A Strategic Predictive Distribution

for Tests of Probabilistic Calibration

James W. Taylor

Saïd Business School

University of Oxford

International Journal of Forecasting, 2020, 36(4), 1380-1388.

Address for Correspondence:

James W. Taylor
Saïd Business School
University of Oxford
Park End Street
Oxford OX1 1HP, UK
Tel: +44 (0)1865 288927
Email: james.taylor@sbs.ox.ac.uk

A Strategic Predictive Distribution for Tests of Probabilistic Calibration

Abstract

Forecasts of probability distributions are needed to support decision making in many applications. The accuracy of predictive distributions should be evaluated by maximising sharpness subject to calibration. Sharpness relates to the concentration of the predictive distributions, while calibration concerns their statistical consistency with the data. This paper focuses on calibration testing. It is important that a calibration test cannot be gamed by forecasts that have been strategically designed to pass the test. The widely-used tests of probabilistic calibration for predictive distributions are based on the probability integral transform. Drawing on previous results for quantile prediction, we show that strategic distributional forecasting is a concern for these tests. To address this, we provide a simple extension of one of the tests. We illustrate ideas using simulated data.

Key words: Predictive Distributions; Calibration Testing; Probability Integral Transform; Strategic Forecasting.

1. Introduction

Forecasts of probability distributions are needed to support decision making in many applications. For example, predictive distributions are needed for macroeconomic variables to inform policy making (Proietti et al., 2017), and for weather variables to reduce the impact of extreme weather on society (Berrocal et al., 2010). Forecasts of distributions provide predictions of quantiles and other functionals, which are needed, for example, in financial risk management (Nieto and Ruiz, 2016), energy trading (Gianfreda and Bunn, 2018), and for setting safety stock in supply chains (Kolassa, 2016).

The aim of distributional forecasting is to maximise sharpness subject to calibration (Gneiting et al., 2007). Sharpness relates to the concentration of the predictive distributions, while calibration concerns their statistical consistency with the data. If a predictive distribution is calibrated, randomly sampled values from it will be indistinguishable from the observations (Gneiting and Katzfuss, 2014). A scoring rule summarises calibration and sharpness, and is proper if minimised when the forecast is the true distribution. Proper scoring rules encourage honest reporting by forecasters (Gneiting and Raftery, 2007). While scores enable forecasters to be ranked, calibration tests can provide insight leading to improved accuracy.

Quantile forecasts are also evaluated in terms of scores and calibration tests. A consistent quantile score is one that is minimised by the true quantile. A forecast of the α quantile is conditionally calibrated if the conditional probability of an observation falling below the forecast is equal to α. A binary variable, indicating exceedance, should have no autocorrelation, and a mean of α, and this has been the focus of calibration tests. However, Engle and Manganelli (2004) present a quantile forecast that, although very poor, is able to pass such a test. It can be viewed as a dishonest forecaster that has strategically manipulated the forecasts in order to pass the test. Just as consistency is necessary for a quantile score to ensure honest reporting, calibration tests should not permit strategic behaviour. Engle and

Manganelli (2004) provide a regression-based test that cannot be gamed by their strategic forecasts. Strategic forecasting has also been considered in a variety of other settings (see, for example, Olszewski, 2015; Ottaviani and Sørensen, 2006; Lichtendahl et al., 2013).

We show that strategic forecasting is a concern for the widely-used calibration tests for predictive distributions, including the regression-based test of Berkowitz (2001). To overcome this, we draw on the work of Engle and Manganelli (2004) for quantiles to propose an augmented version of the test of Berkowitz (2001), which simply involves the inclusion of an additional regressor in the test. This new test has similarities to a calibration test proposed by Tsyplakov (2014).

We acknowledge that strategic probabilistic forecasts are likely to be exposed as very poor by a visual check. However, a visual check is often not performed. For example, it is impractical when there are many methods or time series, which is typically the case in forecasting competitions. We also acknowledge that strategic forecasts are likely to perform relatively poorly in terms of commonly-used scores. However, there are several reasons why it remains a concern that a calibration test can be gamed. First, comparing forecasting methods may involve a trade-off between the results of a calibration test and a score, and so a seemingly calibrated strategic forecasting method may be viewed as dominating a competitor that has a better score but fails the calibration test. In fact, with best practice being to maximise sharpness subject to calibration, a method that fails a calibration test should not really be considered further. Second, predictive distributions are sometimes evaluated using only calibration. This may be due to the focus being on model specification (Rossi and Sekhposyan, 2014), or it could be due to tradition, the intuitive simplicity and informative nature of calibration tests, or computational reasons. Third, a method that has strategic behaviour to some extent, or for some of the time, may be competitive in terms of both a calibration test and score.

Section 2 lays the foundation for the paper by discussing calibration testing for quantile forecasts. Section 3 shows how the widely-used calibration tests for predictive distributions can be gamed, and presents a simple extension of the test of Berkowitz (2001) to address this problem. Section 4 uses simulated data to illustrate the ideas.

2. Calibration Testing for Quantile Forecasts

A forecast of the α quantile $q_t(\alpha)$ is calibrated if the probability of an observation y_t falling below the forecast is equal to α. More formally, consider the variable $Hit_t = \alpha - I(y_t \leq \hat{q}_t(\alpha))$, where $\hat{q}_t(\alpha)$ is the forecast and $I(\cdot)$ is the indicator function. $\hat{q}_t(\alpha)$ is *unconditionally calibrated* if Hit_t has zero unconditional expectation, and is *conditionally calibrated* if Hit_t has zero conditional expectation, conditional on information available at time *t*-1 (Nolde and Ziegel, 2017).

