NÁRODNÁ BANKA SLOVENSKA
EUROSYSTÉM

Assessing Distributional Properties of Forecast Errors

Marián Vávra

Working Paper 3/2018

www.nbs.sk
Imricha Karvaša 1
813 25 Bratislva

research@nbs.sk

July 2018

ISSN 2585-9269

NÁRODNÁ BANKA SLOVENSKA
EUROSYSTÉM

Assessing Distributional Properties of Forecast Errors[1]

Working paper NBS

Marián Vávra[2]

Abstract

This paper considers the problem of assessing the distributional properties (normality and symmetry) of macroeconomic forecast errors of G7 countries for the purpose of fan-chart modelling. Test statistics based on a Cramér von-Mises distance are used with critical values obtained via a bootstrap. Our results indicate that the assumption of symmetry of the marginal distribution of forecast errors is reasonable whereas the assumption of normality is not.

JEL classification: C12; C15; C22; C53
Key words: Normality; Symmetry; Forecast errors; Prediction intervals; Bootstrap

[1] I would like to thank Zacharias Psaradakis, Ron Smith, and Peter Tulip for useful comments and interesting suggestions. All remaining errors are only mine.
[2] Marián Vávra, Research Department of the NBS.

1. INTRODUCTION

Due to uncertainty surrounding point forecasts, there is a general consensus in the literature that a central bank maximizing the probability of achieving its goal should adopt some form of "density forecasting" when conducting monetary policy (see Greenspan (2003)). Many central banks thus nowadays calculate and officially publish prediction intervals for key economic variables (e.g. inflation and output) in order to express and communicate perceived forecast risks with professionals and the general public.[3] Two approaches have become popular in the forecasting industry: (i) asymmetric prediction bands[4]; and (ii) Gaussian prediction bands[5]. In the former case, the marginal prediction bands are calculated using a mixture of two Gaussian distributions with the same means but different variances. Apart from being intuitive and easy to calculate, this approach allows us to take into account unbalanced risks of the future development of exogenous factors. However, this approach implicitly relies on marginal symmetry of historical forecast errors. In the latter case, marginal prediction bands explicitly rely on an assumption that forecast errors are normally distributed. Clearly, if the assumption of symmetry and/or normality is violated, then the prediction intervals are subject to misspecification. This fact can, in turn, give rise to economic policy misperception and erroneous policy decisions. For example, based on the officially reported prediction bands prior to the Great Recession period, most economists and central bankers did not view price deflation and the zero lower bound of interest rates as a problem (see Tetlow and Tulip (2008)). We hold the view that using empirical quantiles from the appropriate distribution (altogether with correctly calculated long-run variance of forecast errors) can substantially improve the performance of fan-charts.

Unfortunately, many practitioners are reluctant to test for the distributional assumptions when calculating prediction intervals. We presume that this reluctance has something to do with the fact that both normality and symmetry tests with appropriate critical values valid under (weak) dependence of observations have not yet been fully implemented in widely used software packages. Given these considerations, it is desirable to provide reliable empirical evidence about the distributional properties of macroeconomic forecasting errors which can then be used for fan-chart modelling.

Although some work on testing for normality of forecast errors has already been done in the literature (see, e.g., Lahiri and Teigland (1987), Makridakis and Winkler (1989), Harvey and Newbold (2003), Reifschneider and Tulip (2007)), the existing results should be treated with caution. For example, Reifschneider and Tulip (2007) assess normality of the US Federal Re-

[3] Hammond et al. (2012) surveys the (inflation) reports of 27 central banks out of which 20 banks provide prediction intervals officially.

[4] This approach has been implemented in some way at the Bank of England, National Bank of Slovakia, South African Reserve Bank, National Bank of Poland, Hungarian National Bank, International Monetary Fund, World Bank, among others.

[5] This approach has been implemented in some way at the Bank of Canada, Czech National Bank, Riksbank, Norges Bank, European Central Bank, among others.

serve System forecast errors using the skewness-kurtosis test based on the asymptotic critical values derived for independently and identically distributed (i.i.d.) observations. As a result, the test gives very likely incorrect inference for dependent observations, including forecast errors where serial correlation increases with the forecast horizon (see Section 4 for details). Using the original skewness-kurtosis test might be justified only for serially uncorrelated forecast errors but not in general.[6] Harvey and Newbold (2003) assess normality of both individual and aggregated errors from the US Survey of Professional Forecasters based on formal testing for excess kurtosis with the Monte Carlo-based critical values. These may improve small sample properties of the test in the case of i.i.d. observations but fail in the case of serial dependence which is clearly the case of empirically observed macroeconomic forecast errors. At least to the best of our knowledge, no results for assessing symmetry of the marginal law of macroeconomic forecast errors are available in the literature.

This study makes two contributions. First, we assess both normality and symmetry of an international panel of survey-based macroeconomic forecast errors using the Cramér von-Mises type test statistics with appropriate critical values obtained via a bootstrap. The main task here is to provide reliable empirical evidence about the distributional properties of macroeconomic forecasting errors which can then be used for fan-chart modelling. The dataset employed in our study represents a unique data source which enables us to analyze forecast errors of two key macroeconomic variables for G7 countries over a long time period, something which is of practical importance for central banks and other forecasting institutions.[7] Second, a MATLAB code is made publicly available to researchers.

The paper is organized as follows. The statement of the problem and the relevant test statistics are discussed in Section 2. The bootstrap method to obtain appropriate critical values is described in Section 3. An international dataset of forecast errors is discussed in Section 4. The empirical results are presented in Section 5. Section 6 summarizes and concludes.

2. Test Statistics

Suppose $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ are consecutive observations from a stationary stochastic process $\mathcal{X} = \{X_t\}_{t\in\mathbb{Z}}$ satisfying

$$X_t - \mu = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad t \in \mathbb{Z}, \tag{1}$$

for some $\mu \in \mathbb{R}$, where $\{\psi_j\}_{j\in\mathbb{Z}^+}$ is a square-summable sequence of real numbers (with $\psi_0 = 1$) and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a sequence of i.i.d., real-valued, zero-mean random variables with variance

[6] It is only fair to note that the authors are aware of this shortcoming (see Reifschneider and Tulip (2007, pp. 19-20)).

[7] Theoretically, it would be more policy relevant to assess the distributional properties of the central banks' forecast errors. Practically, it would be infeasible to compile a dataset comparable with the one employed in the study.

$s^2 \in (0, \infty)$.

