Empirical Economics (2020) 59:2841–2858
https://doi.org/10.1007/s00181-019-01726-0

Assessing distributional properties of forecast errors for fan-chart modelling

Marián Vávra[1,2]

Received: 10 December 2018 / Accepted: 1 July 2019 / Published online: 4 July 2019

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. Our results indicate that the assumption of symmetry of the marginal distribution of forecast errors is reasonable, whereas the assumption of normality is not, making symmetric prediction intervals clearly preferable.

Keywords Normality · Symmetry · Forecast errors · Prediction interval · Fan-chart · Sieve bootstrap

JEL Classification C12 · C15 · C22 · C53

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; Haldane and Nelson 2012). 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.[1] Gaussian-like prediction bands have become a workhorse

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

The author is grateful to Zacharias Psaradakis, Ron Smith, Peter Tulip, and three referees for helpful comments and suggestions in improving an earlier version of the paper.

✉ Marián Vávra
marian.vavra@gmail.com

[1] Faculty of Social and Economic Sciences, Institute of Economics, Comenius University in Bratislava, Mlynské luhy 4, 821 05 Bratislava, Slovakia

[2] Research Department, National Bank of Slovakia, Imricha Karvaša 1, 813 25 Bratislava, Slovakia

in fan-chart modelling (see, e.g., Bank of Canada, Sveriges Riksbank, Norges Bank, Czech National Bank, European Central Bank, just to name a few). An alternative approach, gaining increasing popularity in recent years, is to calculate prediction bands that assume symmetry but not the stronger assumption of normality (see, e.g., Reserve Bank of Australia, National Bank of Slovakia).

The accuracy of both types of prediction bands critically depends on the validity of the underlying distributional properties of forecast errors. In the former case, prediction bands explicitly rely on an assumption that forecast errors are normally distributed, whereas in the latter case on an assumption of symmetrically distributed errors. Clearly, if the distributional assumption is violated, then the intervals are subject to systematic 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).

Unfortunately, many practitioners are reluctant to test for the distributional assumptions when calculating prediction intervals. We suspect 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.[2] For example, Reifschneider and Tulip (2007) assess normality of the US Federal Reserve 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. Using the original skewness–kurtosis test might be justified only for serially uncorrelated forecast errors but not in general.[3] 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.

The main contribution of the paper is to provide reliable empirical evidence about the distributional properties of key macroeconomic forecast errors which can then be used for fan-chart modelling. We do so by assessing both normality and symmetry

[2] The only exception, the author is aware of, is Reifschneider and Tulip (2019) where the appropriate Monte Carlo critical values are used when testing for normality of the US forecast errors.

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

of an international panel of survey-based macroeconomic forecast errors using the test statistics based on empirical (robust) standardized cumulants with appropriate critical values obtained via a sieve bootstrap. 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.[4]

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

2 Assumptions and 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 strictly stationary and ergodic sequence of real-valued random variables with a finite fourth absolute moment such that $\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ and $\mathbb{E}(\varepsilon_t^2|\mathcal{F}_{t-1}) = s^2 > 0$ for all t, $\mathcal{F}_{t-1}$ being the sigma-algebra generated by $\{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$.

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 is non-Gaussian.[5] Testing whether a sample of observations comes from a Gaussian distribution is a problem that has attracted much attention over the years. Although many different test statistics have been developed for this purpose (see Thode 2002 for a review), we focus on a test for normality proposed by Bowman and Shenton (1975) and Jarque and Bera (1980, 1987) based on the empirical standardized third and fourth cumulants, exploiting the fact that for a normal distribution all cumulants of order higher than the second are zero. This classical skewness–kurtosis statistic has become arguably the most popular test for normality in the literature and available in many statistical packages (e.g., EViews or MATLAB). The test statistic is given by

[4] A MATLAB code is available to researchers upon request from the author.

[5] 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.

$$\mathcal{T}_N = \frac{n}{6}\left(\frac{1}{n}\sum_{t=1}^{n}\hat{Z}_t^3\right)^2 + \frac{n}{24}\left(\frac{1}{n}\sum_{t=1}^{n}\hat{Z}_t^4 - 3\right)^2, \tag{3}$$

where $n^{-1}\sum_{t=1}^{n}\hat{Z}_t^3$ is the coefficient of skewness, $n^{-1}\sum_{t=1}^{n}\hat{Z}_t^4$ is the coefficient of kurtosis, and $\hat{Z}_t = (X_t - \bar{X})/\hat{\sigma}$ with $\bar{X}$ and $\hat{\sigma}$ denoting sample mean and sample standard deviation calculated from $\mathcal{X}_n$. It can be shown that for Gaussian independent and identically distributed observations, $\mathcal{T}_N$ is approximately χ^2 distributed as $n \to \infty$. However, when applied to weakly dependent data, the test suffers from a serious size distortion which increases with persistence, sample size, and complexity of stochastic processes (see Psaradakis and Vávra 2019, pp. 9–19).