Unconditional calibration implies that the proportion of observations falling below the quantile forecasts is α. Deviations from this provide insight into how to improve the forecasts. However, one can "game the system" to achieve unconditional calibration by using forecasts equal to unattainably high and low values for proportions α and 1-α, respectively, of the observations. Conditional calibration implies that $\Pr\left(y_t \leq \hat{q}_t(\alpha)\right) = \alpha$. Christoffersen (1998) presents a likelihood ratio test for conditional calibration, which amounts to testing whether Hit_t has zero mean and no autocorrelation. However, Engle and Manganelli (2004) introduce the strategically designed quantile forecast of expression (1), which is clearly very poor, but passes this test for any data generating process (DGP).

$$\hat{q}_t^s(\alpha) = \begin{cases} B_t & \text{if } v_t = 0 \\ A_t & \text{if } v_t = 1 \end{cases} \tag{1}$$

A_t and B_t are values chosen to be above and below the range of possible values for y_t; and the v_t are independent Bernoulli trials, each with probabilities of α and 1-α for outcomes 1 and 0, respectively. Christoffersen's (1998) test is passed because Hit_t has zero mean and no autocorrelation. However, $\hat{q}_t^s(\alpha)$ is not conditionally calibrated because the conditional expectation of Hit_t is not 0, and this can be seen by noting that, once $\hat{q}_t^s(\alpha)$ is known, the value of Hit_t is known.

To address the strategic quantile forecast of expression (1), Engle and Manganelli (2004) develop the *dynamic quantile* test. This test uses a regression framework to perform a joint test of whether Hit_t has zero mean, no autocorrelation, and is independent of the quantile forecast $\hat{q}_t(\alpha)$. Using Hit_{t-1} and $\hat{q}_t(\alpha)$ as regressors, the test's regression is:

$$Hit_t = c + \rho_1 Hit_{t-1} + \rho_2 \hat{q}_t(\alpha) + \varepsilon_t \,, \tag{2}$$

where ε_t is a discrete i.i.d. process. Engle and Manganelli (2004) present a χ^2 test for the null hypothesis of c=0 and ρ_i=0 for all i, which implies conditional calibration. For the strategic forecast of expression (1), ρ_2=0 would be rejected. Pelletier and Wei (2016) suggest that the quantile forecast could be used as the sole regressor in expression (2), because a quantile forecast that reacts too slowly to changing features of the time series will be informative about the probability of the quantile forecast being exceeded.

3. Calibration Testing for Predictive Distributions

In this section, we first provide a brief review of calibration testing, including the widely-used test of Berkowitz (2001). After showing that this test can be gamed, we present an augmented test to overcome this problem.

3.1. Established Tests for Calibration of a Predictive Distribution

The probability integral transform (PIT) is the value of the predictive distribution $\hat{F}_t$ at the observation y_t. It is computed as $p_t = \hat{F}_t(y_t)$, and this is illustrated in Fig. 1. Rosenblatt (1952) observes that a necessary condition for $\hat{F}_t$ to be a correct forecast is that the PIT is i.i.d. U(0,1). In view of this, Diebold et al. (1998) propose that predictive distributions are evaluated by testing the PITs, and that for some applications a pragmatic approach is sufficient, involving just a visual check for uniformity of the histogram of PITs and an inspection of correlograms of the PITs. Gneiting et al. (2007) explain that, for the histogram, a hump shape indicates that $\hat{F}_t$ is, on average, too wide, a U-shape is indicative of $\hat{F}_t$ being too narrow, and a triangular-shape implies that $\hat{F}_t$ is biased. Calibration can, therefore, provide insight into how a predictive distribution can be improved.

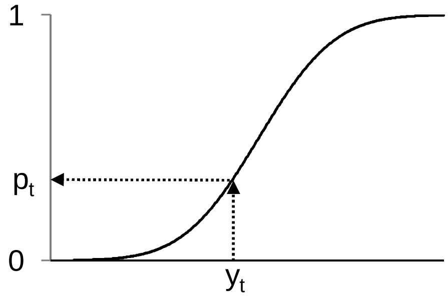

Fig. 1. Generation of the PIT p_t for observation y_t and predictive distribution $\hat{F}_t$.

Gneiting et al. (2007) describe different forms of calibration for a predictive distribution. Uniformity of the PITs is defined as *probabilistic calibration*. Tsyplakov (2011, 2014) explains that definitions of calibration should be clear in terms of conditioning. If an unconditional test is used for the uniformity of the PITs, such as a Kolmogorov-Smirnov test, this would be described as a test for *unconditional probabilistic calibration*. The definition of full probabilistic calibration has the additional requirement that the PITs are independent of information used to produce the forecast, which prompts conditional tests of calibration. Related to this, Mitchell and Wallis (2011) emphasise the importance of testing the independence of the PITs. They define *complete calibration* as the case where the PITs are both uniform and independent. In their empirical analysis, they use a Ljung-Box test for autocorrelation in the PITs, as well as the well-established calibration test of Berkowitz (2001).

Berkowitz's (2001) test involves first transforming the PITs to give $z_t = \Phi^{-1}(p_t)$, where

Φ is the standard normal distribution function. The following regression is then performed[1]:

$$z_t = c + \rho_1 z_{t-1} + \varepsilon_t . \tag{3}$$

If the PITs are i.i.d. U(0,1), c and ρ_1 will be zero, and ε_t will be Gaussian with var(ε_t)=1. Berkowitz (2001) tests these conditions using a likelihood ratio test. The advantage of transforming the PITs to the variable z_t is that there are more tests available to test for normality than uniformity, it is easier to test for autocorrelation under normality than uniformity, and the likelihood ratio test can be based on the commonly used normal likelihood function (Mitchell and Wallis, 2011). As the test of Berkowitz (2001) only has power to test normality through the mean and variance, an additional test for normality should also be performed (see, for example, Proietti et al., 2017). Bao et al. (2007) relax the Gaussian assumption in the test by using a semi-parametric distribution that nests the normal distribution as a special case.