The first objective is to test the null hypothesis that the one-dimensional marginal distribution F of $\mathcal{X}$ is Gaussian, that is

$$\mathcal{H}_0^N : F = N(\mu, \sigma^2). \tag{2}$$

The alternative hypothesis is that distribution F of $\mathcal{X}$ is non-Gaussian.[8] The second objective is to test the null hypothesis that the one-dimensional marginal distribution F of $\mathcal{X}$ is symmetric around the center μ, that is

$$\mathcal{H}_0^S : F(x - \mu) = 1 - F(\mu - x). \tag{3}$$

The alternative hypothesis is that distribution F of $\mathcal{X}$ is asymmetric.[9]

A popular and consistent class of (nonparametric) tests for assessing both null hypotheses of normality and symmetry of the marginal law is based on the distance between the empirical distribution function and its theoretical counterpart (see, e.g., Rothman and Woodroofe (1972), Aki (1981), Boos (1982), Koziol (1983), Lilliefors (1967), Stephens (1976) for i.i.d. data; Psaradakis (2003), Psaradakis and Vávra (2017) for dependent data). In partiular, the test statistics for the null hypotheses in (2) and (3) considered here are based on the Cramér von-Mises distance

$$\mathcal{D}_N = n \int_{\mathbb{R}} [F_n(\hat{\mu} + \hat{\sigma}x) - \Phi(x)]^2 \, \mathrm{d}\Phi(x), \tag{4}$$

$$\mathcal{D}_S = n \int_{\mathbb{R}} [F_n(x) + F_n(2\hat{\mu} - x) - 1]^2 \, \mathrm{d}F_n(x), \tag{5}$$

where $F_n(x) = n^{-1} \sum_{t=1}^{n} \mathbb{I}(X_t \leq x)$, for each $x \in \mathbb{R}$, is the empirical distribution function of $\mathcal{X}_n$, Φ denotes the standard normal distribution function, $\hat{\mu}$ and $\hat{\sigma}$ denote the estimated location and scale parameters respectively. Simple computing forms of the test statistics can be found in Anderson and Darling (1954) and Rothman and Woodroofe (1972).

The asymptotic null distribution of distance-based statistics such as $\mathcal{D}_N$ and $\mathcal{D}_S$ is complicated to obtain, unless one is dealing with the classical problem of testing a simple null hypothesis (μ and σ known) for i.i.d. data. In our case, inference is complicated by the presence of both estimated parameters and dependence in $\mathcal{X}_n$. To complicate matters further, the distribution of the above distance test statistics depends on estimators of the unknown parameters used in (4) and (5) (see, e.g., Aki (1981)).

As a practical way of circumventing the problems mentioned above, we propose to use an autoregressive sieve bootstrap procedure to obtain P-values and/or critical values for the distance test statistics. The principal advantage of the sieve bootstrap is that it can be used to

[8] Note that the null hypothesis can be alternatively stated as: $\mathcal{H}_0^N : F = N(0, \sigma^2)$ since the forecast errors should be zero mean stochastic processes. However, empirical evidence suggests that the forecast errors are biased in small samples. The official forecasts are thus corrected for historically observed biases in forecast errors. Therefore, we inspect the stochastic properties of the errors beyond the first moment in this study.

[9] Note that the null hypothesis can be alternatively stated as: $\mathcal{H}_0^S : F(x) = 1 - F(-x)$ since the forecast errors should be zero mean stochastic processes. See Footnote 6 for an explanation.

approximate the sampling properties of both $\mathcal{D}_N$ and $\mathcal{D}_S$ without knowledge or estimation of the dependence parameter in data. Moreover, because bootstrap approximations are constructed from replicates of the test statistics themselves, there is no need to derive analytically, nor to make assumptions about, the appropriate norming factors for the distance statistics or their asymptotic null distributions, something which is very convenient in practice.

3. AUTOREGRESSIVE SIEVE BOOTSTRAP APPROXIMATION

The autoregressive sieve bootstrap is motivated by the observation that, under (1) and an additional assumption of invertibility, $\mathcal{X} = \{X_t$ admits the representation

$$\sum_{j=0}^{\infty} \phi_j (X_{t-j} - \mu) = \varepsilon_t, \qquad t \in \mathbb{Z}, \tag{6}$$

for a square-summable sequence of real numbers $\{\phi_j\}_{j\in\mathbb{Z}_+}$ (with $\phi_0 = 1$) such that $\phi(z) = \sum_{j=0}^{\infty} \phi_j z^j$ for $|z| < 1$.[10] The idea is to approximate (6) by a finite-order autoregressive model and use this as the basis of a semi-parametric bootstrap scheme. If the order of the autoregressive approximation is allowed to increase simultaneously with n at an appropriate rate, the distribution of the process in (6) will be matched asymptotically (see Kreiss (1992) and Bühlmann (1997)).

The bootstrap procedure used to approximate the sampling properties of the distance statistics $\mathcal{D}_N$ and $\mathcal{D}_S$ under the null hypotheses can be described in the following steps. (Only for notational simplicity, the distance test statistic is denoted as $\mathcal{D}$. Any computational differences between the $\mathcal{D}_N$ and $\mathcal{D}_S$ are stated explicitly):

S1. For some integer $p \geq 1$ (chosen as a function of n so that $p^{-1} + n^{-1}p \to 0$ as $n \to \infty$), compute the pth order least-squares estimate $(\hat{\phi}_{p1}, \ldots, \hat{\phi}_{pp})$ of the autoregressive coefficients for $\mathcal{X}$ by minimizing

$$(n-2p)^{-1} \sum_{t=p+1}^{n} \left\{ (X_t - \hat{\mu}) - \sum_{j=1}^{p} \phi_{pj} (X_{t-j} - \hat{\mu}) \right\}^2, \tag{7}$$

where $\hat{\mu}$ denotes a sample average calculated from $\mathcal{X}_n$ and $\{\hat{\varepsilon}_t\}_{t=p+1}^{n}$ denotes a sequence of the estimated residuals.

S2. By setting initial values $X^*_{-p+1} = \cdots = X^*_0 = \hat{\mu}$, generate bootstrap pseudo-observations

[10] It is important to point out that, as discussed in Poskitt (2007), the autoregressive representation (6) provides a meaningful approximation even if $\psi(z)$ has zeros in the unit disc $|z| < 1$.