Bai and Ng (2005) developed a related test statistic for weakly dependent data (based on replacing constants 6 and 24 in (3) by respective long-run asymptotic variances of the coefficients of skewness and kurtosis. These are constructed using a nonparametric kernel estimator with a data-driven bandwidth). However, the finite sample properties of the Bai–Ng test for normality are not satisfactory. The test suffers from a serious size distortion and a power loss which significantly increases with dependence in data (see Bai and Ng 2005, p. 57). Both distortions are of the magnitude making the test unreliable unless a large sample of observations is available.

There are two reasons for the poor performance of the Bai–Ng-type statistic. First, Bao (2013) showed that both sample skewness and kurtosis can be severely biased, which causes a non-negligible size distortion in finite samples when relying on the asymptotic distribution. The bias depends on higher-order cumulants as well as the dependency structure in data. Second, although consistency of nonparametric long-run variance estimators is well established in the literature (see Andrews 1991, Thm. 1), it is highly inaccurate for weakly dependent data in small samples (see Müller 2014).

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

$$\mathcal{H}_0^S : F(x - \mu) = 1 - F(\mu - x). \tag{4}$$

The alternative hypothesis is that distribution F is asymmetric.[6] However, testing for marginal symmetry of weakly dependent data is even more peculiar and challenging than testing for normality. In line with (3), a natural choice would be to implement a test based on the coefficient of skewness $n^{-1}\sum_{t=1}^{n}\hat{Z}_t^3$. However, it is nowadays well understood that the classical moment-based measure of skewness is adversely affected by leptokurtosis and outliers (see, e.g., Horsewell and Looney 1993; Rayner et al. 1995; Kim and White 2004). Therefore, robust measures of symmetry are particularly useful when the underlying distribution is expected to be heavy-tailed or there are extreme observations in the sample, a characteristic feature of many economic datasets (see Balke and Fomby 1994 for empirical evidence). For this reason, we focus on a test statistic proposed by Premaratne and Bera (2005) based on the Pearson type IV family of distributions taking asymmetry and excess kurtosis into account explicitly.

[6] 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 4 for an explanation.

This family is fairly large and includes, for example, normal and (skew-) Student distributions, among others. The test statistic is given by

$$\mathcal{T}_S = n\left(\frac{1}{n}\sum_{t=1}^{n}\tan^{-1}(\hat{Z}_t)\right)^2, \tag{5}$$

where again $\hat{Z}_t = (X_t - \bar{X})/\hat{\sigma}$ with $\bar{X}$ and $\hat{\sigma}$ denoting a sample mean and sample standard deviation calculated from $\mathcal{X}_n$. The $\tan^{-1}(\cdot)$ function is odd, continuous, and bounded (on the real line) which makes the test statistic robust to heavy-tailed observations. The authors show that for independent and identically distributed observations, the (rescaled) version of the $\mathcal{T}_S$ test statistic is asymptotically χ^2 distributed as $n \to \infty$. Although the limiting distribution of the (rescaled) statistic $\mathcal{T}_S$ can be derived for weakly dependent data (see Chen and Lin 2008), the approximation might be expected to perform poorly in small samples due to the inaccuracy of the long-run variance estimators (see Müller 2014).

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 moment-based test statistics. The principal advantage of the sieve bootstrap is that it can be used to approximate the sampling properties of $\mathcal{T}_N$ and $\mathcal{T}_S$ 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, $\mathcal{X} = \{X_t\}_{t\in\mathbb{Z}}$ 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 $|z| < 1$.[7] 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; Bühlmann 1997).

The bootstrap procedure used to approximate the sampling properties of the test statistics $\mathcal{T}_N$ and $\mathcal{T}_S$ under the null hypotheses can be described in the following steps.

[7] 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 disk $|z| < 1$.

(Only for notational simplicity, the test statistic is denoted as $\mathcal{T}$. Any computational differences between $\mathcal{T}_N$ and $\mathcal{T}_S$ are stated explicitly).