Berkowitz (2001) discusses how the test can be extended to examine higher-order or nonlinear dependence by including additional regressors, and this is considered in the empirical study of Mitchell and Wallis (2011). Clements (2004) describes a version of the Berkowitz (2001) test that can be used to test for unconditional probabilistic calibration when the PITs are potentially serially correlated. For this, the null hypothesis is c=0 and var(ε_t)=(1-ρ_1^2), which implies that var(z_t)=1. In a similar vein, Knüppel (2015) and Rossi and Sekhposyan (2019) present tests of unconditional probabilistic calibration that are robust to potential serial correlation in the PITs. Such serial correlation is likely when dealing with multi-step-ahead prediction, and this is the motivation of Knüppel (2015) who proposes a test based on the raw moments of the standardised PITs. Rossi and Sekhposyan (2019) present a new form of goodness-of-fit test for the distribution of the PITs.

[1] The parameters are estimated by maximising the log-likelihood presented in Appendix 1.

3.2. Gaming the Tests for Calibration of a Predictive Distribution

Hamill (2001) and Gneiting et al. (2007) provide examples of DGP's for which there exist relatively poor predictive distributions that have PITs that are U(0,1). As these DGP's contain no time series dynamics, the PITs are in fact i.i.d. U(0,1), and so it is clear that, although the PITs being i.i.d. U(0,1) is a necessary condition for forecast adequacy, it is not a sufficient condition. However, Mitchell and Wallis (2011) argue that the DGP's used in these examples bear little resemblance to the type of data typically seen in time series forecasting applications. In this section, we do not contribute to this particular debate, but instead show that, regardless of the DGP, it is possible to produce predictive distributions that are clearly very poor, but that can "game the system" to pass tests of probabilistic calibration, such as the Berkowitz (2001) test, as well as simpler tests for independence and uniformity of the PITs.

Consider the following new strategic predictive distribution $\hat{F}_t^s$ and corresponding strategic (discrete) density forecast $\hat{f}_t^s$, which are produced in period *t*-1:

$$\hat{F}_t^s(y) = \begin{cases} 0 & y < B_t \\ u_t & B_t \le y < A_t \\ 1 & A_t \le y \end{cases} \tag{4}$$

$$\hat{f}_t^s(y) = \begin{cases} u_t & y = B_t \\ 1 - u_t & y = A_t \\ 0 & otherwise \end{cases} \tag{5}$$

where A_t and B_t are unattainable upper and lower bounds for the observation y_t; and u_t is a value sampled independently each period from U(0,1). $\hat{F}_t^s$ and $\hat{f}_t^s$ are shown in Figs. 2 and 3, respectively. We discuss the practical issue of choosing A_t and B_t later in this section. We acknowledge that it is perhaps a strange choice of predictive distribution, but we should emphasise that we have chosen it with the specific strategic aim of "gaming" tests for probabilistic calibration, such as the Berkowitz (2001) test.

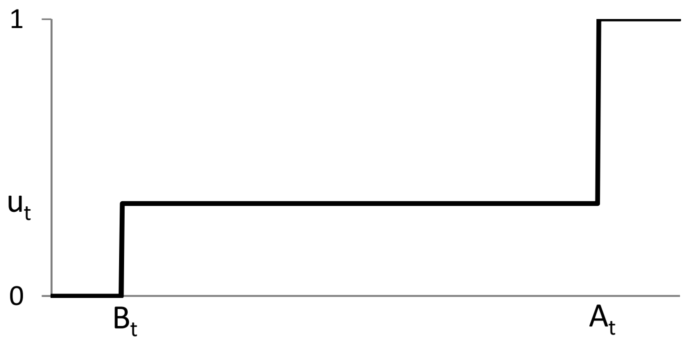

Fig. 2. Strategic predictive distribution $\hat{F}_t^s$ of expression (4).

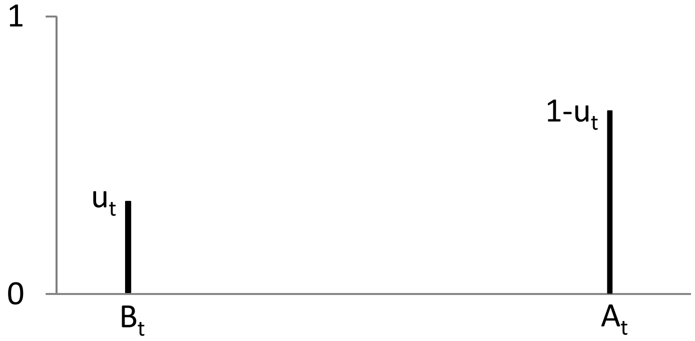

Fig. 3. Strategic (discrete) density forecast $\hat{f}_t^s$ of expression (5).

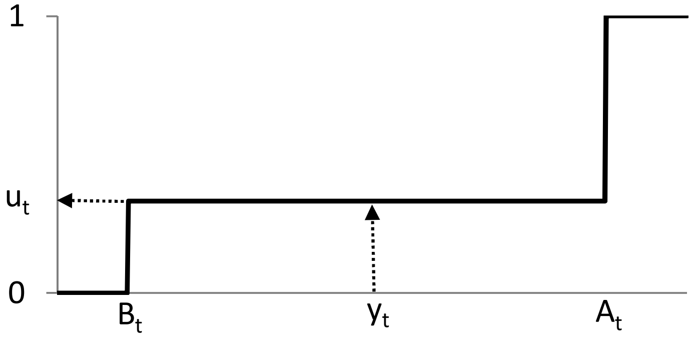

Fig. 4. Generation of the PIT for observation y_t and the strategic predictive distribution $\hat{F}_t^s$ of expression (4).