$(X_1^*, \ldots, X_n^*)$ via the recursion

$$X_t^* - \hat{\mu} = \sum_{j=1}^{p} \hat{\phi}_{pj}(X_{t-j}^* - \hat{\mu}) + a_t^*, \qquad t = 1, 2, \ldots, n+b, \tag{8}$$

where the a_t^*'s are i.i.d. random variables having mean zero and drawn from the empirical distribution function which is selected based on the purpose of the analysis:

- in the case of the null of normality (i.e. $\mathcal{H}_0^N$), $\{a_t^*\}$ are i.i.d. errors drawn from $N(0, \hat{s}_p^2)$, where $\hat{s}_p^2$ is the minimum value of (7);
- in the case of the null of symmetry (i.e. $\mathcal{H}_0^S$), $\{a_t^*\}$ are i.i.d. errors drawn from the symmetrized empirical distribution function of residuals given by

$$\hat{G}_n(x) = (n-p)^{-1} \sum_{t=p+1}^{n} \mathbf{I}(\zeta_t \hat{\varepsilon}_t \leq x), \quad \text{for} \quad x \in \mathbb{R},$$

  where $\{\hat{\varepsilon}_t\}$ is a sequence of the estimated residuals from step S1 and $\{\zeta_t\}$ is a sequence of i.i.d. random variables drawn from the discrete uniform distribution on -1 and 1.

Then discard the initial b replicates to eliminate start-up effects (this procedure, with $b = 100$, is used in the sequel). Define the bootstrap analogue of $\mathcal{D}$ by the plug-in rule as $\mathcal{D}^*$ calculated using the appropriate test statistic with $\mathcal{X}_n^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$ replacing $\mathcal{X}_n$.

S3. Repeat step S2 independently B times to obtain a collection of B replicates $\{\mathcal{D}_1^*, \ldots, \mathcal{D}_B^*\}$ of $\mathcal{D}^*$. The sampling distribution of $\mathcal{D}$ is then approximated by the empirical distribution function associated with $\{\mathcal{D}_1^*, \ldots, \mathcal{D}_B^*\}$, that is $\hat{H}^*(x) = B^{-1} \sum_{i=1}^{B} \mathbb{I}(\mathcal{D}_i^* \leq x)$, for $x \in \mathbb{R}$. Then, a bootstrap test rejects the null hypothesis at the significance level α if $\mathcal{D} > \inf\{x : \hat{H}^*(x) \geq 1-\alpha\}$, where $\mathcal{D}$ is a value of the test statistic obtained from the observed sample $\mathcal{X}_n$.

Consistency of the sieve bootstrap estimator of the null sampling distribution of $\mathcal{D}$ follows from Lemma 1, Theorem 2 and Remark 2 of Poskitt (2008) under a suitable assumption about the rate of increase of p and the fractional parameter d ($d = 0$ in our setup). More specifically, let $\rho(H, H^*) = \sqrt{\int_0^1 |H^{-1}(u) - H^{*-1}(u)|^2 \, \mathrm{d}u}$ stand for the Mallows–Wasserstein distance between the distribution function H of $\mathcal{D}$ and the conditional distribution function H^* of $\mathcal{D}^*$ given $\mathcal{X}_n$ (where $g^{-1}(u) = \inf\{x : g(x) \geq u\}$ for any non-decreasing function g).[11] Then, if $\mathcal{X}$ satisfies (1), the distribution of ε_0 is either Gaussian or symmetric, and $p \to \infty$ and $(\log n)^{-\nu} p = O(1)$ as $n \to \infty$ for some $\nu \geq 1$, we have $\rho(H, H^*) \to 0$ with probability 1 as $n \to \infty$.

Some remarks about the bootstrap procedure are in order.

[11] While H^* is unknown, an approximation (of any desired accuracy) can be obtained by Monte Carlo simulation as $B \to \infty$.

(i) The order of the autoregressive sieve may be selected as the minimizer of Akaike's information criterion $\mathrm{AIC}(p) = \log \hat{s}_p^2 + 2n^{-1}p$ over $1 \leq p \leq p_{\max}$ for some suitable maximal order $p_{\max}$. Under mild regularity conditions, a data-dependent choice of p based on AIC is asymptotically efficient (in the sense of Shibata (1980)) and satisfies (with probability 1) the growth conditions required for the asymptotic validity of the sieve bootstrap for a large class of statistics (Psaradakis (2016)). Alternative criteria for order selection that may be used include, among many others, the Bayesian information criterion and the Mallows criterion.

(ii) Although least-squares estimates $(\hat{\phi}_{p1}, \ldots, \hat{\phi}_{pp})$ of the parameters of the approximating autoregression are used in the above algorithm to construct X_t^*, asymptotically equivalent estimates, such as those obtained from the empirical Yule–Walker equations, may alternatively be used. The Yule–Walker estimator is theoretically attractive because it guarantees that the bootstrap pseudo-observations $(X_1^*, \ldots, X_n^*)$ are generated from a causal (bootstrap) autoregressive process, but is known to be biased in small samples compared to the least-squares estimator (see, e.g., Tjøstheim and Paulsen (1983) and Paulsen and Tjøstheim (1985)).

(iii) It is worth noting that, by requiring a_t^* in (8) to be either normally or symmetrically distributed, $\mathcal{X}^*$ is constructed in a way which reflects the particular null hypothesis under test even though $\mathcal{X}$ may not satisfy it. This is important for ensuring that the bootstrap test has reasonable power against departures from the null (see Lehmann and Romano (2005, Sect. 15.6)).

We conclude this section by remarking that the linear structure imposed by (1) and (6) may arguably be considered as somewhat restrictive. However, the results of Bickel and Bühlmann (1997) suggest that linearity may not be too onerous a requirement in the sense that the closure (with respect to the total variation metric) of the class of linear processes is quite large; roughly speaking, for any stationary non-linear process, there exists another process in the closure of linear processes having identical sample paths with probability exceeding 0.36. This suggests that the autoregressive sieve bootstrap is likely to yield reasonably good approximations within a class of processes larger than that associated with (1) or (6).

4. Forecasts Errors

Since 1989, Consensus Economics Inc. (CE) has been conducting surveys which poll around 10 – 30 fairly diverse economists and financial analysts in each country on their views about the expected development of the selected macroeconomic and financial variables. CE currently operates with more than 1000 economic variables from over 85 countries. The CE output is often seen as a forecast benchmark by investment and planning managers, as well as government institutions. In addition to their annual (fixed-event) forecasts, the company regularly asks country panellists to provide also quarterly (fixed-horizon) economic forecasts for one up to eight quarters ahead. The aggregate CE forecast is a sample average of the forecasts provided by country participants for each economic variable. These quarterly forecasts are

updated every March, June, September, and December.

Here we focus on assessing normality and symmetry of the marginal distribution of the aggregate CE forecast errors for the real GDP growth rate (denoted as GDP) and the inflation rate (denoted as CPI) of G7 countries (United States, Japan, Germany, France, United Kingdom, Italy, Canada). Forecasts of both economic variables are reported in the form of year-on-year percentage changes.[12] The forecast horizon of variables under consideration is from one quarter to six quarters ahead. The dataset of international forecast errors is a balanced panel spanning the period Q4 1994 – Q3 2015 (i.e. 84 observations).