Step 1 Select an appropriate lag order p of an AR model using the Akaike information criterion (AIC), where the lag order is restricted by $0 \leq p < 5\log_{10}(n)$, where n denotes the sample size. Note that other lag order selection criteria can be used, but since the process $\{X_t\}$ in (6) is not assumed to be of finite dimension, the AIC is asymptotically efficient (see Shibata 1980).

Step 2 Obtain the unknown AR(p) model parameters $(\hat{\phi}_1, \ldots, \hat{\phi}_p)$ by the ordinary least squares (OLS) method for the mean corrected series $\{X_t - \bar{X}\}_{t=1}^n$, where $\bar{X} = n^{-1}\sum_{t=1}^n X_t$ is a sample average. In contrast to Bühlmann (1997), who implemented the Yule–Walker (YW) estimator, we rely on the standard OLS estimator. The main reason for doing so is that the OLS estimator produces superior results as compared to the YW estimator (see Tjøstheim and Paulsen 1983; Paulsen and Tjøstheim 1985).

Step 3 Construct a sequence of the estimated residuals $\{\hat{\varepsilon}_t\}_{t=p+1}^n$ by the recursion

$$\hat{\varepsilon}_t = X_t - \bar{X} - \sum_{j=1}^{p} \hat{\phi}_j (X_{t-j} - \bar{X}), \qquad t = p+1, 2, \ldots, n.$$

Step 4 By setting initial values $X^*_{-p+1} = \cdots = X^*_0 = \bar{X}$, generate bootstrap pseudo-observations $(X^*_1, \ldots, X^*_{n+b})$ via the recursion (b is some positive integer)

$$X^*_t - \bar{X} = \sum_{j=1}^{p} \hat{\phi}_j (X^*_{t-j} - \bar{X}) + a^*_t, \qquad t = 1, 2, \ldots, n+b, \tag{7}$$

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 $\hat{s}^2 = (n - 2p - 1)^{-1}\sum_{t=p+1}^n \hat{\varepsilon}_t^2$;
- 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 3 and $\{\zeta_t\}$ is a sequence of i.i.d. random variables drawn from the discrete uniform distribution on -1 and 1. Note that multiplying the estimated residuals by the uniform random variable ζ in this form ensures that the marginal distribution of model innovations is symmetric (see Berg et al. 2010).

Then discard the initial b replicates to eliminate start-up effects (see Swanepoel and Van Wyk 1986). Define the bootstrap analogue of $\mathcal{T}$ by the plug-in rule as $\mathcal{T}^*$ calculated using the appropriate test statistic with $\mathcal{X}_n^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$ replacing $\mathcal{X}_n$.

It is worth noting that, by requiring a_t^* in (7) 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).

Step 5 Repeat Step 4 independently B times to obtain a collection of B replicates $\{\mathcal{T}_1^*, \ldots, \mathcal{T}_B^*\}$ of $\mathcal{T}^*$. The sampling distribution of $\mathcal{T}$ is then approximated by the empirical distribution function associated with $\{\mathcal{T}_1^*, \ldots, \mathcal{T}_B^*\}$, that is $\hat{H}^*(u) = B^{-1} \sum_{i=1}^{B} \mathbb{I}(\mathcal{T}_i^* \leq u)$, for $u \in \mathbb{R}$. Then, a bootstrap test rejects the null hypothesis at the significance level α if $\mathcal{T} > \inf\{q : \hat{H}^*(q) \geq 1 - \alpha\}$, where $\mathcal{T}$ 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 $\mathcal{T}$ 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 $\mathcal{T}$ and the conditional distribution function H^* of $\mathcal{T}^*$ given $\mathcal{X}_n$ (where $g^{-1}(u) = \inf\{x : g(x) \geq u\}$ for any non-decreasing function g).[8] 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 $n \to \infty$.

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 nonlinear 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 Macroeconomic 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

