability mass to the observations in the left tail.

In Figure 4 (third row), we plot the pairs of PITs (u_t, u_{t-1}) from the bootstrap conditional densities (first column) and from the assumed GSNP density (second column). We observe that the PITs obtained by the MEM-GSNP model are not uniformly distributed on the unit-square surface. The G-ACR statistics reported in Table 10 confirm these conclusions. The MEM-GSNP model is clearly rejected for almost all autocontours. Regarding the PITs from the bootstrap densities, they seem to be more uniformly distributed in the unit square though we observe some concentration of PITs in the corners, which indicates that some additional conditional heteroscedasticity model may be needed. In Table 10, 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.

our estimation. Estimation results are reported in the supplementary material.

7 Conclusions

We have proposed 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, yet the information on the bootstrap density contributes to the good properties of the proposed tests. 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 have evaluated the adequacy of conditional densities of the daily VIX index derived from the HAR and MEM models. We have strongly rejected 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 the most successful density model should take into account conditional heteroscedasticity of the error for an adequate construction of the conditional densities regardless of the specification used for the conditional mean.

Given that the proposed tests are based on the information contained in the vector of PITs that eventually is condensed into an indicator, they could be extended into a multivariate framework using the multivariate bootstrap procedures of Fresoli et al. (2015) and Fresoli and Ruiz (2016) for VARMA and multivariate GARCH models, respectively. It is also important to note that, in a multivariate context, the PITs with respect to a multivariate conditional density are not longer independent and uniform even if the model is correctly specified; see, for example, Chen and Hong (2014). In the context of multivariate GARCH models, Bai and Chen (2008) propose evaluating the distribution by using the

PITs of each individual component. However, this test may miss important information on the joint distribution and, in particular, may fail to detect misspecification in the joint dynamics.

Finally, the residual bootstrap implemented in this paper to obtain h-step-ahead predictive densities can be modified in several directions. First, one can extend it to cope with lag-order uncertainty of the ARMA lags by implementing the procedures of Kilian (1998), Alonso et al. (2004, 2006) and Fenga and Politis (2011). Another alternative is substituting the basic residual bootstrap to obtain the sample distribution of the parameters by the subsampling procedure proposed by Hall and Yao (2003). Alternatively, one can implement the block bootstrap based on resampling the likelihood proposed by Corradi and Iglesias (2008). Although we do not expect the results to change qualitatively, the asymptotic validity of the bootstrap could be easier to prove in the case of GARCH errors. A very interesting research extension could be implementing the new bootstrap approach proposed by Pan and Politis (2016) for bootstrap prediction intervals in linear AR models. This new approach is computationally fast because it does not need to generate pseudo-series alleviating the computational burden associated with bootstrapping to obtain the parameter estimator distribution and the variance of the BG-ACR statistics.

SUPPLEMENTARY MATERIAL

Supplementary Tables: Tables A and B report Monte Carlos results on the in-sample size of the BG-ACR tests when the DGP is the AR(1) with $\phi = 0.5$ and $\varepsilon_t \sim N(0,1)$. Tables C and D report Monte Carlo results on the in-sample power of the BG-ACR tests when the DGP is the AR(2) with $\varepsilon_t \sim N(0,1)$. Table E reports Monte Carlo results on the out-of-sample size of the BG-ACR tests when the DGP is the AR(1) with $\phi = 0.95$ and $\varepsilon_t \sim N(0,1)$. Table F report estimation results of the parameters of the HAR, HAR-GJR and MEM-GSNP models for the VIX index while Table G reports the corresponding residual diagnosis together with descriptive moments of the VIX. (Supplementary tables.pdf)

References

Alonso, A. M., D. Peña, and J. Romo (2004). Introducing model uncertainty in time series bootstrap. *Statistica Sinica 14*, 155–174.

Alonso, A. M., D. Peña, and J. Romo (2006). Introducing model uncertainty by moving blocks bootstrap. *Statistical Papers 47*, 167–179.

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

Bai, J. (2003). Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics 85*(3), 531–549.

Bai, J. and Z. Chen (2008). Testing multivariate distributions in GARCH models. *Journal of Econometrics 143*(1), 19–36.

Bollerslev, T., U. Kretschmer, C. Pigorsch, and G. Tauchen (2009). A discrete-time model for daily S & P500 returns and realized variations: Jumps and leverage effects. *Journal of Econometrics 150*(2), 151–166.

Caporin, M., E. Rossi, and P. Santucci de Magistris (2016). Volatility jumps and their economic determinants. *Journal of Financial Econometrics 1*(14), 28–29.

Chen, B. and Y. Hong (2014). A unified approach to validating univariate and multivariate conditional distribution models in time series. *Journal of Econometrics 178*, 22–44.

Corradi, V. and E. M. Iglesias (2008). Bootstrap refinements for QML estimators of the GARCH(1,1) parameters. *Journal of Econometrics 144*(2), 500–510.

Corradi, V. and N. Swanson (2006). Predictive density evaluation. In G. Elliot, C. Granger, and A. Timmerman (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 197–284. London: Elsevier.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 2*(7), 174–196.

Corsi, F. and R. Renò (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics 30*(3), 368–380.

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

Diebold, F. X., J. Hahn, and A. S. Tay (1999). Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange. *The Review of Economics and Statistics 81*(4), 661–673.

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

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association 82*(397), 171–185.

Engle, R. F. and G. M. Gallo (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics 131*(1), 3–27.

Fenga, L. and D. N. Politis (2011). Bootstrap-based arma order selection. *Journal of Statistical Computation and Simulation 81*(7), 799–814.