The aggregate CE forecast error is calculated as: $X_t(h) = Y_{t+h} - \hat{Y}_t(h)$, where Y_{t+h} denotes the realization of a given variable (e.g. GDP) at $t+h$ and $\hat{Y}_t(h)$ denotes the aggregate conditional forecast (a sample average of country participants' forecasts) made at time t for h periods ahead. It is far from clear which version of the realization should be actually used for our analysis. Our approach relies on the last vintage of data (final data) as the most accurate realization available to researchers. In particular, the March 2017 vintage of all macroeconomic variables is used when calculating forecast errors. Data are obtained from the OECD database.

The aggregate CE forecast errors of both macroeconomic variables for the selected horizons $h = 1, 3, 5$ are depicted in Figure A in Appendix A. Although both CPI and GDP errors display slightly different patterns, a characteristic feature is their high persistence which increases with the forecast horizon h – see Figure 2 where the first-order autocorrelation coefficients of forecast errors are depicted in the form of box-and-whisker diagrams. It is worth remarking here that the relatively high persistence of one-step ahead forecasts (i.e. $h = 1$) is caused by the fact that variables are expressed in (more policy relevant) year-on-year rather than quarter-on-quarter form. The figure clearly demonstrates that using test statistics based on the assumption of serially uncorrelated observations would provide very likely misleading inference about the distributional properties of forecast errors. It is shown in Appendix B that the distance-based tests with appropriate critical values obtained via the sieve bootstrap perform very well under both the null and alternative hypotheses even in sample sizes encountering macroeconomic applications.

5. Empirical Results

In this section, we apply the distance-based $\mathcal{D}_N$ and $\mathcal{D}_S$ tests based on (4) and (5) in order to assess normality and symmetry of the marginal law of macroeconomic forecast errors. A sample mean and standard deviation calculated from the sample $\mathcal{X}_n$ are used as estimators of location μ and scale σ. Recall that $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ represents a sample of the aggregate CE forecast errors spanning the period Q4 1994 – Q3 2015 ($n = 84$) for each variable (CPI,

[12] Note that quarter-on-quarter percentage changes of economic variables are not considered here since they are available only for GDP but not for CPI.

GDP) and each forecast horizon ($h = 1, \ldots, 6$). The bootstrap P-values of the distance tests are reported in Tables 1 – 2. These are computed from 1000 bootstrap replications with the data-dependent sieve order p determined using the AIC over $1 \leq p \leq \lfloor 5 \log_{10} n \rfloor$, where $\lfloor \cdot \rfloor$ denoting the greatest-integer function.[13]

The null hypothesis of normality is rejected (at the 10% significance level) in 52% (26%) of the GDP (CPI) forecast errors (see Table 1).[14] A lower rejection rate in the case of CPI errors can be explained to some extent by (in general) higher persistence of inflation forecast errors, which results in a lower power of the test statistics. Two interesting conclusions emerge when focusing on the normality results over the forecast horizon h. First, the GDP and CPI results are, rather surprisingly, in complete opposite for the one-step ahead errors (i.e. $h = 1$): the null is rejected in 70% of the CPI errors, whereas no rejection of the null for GDP errors. Second, apart from the one-step ahead errors, the normality results are quite stable over the forecast horizon, although the rejection rates for GDP are substantially higher than those for CPI. See Figure 1 (a) where the rejection frequencies (aggregated over all seven countries) are depicted across different forecast horizons. It is worth remarking here that only in two cases (i.e. GDP for Canada and CPI for Japan) the null hypothesis of normality is not rejected in either of the forecast horizon.

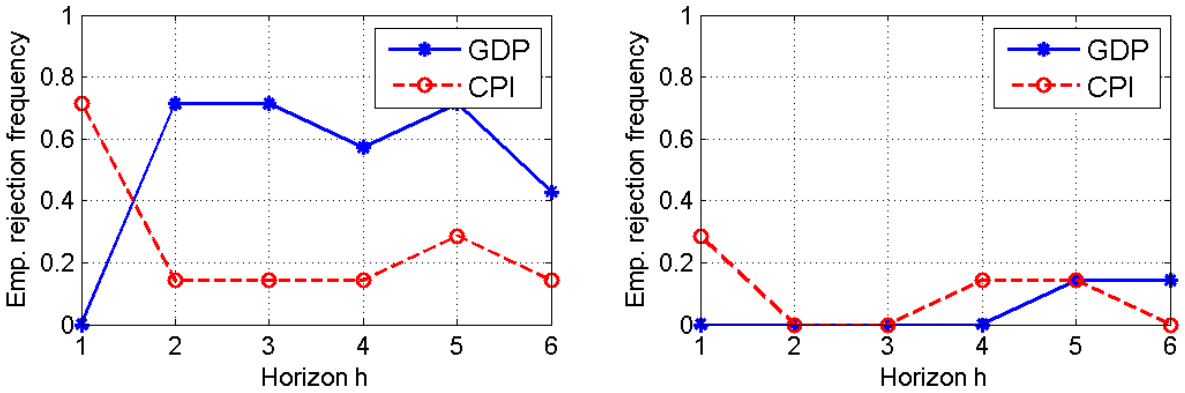
Noticeably different results are obtained when testing for the null of symmetry of the marginal distribution of macroeconomic forecast errors. In particular, the overall rejection frequency of symmetry is 5% (10%) for the GDP (CPI) errors (see Table 2). Interestingly, in most of the cases, the null is rejected for the UK forecast errors. The rejection being so low, it is quite clear that there is little to say about the behaviour of the test results over the forecast horizon. These are depicted in Figure 1 (b) for completeness.

It can be concluded from the previous results that normality is rejected mainly due to presence of heavy-tails in macroeconomic forecast errors. Therefore, using the Student distribution might be preferred against the Gaussian one for the purpose of fan-chart modelling. Of course, an important practical consideration for the application of the Student distribution is the selection of the right degrees of freedom. This can be done in a two step procedure as follows: in the first step, calibrate the degrees of freedom based on sample kurtosis calculated from forecast errors; in the second step, use a test similar to (4), where $\Phi(\cdot)$ is replaced with the Student distribution with the calibrated degrees of freedom, to test the null hypothesis of the adequacy of the Student distribution.

[13] It is worth remarking here that in cases where forecast errors are obtained from a parametric model (e.g. an ARMA model) then one observes $\{X_t(\hat{\boldsymbol{\theta}})\}$ rather than $\{X_t\}$, where $\hat{\boldsymbol{\theta}}$ denotes a vector of the estimated parameters. In such cases, parameter uncertainty should be taken into account when calculating the bootstrap P-values. Since survey-based data are not outcomes from a particular parametric model, no parameter uncertainty correction is considered in this paper.