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

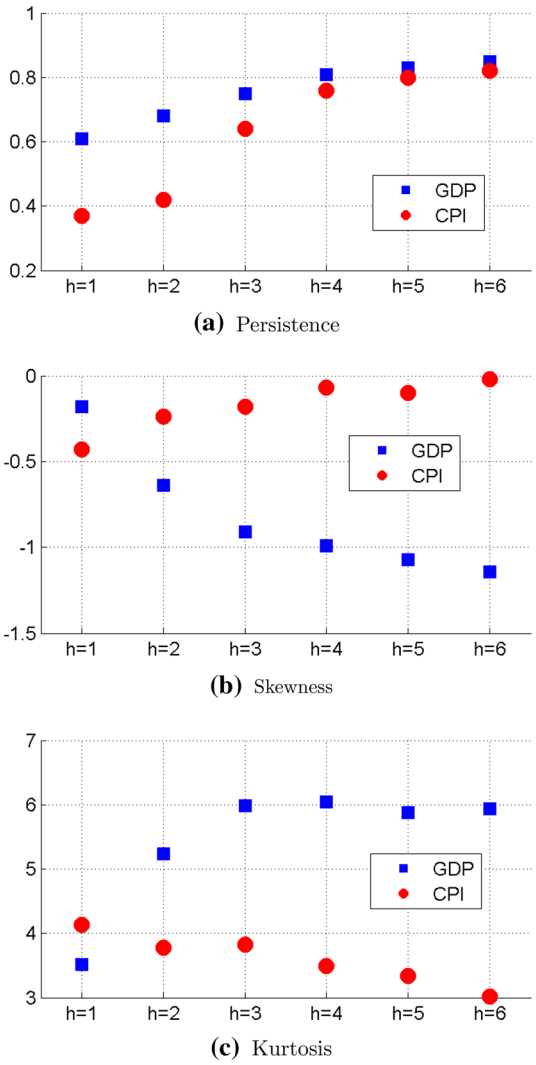

Fig. 1 Horizon-dependent persistence, skewness, and kurtosis of aggregate CE forecast errors (final data)

85 countries. The CE output is often seen as a forecast benchmark by investment and planning managers, as well as government and public sector institutions. In addition to their annual (fixed-event) forecasts, the company provides 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.[9]

[9] 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 comparable dataset of central banks' forecast errors with the one employed in the study.

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 (the USA, Japan, Germany, France, the UK, Italy, Canada). Forecasts of both economic variables are reported in the form of year-on-year percentage changes.[10] The forecast horizon of variables under consideration is from one quarter to six quarters ahead.

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 quarters ahead. (The index h is sometimes omitted for the notational simplicity.) Since macroeconomic variables are subject to (sometimes substantial) revisions, it is far from clear which version of the realization should be actually used for our analysis. On the one hand, final data can be preferred since they represent the most accurate data available to researchers. On the other hand, final data are available with a considerable lag taking even several years and thus first-release data might be more relevant for real-time analysis (see Aruoba 2008 for details about statistical revisions). For these reasons, we examine the distributional properties of forecasts errors calculated using both first-release (unrevised) and final (fully revised) macroeconomic data. The dataset of international forecast errors is a balanced panel spanning the period Q4 1994–Q3 2015 in the case of final data (Q2 1999–Q3 2015 in the case of first-release data). Revised actuals for all macroeconomic series are obtained from the Real-Time OECD Database (visit https://stats.oecd.org).

The aggregate CE forecast errors of both macroeconomic variables for the selected horizons $h = 1, 3, 5$ are depicted in Fig. 2 in Appendix A. Some comments about the stochastic properties of the errors are in order. A characteristic feature of both GDP and CPI errors is their high persistence which increases with the forecast horizon h—see Fig. 1a where averages of the first-order autocorrelation coefficients of forecast errors calculated over individual countries are depicted. The figure clearly demonstrates that using test statistics based on the assumption of serially uncorrelated observations (as in studies cited earlier) would provide very likely misleading inference about the distributional properties of forecast errors. The interested reader is referred to Psaradakis and Vávra (2019, pp. 9–19) for Monte Carlo evidence. As for higher moments, relevant for testing normality, GDP forecast errors clearly exhibit higher degree of non-normality in terms of both sample skewness and kurtosis—see Fig. 1b, c.

[10] 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.

5 Empirical results

In this section, we assess normality and symmetry of the marginal law of survey-based macroeconomic forecast errors using the above-described $\mathcal{T}_N$ and $\mathcal{T}_S$ test statistics. It is shown in Appendix B that the moment-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. Recall that, consistently with a notation introduced in Sect. 2, $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ denotes a sample of the aggregate CE forecast errors for each variable (CPI and GDP), each forecast horizon ($h = 1, \ldots, 6$), and each data release (final data and first-release data). A sample mean and standard deviation calculated from the sample $\mathcal{X}_n$ are used as estimators of location μ and scale σ. (Both estimators are perfectly justifiable under the null hypothesis of normality.) The bootstrap P-values of the distance tests are reported in Tables 1, 2, 3, and 4. These are computed from 1,000 bootstrap replications with the data-dependent sieve order p determined using the AIC over $1 \leq p < 5 \log_{10}(n)$.