As the PIT is the value of the predictive distribution at the observation y_t, for the strategic predictive distribution of expression (4), the PIT is equal to u_t, regardless of the value of y_t, and this is illustrated in Fig. 4. As the u_t are generated as i.i.d. U(0,1), it follows that the PITs will be i.i.d. U(0,1). Therefore, the PITs for the strategic predictive distribution will pass tests of probabilistic calibration, including the Berkowitz (2001) test of expression (3). Although our strategic predictive distribution may not pass a visual check, as we said in Section 1, it remains a concern that a strategic predictive distribution exists that will pass probabilistic calibration tests, regardless of the DGP.

We acknowledge that if y_t has unbounded support, unattainable upper and lower bounds cannot be found. However, in finite samples, it is likely that extreme values can be chosen for A_t and B_t that will be exceeded with very low probability, making it virtually impossible to reject the null hypothesis of correct calibration.

Consider also a forecaster who has an accurate forecast for only the mean μ_t. We denote this prediction as $\hat{\mu}_t$. For this situation, it is straightforward for the forecaster to produce a predictive distribution that has mean $\hat{\mu}_t$, and that passes the Berkowitz (2001) test. This is achieved using the strategic predictive distribution of expression (4), with A_t and B_t defined as unattainable upper and lower bounds that satisfy the following expression:

$$u_t B_t + \left(1 - u_t\right) A_t = \hat{\mu}_t \,. \tag{6}$$

PIT values of 0 or 1 are problematic because they cannot be transformed using the standard normal distribution, prior to the application of the Berkowitz (2001) test. To avoid PIT values of 0 or 1 in our simulation study, we replaced the strategic predictive distribution of expression (4) with the following strategic predictive distribution, which is a mixture of Gaussian distributions:

$$\hat{F}_t^{s}\left(y\right) = u_t F_t^{B_t}\left(y\right) + \left(1 - u_t\right) F_t^{A_t}\left(y\right) \tag{7}$$

where $F_t^{A_t}\left(y\right)$ and $F_t^{B_t}\left(y\right)$ are Gaussian distributions with low variance and means equal to A_t and B_t, respectively. An example of this strategic predictive distribution is presented in Fig. 5, which shows that the distribution is similar to the strategic predictive distribution of expression (4) and Fig. 2.

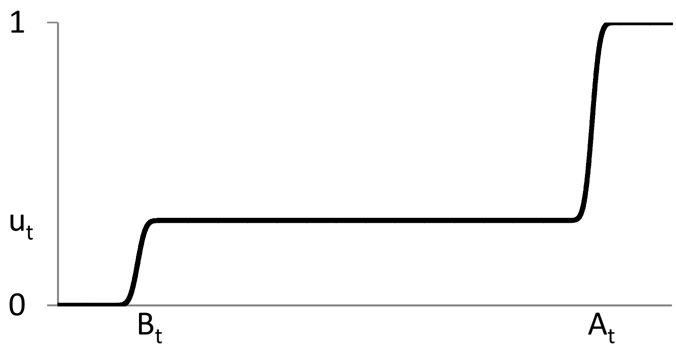

Fig. 5. Strategic predictive distribution $\hat{F}_t^s$ of expression (7).

In our simulation study, we chose A_t and B_t as in the following expressions:

$$A_t = A = \hat{\mu}_y + k\hat{\sigma}_y$$
$$B_t = B = \hat{\mu}_y - k\hat{\sigma}_y$$

where $\hat{\mu}_y$ and $\hat{\sigma}_y$ are the mean and standard deviation of the in-sample observations, and k is a constant factor. We selected k=100, which according to the Chebyshev inequality, implies that at least 99.99% of the observations will fall between A_t and B_t. For common sizes of out-of-sample periods, this percentage is likely to be large enough to ensure that statistical tests will not have sufficient power to reject the null hypothesis of correct calibration. For the Gaussian distributions, $F_t^{A_t}(y)$ and $F_t^{B_t}(y)$ in expression (7), which are centred at A_t and B_t, respectively, we set the standard deviation to be 0.01 multiplied by $\hat{\sigma}_y$.

3.3. Augmenting the Berkowitz Test to Address Strategic Prediction

The strategic predictive distributions of expressions (4) and (7) pass the Berkowitz (2001) test because it tests only for the PITs being i.i.d. U(0,1), when ideally the PIT in period *t* should also be independent of all information known in period *t*-1, the forecast origin. This is not the case for the PIT from the strategic predictive distributions because they are dictated by u_t, which is generated in period *t*-1, and, as shown in Fig. 4, u_t is also the PIT.

In Section 2, the need to ensure independence of all information available at the forecast origin was also apparent for quantile conditional calibration testing. In Engle and Manganelli's

(2004) regression-based test of expression (2), their strategic quantile forecast was exposed as poor by including the forecast itself as a regressor. Similarly, for a predictive distribution, the PIT should be independent of the distributional forecast. In view of this, and given the form of the strategic predictive distributions of expressions (4) and (7), we propose the inclusion of the median or skewness of the distributional forecast as an additional regressor in the Berkowitz (2001) test. The following expression presents the test's regression model, augmented with the median[2] $\hat{m}_t$ of the predictive distribution to give an *augmented Berkowitz test*[3]:

$$z_t = c + \rho_1 z_{t-1} + \rho_2 \hat{m}_t + \varepsilon_t \,. \tag{8}$$

For a calibrated distributional forecast, c=0, ρ_1=0, ρ_2=0, $var(\varepsilon_t)$=1, and ε_t will be Gaussian. Following Berkowitz (2001), a likelihood ratio test can be used to test for these conditions, along with a test for normality. For the strategic predictive distributions of expressions (4) and (7), a relatively high value of the median (or large negative skewness) will correspond to a relatively low value of the PIT, and a relatively low median (or large positive skewness) will correspond to a relatively high PIT. This implies that the hypothesis ρ_2=0 will be rejected, revealing the strategic predictive distribution as being of poor quality.