[14] We hold the view that the 10% significance level is probably more appropriate in forecasting applications where a limited number of observations is available to researchers.

Figure 1: Aggregate Empirical Rejection Rates



(a) Normality (b) Symmetry

6. Conclusion

The distributional properties of the forecast errors play a crucial role in calculating reliable prediction intervals. This paper has considered the problem of testing both for normality and asymmetry of the marginal law of an international panel of survey-based macroeconomic forecast errors using the distance test statistics with the critical values obtained via the sieve bootstrap. Our results indicate that the assumption of symmetry of the marginal distribution of forecast errors is reasonable whereas the assumption of normality is not. Therefore, using the Student distribution might be preferred against the Gaussian one for the purpose of fan-chart modelling.

REFERENCES

Aki, S. (1981). symptotic distribution of a Cramer-von Mises type statistic for testing symmetry when the center is estimated. *Annals of the Institute of Statistical Mathematics 33*, 1–14.

Anderson, T. and D. Darling (1954). A test of goodness of fit. *Journal of the American Statistical Association 49*, 765–769.

Bai, J. and S. Ng (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business and Economic Statistics 23*, 49–60.

Bickel, P. J. and P. Bühlmann (1997). Closure of linear processes. *Journal of Theoretical Probability 10*, 445–479.

Boos, D. D. (1982). A test for asymmetry associated with the Hodges-Lehmann estimator. *Journal of the American Statistical Association 77*, 647–651.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli 3*, 123–148.

Davison, A. and D. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Greenspan, A. (2003). Monetary Policy under Uncertainty - Remarks by Chairman Alan Greenspan at a symposium of the Federal Reserve Bank of Kansas City. *The Federal Reserve Board, Washington, D.C*.

Hammond, G. et al. (2012). State of the art of inflation targeting. *Centre for Central Banking Studies, Bank of England*.

Harvey, D. I. and P. Newbold (2003). The non-normality of some macroeconomic forecast errors. *International Journal of Forecasting 19*, 635–653.

Koziol, J. A. (1983). Tests for symmetry about an unknown value based on the empirical distribution function. *Communications in Statistics - Theory and Methods 12*, 2823–2846.

Kreiss, J.-P. (1992). Bootstrap procedures for AR(∞) processes. In K.-H. Jöckel, G. Rothe, and W. Sendler (Eds.), *Bootstrapping and Related Techniques*, pp. 107–113. Heidelberg: Springer-Verlag.

Lahiri, K. and C. Teigland (1987). On the normality of probability distributions of inflation and GNP forecasts. *International Journal of Forecasting 3*, 269–279.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. Springer.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association 62*, 399–402.

Makridakis, S. and R. L. Winkler (1989). Sampling distributions of post-sample forecasting errors. *Applied Statistics 38*, 331–342.

Paulsen, J. and D. Tjøstheim (1985). On the estimation of residual variance and order in autoregressive time series. *Journal of the Royal Statistical Society, B 47*, 216–228.

Poskitt, D. S. (2007). Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases. *Annals of the Institute of Statistical Mathematics 59*, 697–725.

Poskitt, D. S. (2008). Properties of the sieve bootstrap for fractionally integrated and non-invertible processes. *Journal of Time Series Analysis 29*, 224–250.

Psaradakis, Z. (2003). A bootstrap test for symmetry of dependent data based on a Kolmogorov–Smirnov type statistic. *Communications in Statistics - Simulation and Computation 32*, 113–126.

Psaradakis, Z. (2016). Using the bootstrap to test for symmetry under unknown dependence. *Journal of Business and Economic Statistics 34*, 406–415.

Psaradakis, Z. and M. Vávra (2017). A distance test of normality for a wide class of stationary processes. *Econometrics and Statistics 2*, 50–60.

Ramberg, J. and B. Schmeiser (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM 17*, 78–82.

Reifschneider, D. and P. Tulip (2007). Gauging the uncertainty of the economic outlook from historical forecasting errors. *Finance and Economics Discussion Series 2007-60*.

Rothman, E. and M. Woodroofe (1972). A Cramér von-Mises type statistic for testing symmetry. *The Annals of Mathematical Statistics 43*, 2035–2038.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics 8*, 147–164.

Stephens, M. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *The Annals of Statistics 4*, 357–369.

Tetlow, R. and P. Tulip (2008). Changes in macroeconomic uncertainty. *Board of Governors Of the Federal Reserve System*.

Tjøstheim, D. and J. Paulsen (1983). Bias of some commonly-used time series estimates. *Biometrika 70*, 389–399; Corrigendum (1984), 71, 656.