Normality The null hypothesis of normality is rejected (at the conventional 5% significance level) in around 75% (25%) of the GDP (CPI) forecast errors using the final data. A markedly higher rejection rate of GDP errors can be explained by higher degree of non-normality in terms of sample skewness and kurtosis as compared to CPI (see Fig. 1). Two interesting conclusions emerge when focusing on normality over the forecast horizon h and individual countries. First, the null of normality of GDP errors is rejected for all countries. What is more, in five out of seven countries (i.e., UK, JP, DE, FR, and IT), the null is rejected for at least 5 out of 6 forecast horizons. Second, normality of CPI errors is rejected mainly for two countries only (US and JP). It is worth remarking here that no significant differences between rejection rates for the final and first-release data are observed.

Symmetry Noticeably different results are obtained when testing for the null of symmetry of the marginal distribution of forecast errors. The null of symmetry is rejected (again at the 5% significance level) only in 15% (10%) of the GDP (CPI) forecast errors series using the final data. Interestingly, in most of the cases, the null is rejected for the UK and FR forecast errors only. As in the previous case of normality, no significant differences between rejection rates for the final and first-release data are observed. Putting the normality and symmetry results together, we may conclude that when the null normality is rejected, it is mainly due to the presence of excess kurtosis in the forecast errors.

Implications Our results indicate that the assumption of symmetry of the marginal distribution of forecast errors is reasonable, whereas the assumption of normality is not in general, making symmetric prediction intervals clearly preferable. Although the prediction bands that assume symmetry of the marginal distribution of errors might be calculated in different ways, using sample quantiles of the estimated marginal

distribution of (symmetrized) forecast errors seems to be computationally the most attractive approach (see Tulip and Wallace 2012 for details).[11] Unlike the traditional Gaussian approach (based on the root-mean-squared errors calculated from historical errors and the standard normal critical values), the quantile approach is more general (and also nests Gaussian prediction bands as a special case) and consistent for weakly dependent forecast errors under mild regularity conditions (see Sen 1968; Lee 1998; Psaradakis and Vávra 2015 for the asymptotic properties of sample quantiles). Nevertheless, it is important to point out that the behavior of sample quantiles lying in the tails of the error distribution might be erratic due to a limited number of representative observations. In such cases, quantile estimates might be improved using a bootstrap method (see, e.g., Sharipov and Wendler 2013).

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 for both normality and asymmetry of survey-based macroeconomic forecast errors using the distance-based 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 in general, making symmetric prediction intervals clearly preferable.

A Figures and Tables

See Fig. 2 and Tables 1, 2, 3, and 4.

[11] An alternative way could be to use a Student t distribution with the estimated degrees of freedom which are very likely to be horizon/variable dependent as implied from Fig. 1c.