The ideas in this paper relate closely to the work of Tsyplakov (2011, 2014)[4]. As we mentioned in Section 3.1, he formalises definitions of calibration, emphasising the need to be clear about conditioning. He explains that probabilistic calibration requires that the PITs are uniformly distributed and are independent of the information used to produce the forecast. The strategic predictive distributions that we have presented are probabilistically calibrated. However, they are not *auto-calibrated*. A predictive distribution is defined by Tsyplakov (2011, 2014) to be *auto-calibrated* if the PITs are uniform and independent of the information

[2] As the predictive distribution is a one step-ahead forecast, $\hat{m}_t$ is a forecast produced at time *t*-1 for the median at time *t*.

[3] The parameters are estimated by maximising the log-likelihood presented in Appendix 2.

[4] We are grateful to a reviewer for drawing our attention to the papers of Tsyplakov (2011, 2014).

used to produce the predictive distribution as well as the predictive distribution itself. The test of Berkowitz (2001) only tests a “necessary condition of sequential auto-calibration” (Tsyplakov, 2014). The augmented Berkowitz test that we have proposed is a test for auto-calibration, because it uses information from the forecast itself. Interestingly, it has similarities to a test for auto-calibration considered by Tsyplakov (2014), which essentially involves a test of whether the PIT is correlated with the mean of the predictive distribution. We note that, in the augmented Berkowitz test regression of expression (8), the mean should not be used instead of the median, because this could be “gamed” using the strategic predictive distribution discussed in relation to expression (6).[5]

We should point out that it would be too bold to claim that the augmented Berkowitz test cannot with certainty be gamed by some other form of strategic predictive distribution. However, the augmented test does address the only strategic predictive distribution that we can envisage that can pass the Berkowitz (2001) test for any DGP.

In Section 2, we noted that Pelletier and Wei (2016) suggest that the quantile calibration test of expression (2) could be implemented with the quantile forecast as sole regressor, as the forecast itself is a form of summary of information available at the forecast origin. In a similar way, the augmented test of expression (8) could be reduced to an alternative Berkowitz test that has the median forecast as sole regressor, as in the following:

$$z_t = c + \rho_1 \hat{m}_t + \varepsilon_t .$$

[5] It is worth noting that a strategic forecast of the conditional mean can easily be produced to game a common test of bias in which the forecast error is used as dependent variable with just an intercept included in the regression. Bias is assessed by testing for zero intercept. For example, a strategic conditional mean forecast can be generated as a randomly sampled value from a normal distribution with mean set as the unconditional mean of the historical observations, and variance set as the product of a large positive number and the variance of the historical observations. The bias test behaves like a test for probabilistic calibration. Instead, a test of auto-calibration of the conditional mean forecasts could be used in which the forecast itself is included as regressor. This leads to the regression of Mincer and Zarnowitz (1969). We are grateful to a reviewer for highlighting these issues.

4. Illustration with Simulated Data

We now use simulated data to show how tests of probabilistic calibration can be gamed, and how this problem can be overcome by the augmented version of the Berkowitz (2001) test.

4.1. Data Generating Processes

We simulated data using the following autoregressive (AR) process of order 1:

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

where ε_t is i.i.d. N(0,1). We chose ϕ_1 to be either 0.5 or 0.9. With n_1 and n_2 defined as the size of the in-sample and out-of-sample periods, respectively, we considered: n_1=10000 and n_2=1000; n_1=1000 and n_2=100; n_1=n_2=100; n_1=10 and n_2=100; and n_1=100 and n_2=10. For each choice of n_1, n_2 and ϕ_1, we generated 10000 time series. For each series, method parameters were re-estimated on a rolling basis using a moving window of n_1 observations to give a set of n_2 one step-ahead out-of-sample forecasts. In selecting pairs of values for n_1 and n_2, we were interested to experiment with a variety of values that might be considered in practice. However, we should acknowledge that if n_1, n_2 and the ratio n_1/n_2 are small, there can be size distortions in the tests of forecast performance due to estimation error of the type described by West (1996) and West and McCracken (1998).

4.2. Forecasting Methods

The first method that we applied to each series involved the unrealistic assumption that the DGP and its parameters were known. For period t, the predictive distributions were Gaussian with mean of $\phi_1 y_{t-1}$ and unit variance. We refer to this as the *true DGP* method.

We implemented two benchmark forecasting methods. The first used least squares to fit an AR(1) model with intercept, and produced predictive distributions that were Gaussian with mean equal to the one step-ahead forecast, and variance given by the regression forecast error variance expression. We refer to this as the *fitted AR(1)* model. The other benchmark method constructed the predictive distribution as a Gaussian distribution with mean and standard deviation set as the mean $\hat{\mu}_y$ and standard deviation $\hat{\sigma}_y$ of the in-sample observations. As these are estimates of the unconditional mean and standard deviation, we refer to this as the *unconditional* method. This method is also sometimes termed the *climatological* method (see, for example, Gneiting et al., 2007).

We implemented the strategic predictive distribution of expression (7) with A_t, B_t and the standard deviations of $F_t^{A_t}(y)$ and $F_t^{B_t}(y)$ chosen as described in the final paragraph of Section 3.2.