A. TABLES

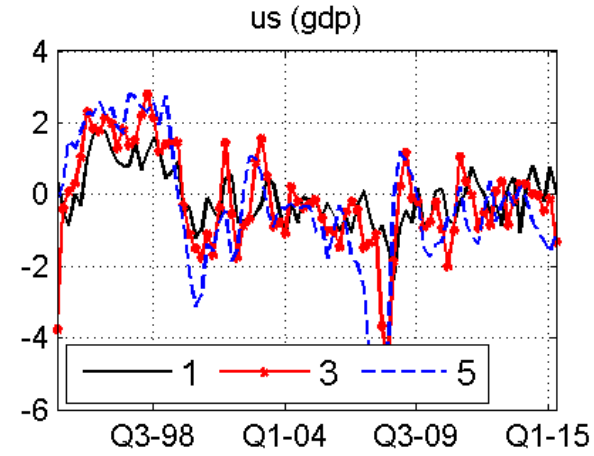
Table 1: P-values of Normality Test

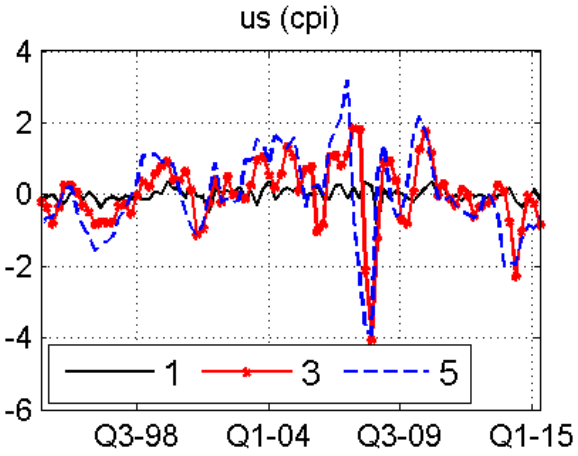
		Horizon					
Variable	Country	1	2	3	4	5	6
GDP	US	0.18	0.04	0.09	0.21	0.18	0.18
	UK	0.47	0.07	0.01	0.00	0.00	0.00
	JP	0.38	0.38	0.20	0.07	0.06	0.04
	DE	0.33	0.04	0.03	0.05	0.04	0.06
	FR	0.21	0.06	0.07	0.10	0.10	0.11
	IT	0.23	0.07	0.07	0.11	0.10	0.12
	CA	0.76	0.56	0.31	0.21	0.29	0.20
CPI	US	0.17	0.00	0.17	0.25	0.20	0.32
	UK	0.00	0.13	0.06	0.01	0.01	0.03
	JP	0.54	0.67	0.54	0.94	0.57	0.62
	DE	0.00	0.93	0.34	0.69	0.79	0.72
	FR	0.00	0.64	0.90	0.83	0.62	0.43
	IT	0.00	0.64	0.91	0.83	0.61	0.42
	CA	0.04	0.87	0.86	0.39	0.09	0.31

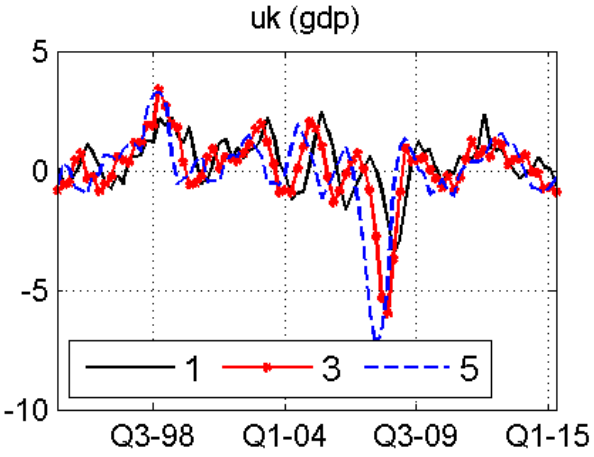
Table 2: P-values of Symmetry Test

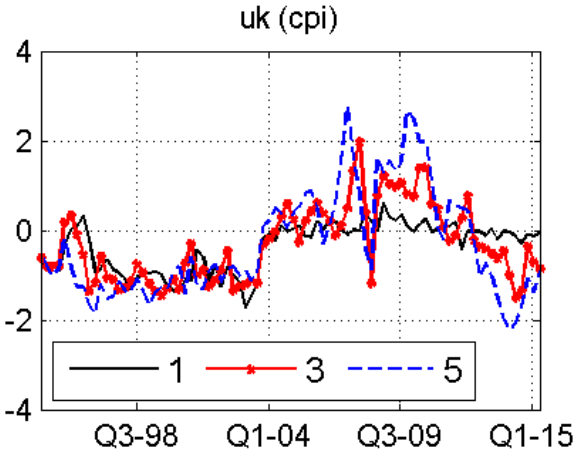
		Horizon					
Variable	Country	1	2	3	4	5	6
GDP	US	0.12	0.12	0.11	0.16	0.37	0.32
	UK	0.46	0.44	0.21	0.16	0.06	0.01
	JP	0.94	0.34	0.48	0.46	0.42	0.37
	DE	0.24	0.82	0.67	0.56	0.68	0.73
	FR	0.63	0.27	0.21	0.14	0.17	0.15
	IT	0.63	0.28	0.20	0.14	0.16	0.14
	CA	0.76	0.56	0.31	0.21	0.29	0.20
CPI	US	0.49	0.20	0.56	0.56	0.32	0.27
	UK	0.04	0.71	0.22	0.08	0.10	0.14
	JP	0.50	0.79	0.60	0.92	0.52	0.69
	DE	0.00	0.84	0.21	0.44	0.61	0.57
	FR	0.24	0.40	0.87	0.77	0.65	0.51
	IT	0.23	0.40	0.86	0.77	0.65	0.51
	CA	0.23	0.76	0.77	0.34	0.22	0.35