Fernandes, M., M. C. Medeiros, and M. Scharth (2014). Modeling and predicting the CBOE market volatility index. *Journal of Banking & Finance 40*, 1–10.

Francq, C. and J.-M. Zakoïan (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli 10*(4), 605–637.

Fresoli, D. and E. Ruiz (2016). The uncertainty of conditional returns, volatilities and correlations in DCC models. *Computational Statistics & Data Analysis 100*, 170–185.

Fresoli, D., E. Ruiz, and L. Pascual (2015). Bootstrap multi-step forecasts of non-Gaussian VAR models. *International Journal of Forecasting 31*(3), 834–848.

Gonçalves, S. and H. White (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *Journal of Econometrics 119*(1), 199–219.

González-Rivera, G., Z. Senyuz, and E. Yoldas (2011). Autocontours: Dynamic specification testing. *Journal of Business & Economic Statistics 29*(1), 186–200.

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

González-Rivera, G. and E. Yoldas (2012). Autocontour-based evaluation of multivariate predictive densities. *International Journal of Forecasting 28*(2), 328–342.

Granger, C. W. J. and M. H. Pesaran (2000a). A decision theoretic approach to forecast evaluation. In W. S. Chan, W. K. Li, and H. Tong (Eds.), *Statistics and Finance: An Interface*, Chapter 15, pp. 261–278. London: Imperial College Press.

Granger, C. W. J. and M. H. Pesaran (2000b). Economic and statistical measures of forecast accuracy. *Journal of Forecasting 19*, 537–560.

Hall, P. and Q. Yao (2003). Inference in ARCH and GARCH models with heavy–tailed errors. *Econometrica 71*(1), 285–317.

Hidalgo, J. and P. Zaffaroni (2007). A goodness-of-fit test for ARCH(∞) models. *Journal of Econometrics 141*(2), 973–1013.

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

Huang, Z., H. Liu, and T. Wang (2016). Modeling long memory volatility using realized measures of volatility: A realized HAR GARCH model. *Economic Modelling 52*, 812–821.

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

Kilian, L. (1998). Accounting for lag order uncertainty in autoregressions: The endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis 19*(5), 531–548.

Konstantinidi, E. and G. Skiadopoulos (2011). Are VIX futures prices predictable? An empirical investigation. *International Journal of Forecasting 27*(2), 543–560.

Konstantinidi, E., G. Skiadopoulos, and E. Tzagkaraki (2008). Can the evolution of implied volatility be forecasted? Evidence from European and US implied volatility indices. *Journal of Banking & Finance 32*(11), 2401–2411.

Manzan, S. and D. Zerom (2008). A bootstrap-based non-parametric forecast density. *International Journal of Forecasting 24*, 535–550.

Martin, I. (2017). What is the expected return on the market? *The Quarterly Journal of Economics 132*(1), 367–433.

McAleer, M., S. Hoti, and F. Chan (2009). Structure and asymptotic theory for multivariate asymmetric conditional volatility. *Econometric Reviews 28*(5), 422–440.

Mencía, J. and E. Sentana (2016). Volatility-related exchange traded assets: An econometric investigation. *Journal of Business & Economic Statistics, in press*.

Mika, M. and P. Saikkonen (2011). Parameter estimation in nonlinear AR–GARCH models. *Econometric Theory 27*(6), 1236–1278.

Mitchell, J. and K. F. Wallis (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics 26*(6), 1023–1040.

Pan, L. and D. N. Politis (2016). Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *Journal of Statistical Planning and Inference 177*, 1–27.

Park, Y.-H. (2016). The effects of asymmetric volatility and jumps on the pricing of VIX derivatives. *Journal of Econometrics 192*(1), 313–328.

Pascual, L., J. Romo, and E. Ruiz (2004). Bootstrap predictive inference for ARIMA processes. *Journal of Time Series Analysis 25*, 449–465.

Pascual, L., J. Romo, and E. Ruiz (2006). Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics & Data Analysis 50*, 2293–2312.

Psaradellis, I. and G. Sermpinis (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.

Reeves, J. J. (2005). Bootstrap prediction intervals for ARCH models. *International Journal of Forecasting 21*(2), 237–248.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*(3), 470–472.

Rossi, B. and T. Sekhposyan (2013). Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics 177*, 199–212.

Rossi, B. and T. Sekhposyan (2016). Alternative tests for correct specification of conditional predictive densities. *Available at SSRN: https://ssrn.com/abstract=2283980*.

Shimizu, K. (2010). *Bootstrapping Stationary ARMA-GARCH Models*. Springer, Vieweg.

Shimizu, K. (2013). The bootstrap does not always work for heteroscedastic models. *Statistics & Risk Modeling 30*(3), 189–204.

Shimizu, K. (2014). Bootstrapping the nonparametric ARCH regression model. *Statistics & Probability Letters 87*, 61–69.

Song, Z. and D. Xiu (2016). A tale of two option markets: Pricing kernels and volatility risk. *Journal of Econometrics 190*(1), 176–196.

Tay, A. S. and K. F. Wallis (2000). Density forecasting: A survey. *Journal of Forecasting 19*, 235–254.

Todorova, N. (2015). The course of realized volatility in the lme non-ferrous metal market. *Economic Modelling 51*, 1–12.

Whaley, R. E. (2000). The investor fear gauge. *Journal of Portfolio Management 26*(3), 12–17.

Whaley, R. E. (2009). Understanding the VIX. *Journal of Portfolio Management 35*(3), 98–105.