4.3. Out-of-Sample Evaluation

Unconditional probabilistic calibration can be assessed by testing the uniformity of the PIT histogram. Fig. 6 enables a visual check of this for the out-of-sample strategic predictive distributions applied to the first of the 10000 series generated from the DGP with ϕ_1=0.9 and n_1=10000 and n_2=1000.

In all the tests that we consider, we use a nominal size of 5%. In other words, we test with a 5% significance level, which implies that 5% is the ideal value for the percentage of the 10000 simulated series for which the null hypothesis was rejected.

Table 1 presents the results of the Kolmogorov-Smirnov test for the distribution of the PITs being U(0,1). This is a test of unconditional calibration. Each column in the table corresponds to one of the different types of simulated series. Each value in the table is the percentage of the 10000 simulated series for which U(0,1) was rejected at the 5% significance level. In Table 1, the value is close to the nominal size of 5% for the true DGP method and the strategic method when n_1=10000 and n_2=1000. The percentages are also reasonably close to 5% elsewhere in the table, except for the unconditional method. This is perhaps surprising because, at least in large samples, this method is unconditionally calibrated by construction. However, our results for this method are consistent with those of Mitchell and Wallis (2011), who explain that the Kolmogorov-Smirnov test for uniformity assumes the data is a random sample, which is not the case with the PITs from this method.

A necessary condition for full probabilistic calibration is that the PIT series is not serially correlated. We tested for this using a Ljung-Box test that examined autocorrelation up to lag 2. Each value in Table 2 is the percentage of the 10000 series for which the null hypothesis of no autocorrelation was rejected at the 5% significance level. The table shows that the results of the strategic method matched those of the true DGP, with values close to the nominal size of 5%, except for the case with just 10 observations in the out-of-sample period.

We implemented the version of the test of Berkowitz (2001) proposed by Clements (2004), which tests for unconditional probabilistic coverage. We discussed this version in Section 3.1. Table 3 reports the percentage of the 10000 series for which unconditional coverage was rejected at the 5% significance level. As in Tables 1 and 2, we see the strategic method performing as well as the true DGP. In assessing the results of Table 3, it is worth noting that the test proposed by Clements (2004) only allows for serial correlation of the transformed PITs if it is an AR(1) process, which may well not be the case.

Table 4 presents the results for the standard calibration test of Berkowitz (2001), which uses expression (3). Each value is the percentage of the 10000 series for which calibration was rejected at the 5% significance level. As expected, the value is close to the nominal size of 5% for the unrealistic true DGP method, although the performance weakens when the out-of-sample period is small. The Berkowitz (2001) test results for the true DGP are matched for all sample sizes by the strategic method, which shows the limitation of this test. Indeed, the strategic method performs better than the fitted AR(1) model for the cases in which the in-sample size is 100 or 10. Comparing the results for the unconditional method in Tables 3 and 4, we see that, while unconditional probabilistic calibration was often not rejected in Table 3, calibration is generally rejected with the Berkowitz (2001) test in Table 4.

Table 5 presents the results for our augmented Berkowitz test, which is based on expression (8). Reassuringly, calibration for the strategic method is rejected in all cases, except when the out-of-sample period is small, in which case calibration was rejected for more than 90% of the series. For the unconditional method, when using a small out-of-sample period, calibration is more often rejected in Table 5 than it was in Table 4, which suggests that the augmented test has more power than the standard Berkowitz (2001) test.

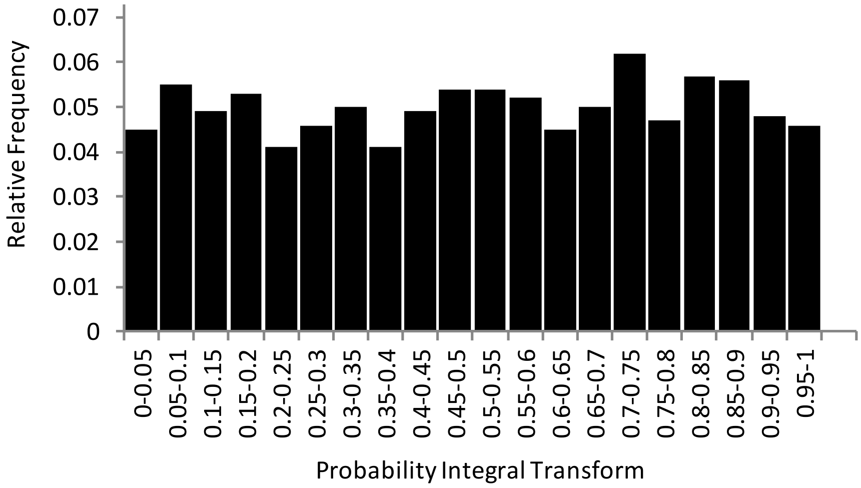

Fig. 6. PIT histogram for the out-of-sample prediction from the strategic method applied to the DGP with AR(1) parameter ϕ_1=0.9, in-sample size n_1=10000 and out-of-sample size n_2=1000. Kolmogorov-Smirnov statistic is 0.0222, which is below the 5% critical value, so unconditional probabilistic calibration is not rejected.

Table 1
Kolmogorov-Smirnov test for unconditional probabilistic calibration (uniformity of PITs). Each value is the percentage of simulated series for which calibration was rejected at 5% sig. level. ϕ_1 is the coefficient of the AR(1) DGP, and n_1 and n_2 are the in-sample and out-of-sample sizes, respectively.