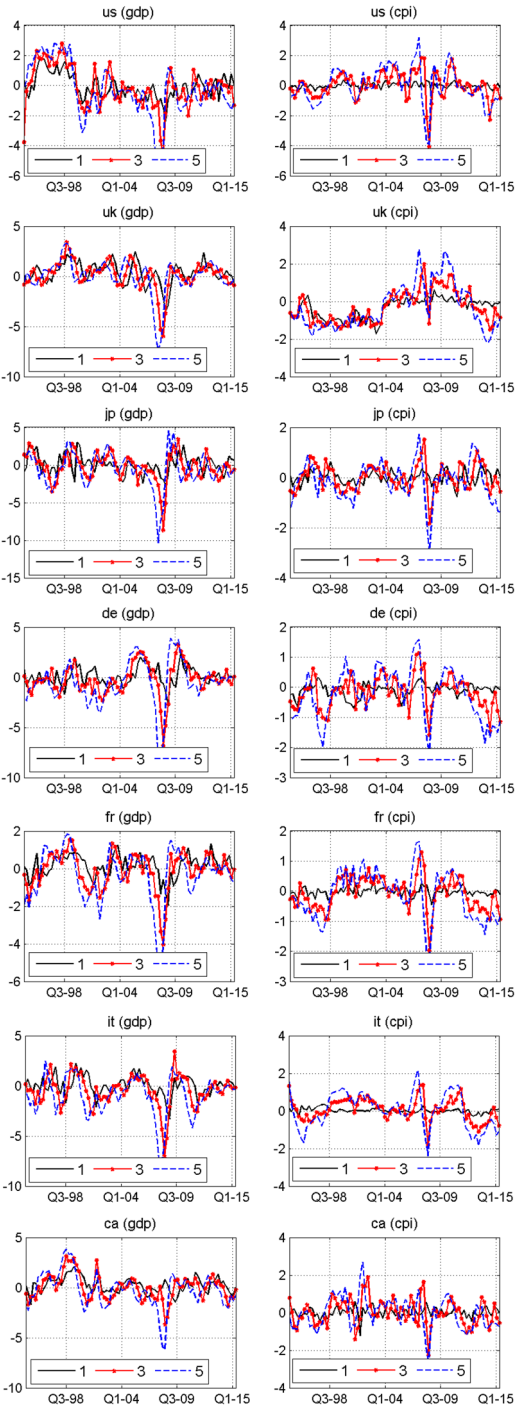

Fig. 2 Aggregate CE forecast errors (final data)

Table 1 P-values of normality test $\mathcal{T}_N$: GDP

Release	Country	Horizon					
		1	2	3	4	5	6
Final	US	0.59	0.04	0.09	0.53	0.18	0.06
	UK	0.01	0.00	0.00	0.00	0.00	0.00
	JP	0.79	0.00	0.00	0.00	0.00	0.00
	DE	0.58	0.00	0.00	0.00	0.00	0.00
	FR	0.01	0.00	0.00	0.00	0.00	0.00
	IT	0.02	0.00	0.00	0.00	0.00	0.00
	CA	0.78	0.28	0.58	0.04	0.01	0.01
First	US	0.50	0.05	0.00	0.00	0.00	0.00
	UK	0.09	0.02	0.00	0.00	0.00	0.00
	JP	0.04	0.00	0.00	0.00	0.00	0.00
	DE	0.00	0.00	0.00	0.00	0.00	0.00
	FR	0.68	0.01	0.00	0.00	0.00	0.00
	IT	0.00	0.00	0.00	0.00	0.00	0.00
	CA	0.06	0.17	0.02	0.01	0.00	0.00

Table 2 P-values of normality test $\mathcal{T}_N$: CPI

Release	Country	Horizon					
		1	2	3	4	5	6
Final	US	0.00	0.00	0.00	0.01	0.07	0.49
	UK	0.07	0.17	0.10	0.06	0.05	0.06
	JP	0.82	0.77	0.15	0.04	0.01	0.02
	DE	0.04	0.39	0.87	0.91	0.75	0.71
	FR	0.31	0.70	0.21	0.59	0.81	0.70
	IT	0.10	0.29	0.74	0.74	0.59	0.40
	CA	0.00	0.64	0.49	0.29	0.16	0.65
First	US	0.00	0.01	0.00	0.00	0.01	0.06
	UK	0.08	0.71	0.45	0.52	0.72	0.87
	JP	0.57	0.80	0.08	0.05	0.01	0.05
	DE	0.09	0.77	0.50	0.59	0.51	0.52
	FR	0.13	0.13	0.02	0.06	0.22	0.43
	IT	0.02	0.32	0.35	0.40	0.41	0.41
	CA	0.00	0.73	0.41	0.39	0.18	0.67

B Simulation study

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

M1: $X_t = 0.5X_{t-1} + \varepsilon_t$,

Table 3 P-values of symmetry test $\mathcal{T}_S$: GDP

Release	Country	Horizon					
		1	2	3	4	5	6
Final	US	0.10	0.33	0.84	0.73	0.89	0.78
	UK	0.27	0.23	0.11	0.06	0.03	0.01
	JP	0.66	0.27	0.26	0.31	0.28	0.24
	DE	0.25	1.00	0.75	0.69	0.67	0.59
	FR	0.06	0.03	0.03	0.02	0.04	0.06
	IT	0.44	0.14	0.13	0.07	0.06	0.08
	CA	0.51	0.27	0.60	0.89	0.64	0.52
First	US	0.25	0.53	0.65	0.74	0.51	0.13
	UK	0.10	0.10	0.09	0.01	0.01	0.01
	JP	0.38	0.12	0.38	0.51	0.38	0.19
	DE	0.97	0.41	0.22	0.17	0.26	0.29
	FR	0.79	0.73	0.11	0.02	0.01	0.01
	IT	0.52	0.32	0.32	0.13	0.05	0.09
	CA	0.10	0.41	0.68	0.76	0.12	0.03