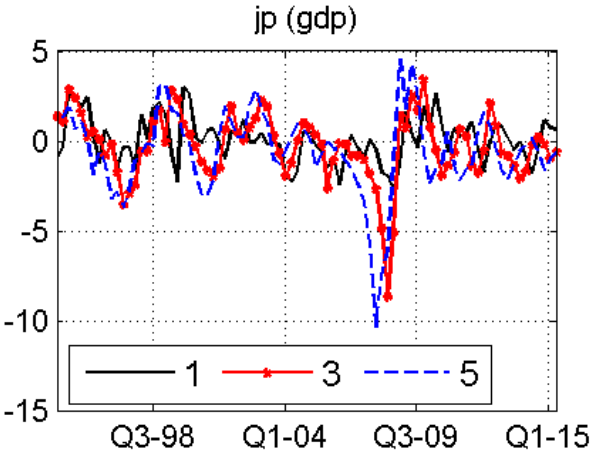
c Aggregate CE Forecast Errors

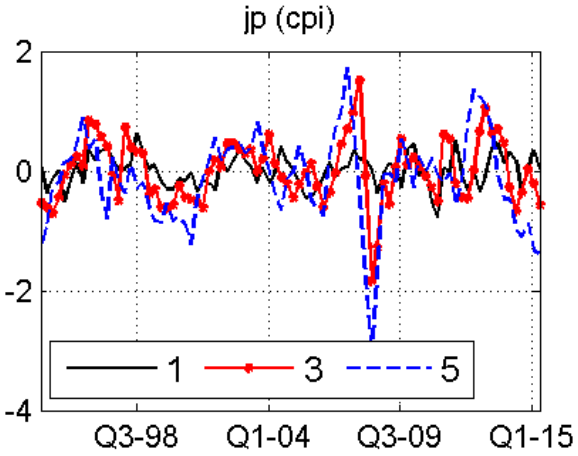


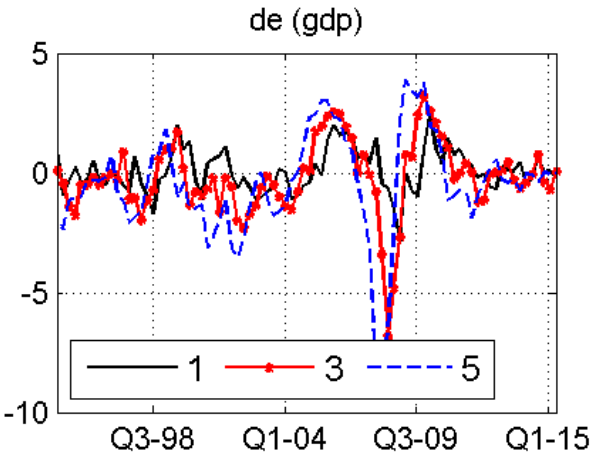


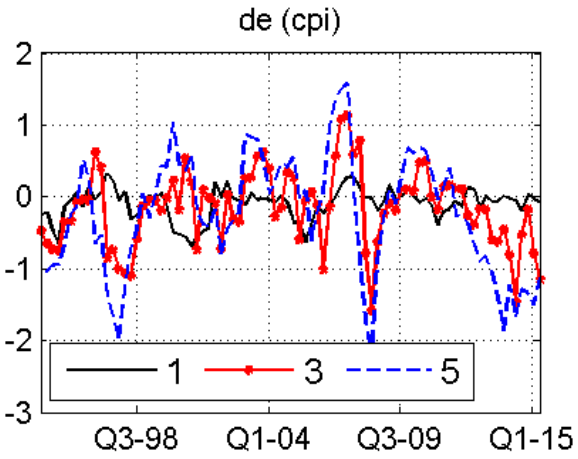


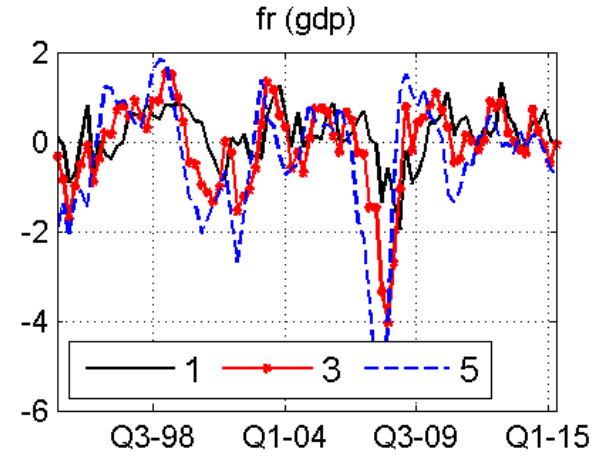


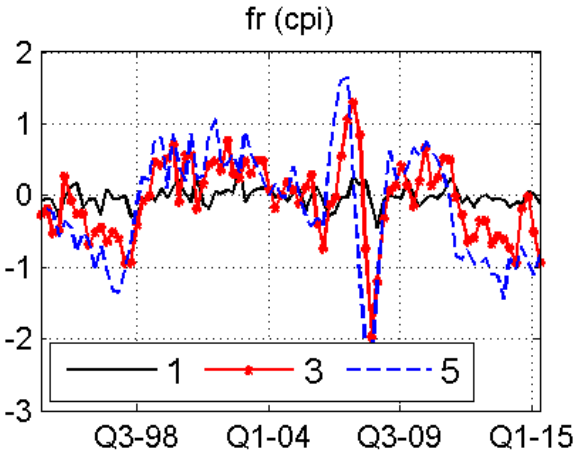


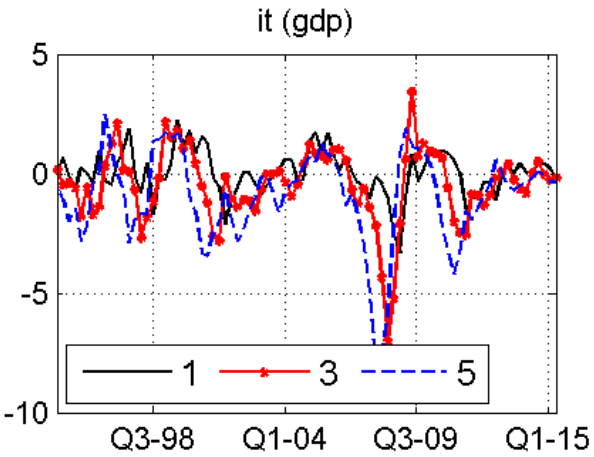


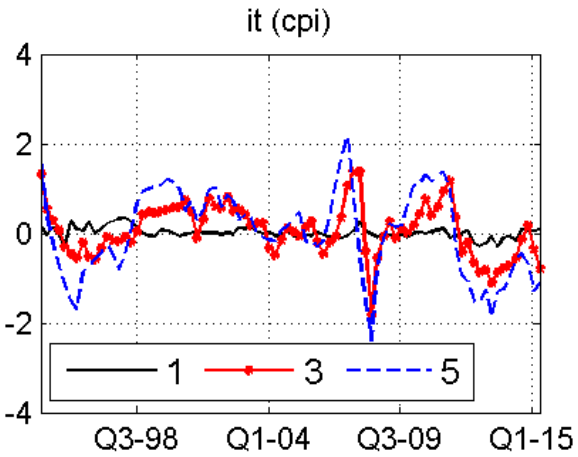

de (gdp)
5
0
-5
-10
1
3
5
Q3-98
Q1-04
Q3-09
Q1-15


de (cpi)
2
1
0
-1
-2
-3
1
3
5
Q3-98
Q1-04
Q3-09
Q1-15


fr (gdp)
2
0
-2
-4
-6
1
3
5
Q3-98
Q1-04
Q3-09
Q1-15


fr (cpi)
2
1
0
-1
-2
-3
1
3
5
Q3-98
Q1-04
Q3-09
Q1-15


it (gdp)
5
0
-5
-10
1
3
5
Q3-98
Q1-04
Q3-09
Q1-15


it (cpi)
4
2
0
-2
-4
1
3
5
Q3-98
Q1-04
Q3-09
Q1-15

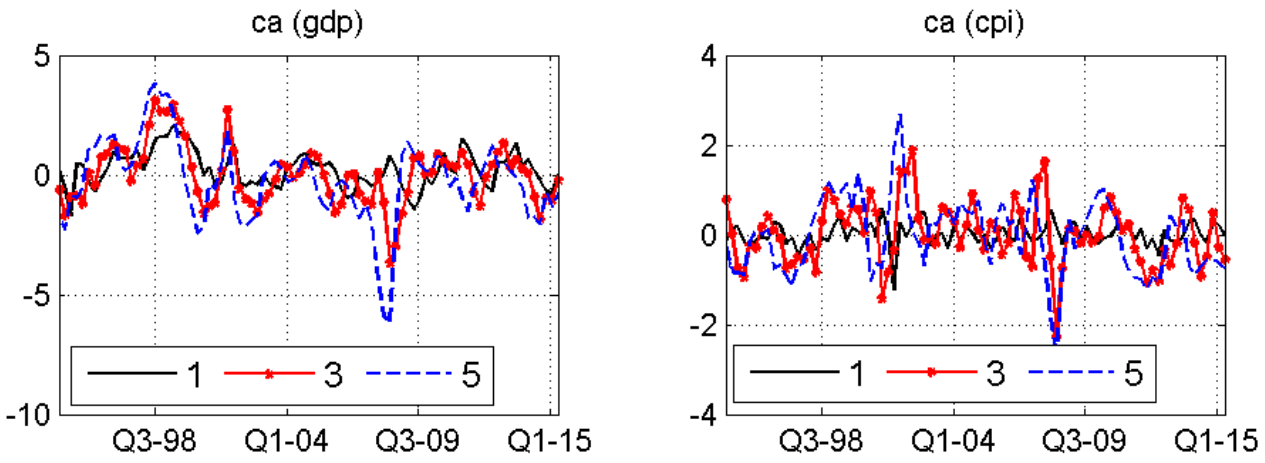
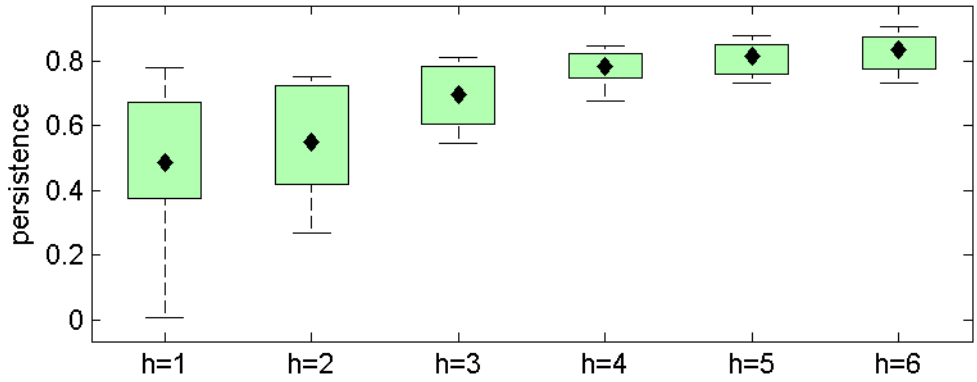


Figure 2: Persistence of Aggregate CE Forecast Errors



Note: The black diamond (placed inside each box) indicates the average value of the first-order autocorrelation coefficient estimated from the forecast errors under consideration, the top and bottom of each box indicate the 25th and 75th percentile, respectively, of the empirical distribution of the first-order autocorrelation coefficients, and the whiskers indicate the 10th and 90th percentiles of the empirical distribution of the first-order autocorrelation coefficients.

B. Simulation Study

In this section, we present and discuss the results of a simulation study examining the small-sample properties of the distance-based tests of normality and symmetry under different patterns of dependence by considering artificial data generated according to the ARMA models

M1: $X_t = 0.8X_{t-1} + \varepsilon_t$,

M2: $X_t = 0.6X_{t-1} - 0.5X_{t-2} + \varepsilon_t$,

M3: $X_t = 0.6X_{t-1} + 0.3\varepsilon_{t-1} + \varepsilon_t$.

Here, and throughout this section, $\{\varepsilon_t\}$ are i.i.d. random variables the common distribution of which is either standard normal (labelled N in the various tables) or generalized lambda with quantile function $Q(w) = \lambda_1 + (1/\lambda_2)\{w^{\lambda_3} - (1-w)^{\lambda_4}\}$, $0 < w < 1$, standardized to have zero mean and unit variance (see Ramberg and Schmeiser (1974)). The parameter values of the generalized lambda distribution used in the experiments are taken from Bai and Ng (2005) and can be found in Table 3, along with the corresponding coefficients of skewness and kurtosis; the distributions N, S1, S2 are symmetric, whereas A1, A2, A3 are asymmetric.