ϕ_1	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**
n_1	**10000**	**10000**	**1000**	**1000**	**100**	**100**	**10**	**10**	**100**	**100**
n_2	**1000**	**1000**	**100**	**100**	**100**	**100**	**100**	**100**	**10**	**10**
True DGP	4.7	4.7	3.5	3.5	3.4	3.4	3.7	3.7	1.9	1.9
Fitted AR(1)	4.6	4.5	3.6	3.6	1.8	2.9	0.4	1.5	2.1	3.0
Unconditional	21.8	73.1	19.1	70.4	11.8	62.8	2.2	66.4	12.2	49.4
Strategic	4.5	4.5	3.6	3.6	3.7	3.7	3.7	3.7	1.8	1.8

Table 2
Ljung-Box test for autocorrelation in the PITs. Each value is the percentage of simulated series for which hypothesis of no autocorrelation was rejected at 5% sig. level. ϕ_1 is the coefficient of the AR(1) DGP, and n_1 and n_2 are the in-sample and out-of-sample sizes, respectively.

ϕ_1	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9
n_1	10000	10000	1000	1000	100	100	10	10	100	100
n_2	1000	1000	100	100	100	100	100	100	10	10
True DGP	4.9	4.9	5.7	5.7	5.5	5.5	5.8	5.8	15.9	15.9
Fitted AR(1)	4.9	4.8	5.5	5.6	3.6	6.0	22.3	63.3	15.6	15.9
Unconditional	100.0	100.0	99.2	100.0	99.2	100.0	97.4	100.0	21.0	46.1
Strategic	4.9	4.9	5.7	5.7	5.7	5.7	5.7	5.7	14.8	14.8

Table 3
Unconditional Berkowitz test for unconditional probabilistic calibration, which is based on expression (3), with null hypothesis c=0 and var(ε_t)=1-ρ_1^2 (i.e. var(z_t)=1). Each value is the percentage of simulated series for which calibration was rejected at 5% sig. level. ϕ_1 is the coefficient of the AR(1) DGP, and n_1 and n_2 are the in-sample and out-of-sample sizes, respectively.

ϕ_1	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9
n_1	10000	10000	1000	1000	100	100	10	10	100	100
n_2	1000	1000	100	100	100	100	100	100	10	10
True DGP	5.3	5.3	5.9	5.9	5.4	5.4	5.9	5.9	11.8	11.8
Fitted AR(1)	5.3	5.3	6.0	6.1	1.8	2.9	50.0	83.4	11.9	14.1
Unconditional	10.6	49.1	11.7	51.4	5.2	39.2	96.1	100.0	21.2	64.4
Strategic	4.9	4.9	5.5	5.5	5.6	5.6	5.5	5.5	11.3	11.3

Table 4
Standard Berkowitz test for probabilistic calibration, which is based on expression (3), with null hypothesis c=ρ_1=0 and var(ε_t)=1. Each value is the percentage of simulated series for which calibration was rejected at 5% sig. level. ϕ_1 is the coefficient of the AR(1) DGP, and n_1 and n_2 are the in-sample and out-of-sample sizes, respectively.

ϕ_1	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9
n_1	10000	10000	1000	1000	100	100	10	10	100	100
n_2	1000	1000	100	100	100	100	100	100	10	10
True DGP	5.0	5.0	5.4	5.4	5.1	5.1	5.5	5.5	6.8	6.8
Fitted AR(1)	5.0	4.9	5.7	5.8	1.2	3.5	62.6	96.4	7.0	8.7
Unconditional	100.0	100.0	99.8	100.0	99.5	100.0	99.6	100.0	28.7	90.7
Strategic	4.9	4.9	5.3	5.3	5.2	5.2	5.2	5.2	6.4	6.4

Table 5
Augmented Berkowitz test for probabilistic calibration, which is based on expression (8), with null hypothesis $c=\rho_1=\rho_2=0$ and var(ε_t)=1. Each value is the percentage of simulated series for which calibration was rejected at 5% sig. level. ϕ_1 is the coefficient of the AR(1) DGP, and n_1 and n_2 are the in-sample and out-of-sample sizes, respectively.

ϕ_1	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**
n_1	**10000**	**10000**	**1000**	**1000**	**100**	**100**	**10**	**10**	**100**	**100**
n_2	**1000**	**1000**	**100**	**100**	**100**	**100**	**100**	**100**	**10**	**10**
True DGP	5.0	5.1	6.3	7.3	6.2	7.1	6.0	6.9	64.5	31.6
Fitted AR(1)	5.2	5.2	6.6	7.4	9.5	4.8	99.9	99.7	68.8	34.6
Unconditional	100.0	100.0	100.0	100.0	99.8	100.0	100.0	100.0	97.9	99.6
Strategic	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.6	93.2

Table 6
CRPS averaged across simulated series. Lower values are better. ϕ_1 is the coefficient of the AR(1) DGP, and n_1 and n_2 are the in-sample and out-of-sample sizes, respectively.

ϕ_1	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.9**
n_1	**10000**	**10000**	**1000**	**1000**	**100**	**100**	**10**	**10**	**100**	**100**
n_2	**1000**	**1000**	**100**	**100**	**100**	**100**	**100**	**100**	**10**	**10**
True DGP	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
Fitted AR(1)	0.56	0.56	0.56	0.56	0.57	0.57	0.66	0.67	0.57	0.57
Unconditional	0.65	1.29	0.65	1.30	0.65	1.29	0.68	0.97	0.66	1.29
Strategic	7.69	15.27	7.68	15.11	7.58	13.70	6.71	7.83	7.58	13.58

Table 6 evaluates the methods using the continuous ranked probability score (CRPS), defined in such a way that lower values are preferable. As this is a proper scoring rule for distributions (Gneiting and Raftery, 2007), it is no surprise to see the strategic method performing badly in this table, as it was designed purely to game the tests for probabilistic calibration. This confirms our comment in Section 1 that, although our strategic predictive distribution passes the tests for calibration, a strategic predictive distribution will always be exposed as very poor by a well-chosen score.