Table 4 P-values of symmetry test $\mathcal{T}_S$: CPI

Release	Country	Horizon					
		1	2	3	4	5	6
Final	US	0.60	0.21	0.44	0.80	0.96	0.83
	UK	0.04	0.92	0.15	0.06	0.04	0.05
	JP	0.57	0.73	0.79	0.86	0.38	0.46
	DE	0.00	0.78	0.48	0.60	0.51	0.71
	FR	0.85	0.89	0.61	0.54	0.56	0.57
	IT	0.38	0.41	0.93	0.71	0.59	0.51
	CA	0.90	0.61	0.63	0.85	0.93	0.44
First	US	0.22	0.17	0.15	0.27	0.35	0.51
	UK	0.11	0.73	0.70	0.51	0.41	0.49
	JP	0.79	0.86	0.95	0.83	0.38	0.37
	DE	0.81	0.25	0.48	0.60	0.45	0.57
	FR	0.57	0.35	0.16	0.11	0.10	0.17
	IT	0.63	0.38	0.42	0.28	0.25	0.32
	CA	0.68	0.95	0.90	0.85	0.68	0.75

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) 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

Table 5 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

Table 6 Empirical rejection frequencies of tests: AIC

Sample	Distribution	$\mathcal{T}_N$			$\mathcal{T}_S$		
		M1	M2	M3	M1	M2	M3
$n = 100$	N	0.05	0.04	0.05	0.05	0.04	0.05
	S1	0.62	0.52	0.38	0.05	0.05	0.06
	S2	0.75	0.65	0.53	0.06	0.06	0.07
	A1	0.82	0.68	0.51	0.71	0.55	0.33
	A2	0.76	0.68	0.53	0.34	0.26	0.22
	A3	1.00	0.97	0.89	0.97	0.89	0.71
$n = 200$	N	0.05	0.06	0.05	0.04	0.04	0.04
	S1	0.85	0.74	0.52	0.04	0.05	0.05
	S2	0.94	0.88	0.72	0.04	0.06	0.05
	A1	0.98	0.95	0.79	0.96	0.90	0.65
	A2	0.95	0.90	0.76	0.61	0.54	0.41
	A3	1.00	1.00	0.99	1.00	0.99	0.97

Bai and Ng (2005) and can be found in Table 5, 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 $n + 100$, with $n \in \{100, 200\}$ (as representative samples for macroeconomic applications), are generated.[12] 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 $\mathcal{T}_N$ and $\mathcal{T}_S$ test statistics. In the case of bootstrap tests, the order of the autoregressive sieve is determined by minimizing the AIC in the range $1 \leq p < 5 \log_{10}(n)$, while the number of bootstrap replications is $B = 499$. We note that using a larger number of bootstrap replications did not change the results substantially (see Davison and Hinkley 1997, pp. 155–156, for an explanation).

The Monte Carlo rejection frequencies of the test statistics at 5% significance level are reported in Table 6. The null rejection probabilities of the tests are generally insignificantly different from the nominal level across all relevant DGPs. Their

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

Table 7 Empirical rejection frequencies of tests: BIC

Sample	Distribution	$\mathcal{T}_N$			$\mathcal{T}_S$		
		M1	M2	M3	M1	M2	M3
$n = 100$	N	0.04	0.04	0.06	0.05	0.06	0.06
	S1	0.64	0.54	0.37	0.06	0.08	0.07
	S2	0.78	0.69	0.52	0.05	0.08	0.07
	A1	0.83	0.74	0.52	0.71	0.60	0.35
	A2	0.78	0.68	0.51	0.34	0.29	0.24
	A3	0.99	0.99	0.89	0.97	0.89	0.67
$n = 200$	N	0.06	0.06	0.05	0.06	0.06	0.05
	S1	0.84	0.78	0.54	0.05	0.07	0.06
	S2	0.95	0.90	0.74	0.05	0.07	0.06
	A1	0.99	0.96	0.79	0.97	0.90	0.63
	A2	0.94	0.91	0.75	0.61	0.54	0.36
	A3	1.00	1.00	1.00	1.00	1.00	0.96

rejection frequencies improve with both the sample size and non-normality in the distribution of innovations, although not uniformly (compare the results for A1 and A2). To assess the sensitivity of results with respect to the method used to determine the order of the autoregressive sieve, we consider selecting the latter by minimizing BIC in addition to AIC. The rejection frequencies are reported in Table 7. It is clear that there is little to choose between AIC and BIC, the rejection frequencies not being notably different across the two criteria for any given combination of noise distribution and the sample size n. It is worth noting that results from experiments based on artificial time series confirm the robustness of the properties of the test procedure with respect to the choice of order selection criterion.