For each design point, 1000 independent realizations of $\{X_t\}$ of length $100 + n$, with $n \in \{100, 200\}$ (as representative samples for macroeconomic applications), are generated.[15] The first 100 data points of each realization are then discarded in order to eliminate start-up effects and the remaining n data points are used to compute the value of the $\mathcal{D}_N$ and $\mathcal{D}_S$ test statistics. In the case of bootstrap tests, the order of the autoregressive sieve is determined by minimizing the AIC in the range $1 \leq p \leq \lfloor 5 \log_{10} n \rfloor$, while the number of bootstrap replications is $B = 199$. We note that using a larger number of bootstrap replications did not change the results substantially.[16]

Table 3: Innovation Distributions

	λ_1	λ_2	λ_3	λ_4	skewness	kurtosis
N	–	–	–	–	0.0	3.0
S1	0.000000	-0.397912	-0.160000	-0.160000	0.0	11.6
S2	0.000000	-1.000000	-0.240000	-0.240000	0.0	126.0
A1	0.000000	-1.000000	-0.007500	-0.030000	1.5	7.5
A2	0.000000	-1.000000	-0.100900	-0.180200	2.0	21.1
A3	0.000000	-1.000000	-0.001000	-0.130000	3.2	23.8

The Monte Carlo rejection frequencies of the distance tests at 10% significance level are reported in Tables 4 – 5. The null rejection probabilities of the tests are generally insignificantly different from the nominal level across all relevant DGPs. Their rejection frequencies improve

[15] The Monte Carlo results for different sample sizes are available from the author upon request.

[16] Using a larger number of bootstrap replications did not change the results significantly (see Davison and Hinkley (1997, pp. 155-156) for an explanation).

both with the sample size and asymmetry in the distribution of innovations, although not uniformly (compare the results for A1 and A2).

Table 4: Rejection Frequencies of Distance Tests: $n = 100$, AIC

	$\mathcal{D}_N$			$\mathcal{D}_S$		
Distribution	M1	M2	M3	M1	M2	M3
N	0.09	0.08	0.10	0.10	0.09	0.10
S1	0.28	0.38	0.30	0.12	0.09	0.09
S2	0.36	0.57	0.46	0.11	0.10	0.13
A1	0.36	0.68	0.53	0.24	0.56	0.39
A2	0.37	0.62	0.50	0.19	0.32	0.26
A3	0.66	0.97	0.91	0.35	0.92	0.74

Table 5: Rejection Frequencies of Distance Tests: $n = 200$, AIC

	$\mathcal{D}_N$			$\mathcal{D}_S$		
Distribution	M1	M2	M3	M1	M2	M3
N	0.11	0.09	0.09	0.10	0.09	0.08
S1	0.31	0.62	0.45	0.11	0.11	0.10
S2	0.46	0.82	0.63	0.12	0.09	0.10
A1	0.53	0.92	0.80	0.40	0.89	0.70
A2	0.53	0.85	0.73	0.28	0.57	0.46
A3	0.90	1.00	1.00	0.70	1.00	0.97