References

Andrews D (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59:817–858

Aruoba S (2008) Data revisions are not well behaved. J Money Credit Bank 40:319–340

Bai J, Ng S (2005) Tests for skewness, kurtosis, and normality for time series data. J Bus Econ Stat 23:49–60

Balke N, Fomby T (1994) Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. J Appl Econom 9:181–200

Bao Y (2013) On sample skewness and kurtosis. Econom Rev 32:415–448

Berg A, Paparoditis E, Politis DN (2010) A bootstrap test for time series linearity. J Stat Plan Inference 140:3841–3857

Bickel PJ, Bühlmann P (1997) Closure of linear processes. J Theor Probab 10:445–479

Bowman K, Shenton L (1975) Tables of moments of the skewness and kurtosis statistics in non-normal sampling. Union Carbide Nuclear Divison Report

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

Chen Y, Lin C (2008) On the robustness of symmetry tests for stock returns. Stud Nonlinear Dyn Econom 12:1–37

Davison A, Hinkley D (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Greenspan A (2003) Monetary policy under uncertainty–remarks by chairman alan greenspan at a symposiom of the federal reserve bank of kansas city, vol 19. The Federal Reserve Board, Washington, DC

Haldane AG, Nelson B (2012) Tails of the unexpected. In: Presentation at the conference the credit crisis five years on: unpacking the crisis–University of Edinburgh Business School

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

Harvey DI, Newbold P (2003) The non-normality of some macroeconomic forecast errors. Int J Forecast 19:635–653

Horsewell RL, Looney SW (1993) Diagnostic limitations of skewness coefficients in assessing departures from univariate and multivariate normality. Commun Stat B 22:437–459

Jarque C, Bera A (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Lett 6:255–259

Jarque C, Bera A (1987) A test for normality of observations and regression residuals. Int Stat Rev 55:163–172

Kim T, White H (2004) On more robust estimation of skewness and kurtosis. Finance Res Lett 1:56–73

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

Lahiri K, Teigland C (1987) On the normality of probability distributions of inflation and gnp forecasts. Int J Forecast 3:269–279

Lee S (1998) On the quantile process based on the autoregressive residuals. J Stat Plan Inference 67:17–28

Lehmann EL, Romano JP (2005) Testing statistical hypotheses. Springer, Berlin

Makridakis S, Winkler RL (1989) Sampling distributions of post-sample forecasting errors. Appl Stat 38:331–342

Müller UK (2014) Hac corrections for strongly autocorrelated time series. J Bus Econ Stat 32:311–322

Paulsen J, Tjøstheim D (1985) On the estimation of residual variance and order in autoregressive time series. J R Stat Soc B 47:216–228

Poskitt DS (2007) Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases. Ann Inst Stat Math 59:697–725

Poskitt DS (2008) Properties of the sieve bootstrap for fractionally integrated and non-invertible processes. J Time Ser Anal 29:224–250

Premaratne G, Bera A (2005) A test for symmetry with leptokurtic financial data. J Financ Econom 3:169–187

Psaradakis Z, Vávra M (2015) A quantile-based test for symmetry of weakly dependent processes. J Time Ser Anal 36:587–598

Psaradakis Z, Vávra M (2019) Normality tests for dependent data: large-sample and bootstrap approaches. Commun Stat Simul Comput (forthcoming)

Ramberg J, Schmeiser B (1974) An approximate method for generating asymmetric random variables. Commun ACM 17:78–82

Rayner JCW, Best DJ, Mathews KL (1995) Interpreting the skewness coefficient. Commun Stat A 24:593–600

Reifschneider D, Tulip P (2007) Gauging the uncertainty of the economic outlook from historical forecasting errors. Finance Econ Discuss Ser, vol 60

Reifschneider D, Tulip P (2019) Gauging the uncertainty of the economic outlook using historical forecasting errors: the Federal Reserve's approach. Int J Forecast. (forthcoming)

Sen PK (1968) Asymptotic normality of sample quantiles for m-dependent processes. Ann Math Stat 39:1724–1730

Sharipov OS, Wendler M (2013) Normal limits, nonnormal limits, and the bootstrap for quantiles of dependent data. Stat Probab Lett 83:1028–1035

Shibata R (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Ann Stat 8:147–164

Swanepoel J, Van Wyk J (1986) The bootstrap applied to power spectral density function estimation. Biometrika 73:135–141

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

Thode H (2002) Testing for normality. Marcel Dekker, NY

Tjøstheim D, Paulsen J (1983) Bias of some commonly-used time series estimates. Biometrika 70:389–399

Tulip P, Wallace S (2012) Estimates of uncertainty around the RBA's forecasts. Research Discussion Paper, vol 7