In this paper, we have implemented the strategic predictive distribution of expression (7) with A_t and B_t chosen so that the distribution is very wide in order to ensure probabilistic calibration. As a result of the distribution being so wide, the CRPS is very poor. A narrower predictive distribution can be chosen that delivers much better CRPS, with only small deterioration in the results of the probabilistic calibration tests. Further improvements in the CRPS can be achieved by using the strategic predictive distribution with A_t and B_t chosen, as in expression (6), so that the mean is equal to an accurate point forecast produced by another method. Although the CRPS would probably still be worse than the fitted AR(1) model in Table 6, it does raise the issue that there are potentially strategic predictive distributions that, in some circumstances, may not be poor in terms of the CRPS. This raises the practical importance of seeking tests for calibration that cannot be gamed.

5. Summary

We have shown the existence of a strategic predictive distribution that can game the established tests for probabilistic calibration, including the regression-based test of Berkowitz (2001). To address this, we have proposed a simple augmented version of this test. Although strategic predictive distributions may well be exposed as poor by well-chosen scores or visual checks, it remains a practical concern if they are able to game a widely-used calibration test. Best practice is to seek probabilistic forecasts that maximise sharpness subject to calibration. This suggests that calibration should be viewed as a necessity, which is problematic when, as in our simulation study, a strategic method is the only forecasting method producing satisfactory results for the established tests of probabilistic calibration. This problem is addressed by using the augmented version of the test.

Appendix 1

As discussed in Section 3.1, the standard test of Berkowitz (2001) involves the estimation of the model $z_t = c + \rho_1 z_{t-1} + \varepsilon_t$. Let us write *var*($\varepsilon_t$)=$\sigma_\varepsilon^2$. The parameters *c*, ρ_1 and σ_ε are estimated by maximising the following exact log-likelihood function:

$$-\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\left[\sigma_\varepsilon^2/(1-\rho_1^2)\right]-\frac{(z_1-c/(1-\rho_1))^2}{2\sigma_\varepsilon^2/(1-\rho_1^2)}$$
$$-\frac{T-1}{2}\log(2\pi)-\frac{T-1}{2}\log(\sigma_\varepsilon^2)-\sum_{t=2}^{T}\left(\frac{(z_t-c-\rho_1 z_{t-1})^2}{2\sigma_\varepsilon^2}\right).$$

Note that the same log-likelihood expression is given by Berkowitz (2001), but it is incorrectly described as corresponding to the model of expression (3) of that paper.

Appendix 2

In Section 3.3, we introduced the augmented Berkowitz test, which involves estimating the model $z_t = c + \rho_1 z_{t-1} + \rho_2 \hat{m}_t + \varepsilon_t$. We write *var*($\varepsilon_t$)=$\sigma_\varepsilon^2$. To obtain the exact log-likelihood, we use the steps described, for example, by Biørn (2011). We first rewrite the model as

$$(1-\rho_1 L)z_t = c + \rho_2 \hat{m}_t + \varepsilon_t .$$

Using this, we can write:

$$z_1 = c/(1-\rho_1) + \rho_2 \hat{m}_1 + \rho_2 \sum_{i=1}^{\infty} \rho_1^i \hat{m}_{1-i} + \sum_{i=0}^{\infty} \rho_1^i \varepsilon_{1-i} . \qquad (9)$$

In order to obtain the mean and variance of z_1, we introduce constant parameters $\mu_{\rho_1\hat{m}}$ and $\sigma_{\rho_1\hat{m}}^2$ for the mean and variance of the penultimate summation of expression (9). The exact log-likelihood function is then written in terms of the parameters *c*, ρ_1, ρ_2, σ_ε, $\mu_{\rho_1\hat{m}}$ and $\sigma_{\rho_1\hat{m}}$ as:

$$-\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\left[\rho_2^2\sigma_{\rho_1\hat{m}}^2+\sigma_\varepsilon^2/(1-\rho_1^2)\right]-\frac{(z_1-c/(1-\rho_1)-\rho_2(\hat{m}_1+\mu_{\rho_1\hat{m}})^2}{2\left[\rho_2^2\sigma_{\rho_1\hat{m}}^2+\sigma_\varepsilon^2/(1-\rho_1^2)\right]}$$
$$-\frac{T-1}{2}\log(2\pi)-\frac{T-1}{2}\log(\sigma_\varepsilon^2)-\sum_{t=2}^{T}\left(\frac{(z_t-c-\rho_1 z_{t-1}-\rho_2\hat{m}_t)^2}{2\sigma_\varepsilon^2}\right).$$

Acknowledgements

We are very grateful to the handling editor, an associate editor and two referees for providing very useful comments on the paper.

References

Bao, Y., Lee, T. H., & Saltoğlu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, *26*(3) 203-225.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, *19*(4) 465-474.

Berrocal, V.J., Raftery, A.E. Gneiting, T., & Steed, R.C. (2010). Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association*, *105*(490) 522-537.

Biørn, E. (2011). Estimation of ARX and VARX models by ML. Unpublished lecture notes. ECON 5101 Advanced Econometrics, Lecture note no. 6.

Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4) 841-862.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *The Economic Journal*, 114(498), 844-866.

Diebold, F.X., Gunther, T.A., & Tay, A. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*(4) 863-883.

Engle, R.F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4) 367-381.

Gianfreda, A., & Bunn, D.W. (2018). A stochastic latent moment model for electricity price formation. *Operations Research*, *66*(5) 1189-1203.

Gneiting, T., Balabdaoui, F., & Raftery, A.E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2) 243-268.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and its Application*, *1*(1) 125-151.

Gneiting, T., & Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477) 359-378.

Hamill, T.M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*(3) 550-560.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, *33*(2), 270-281.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting, 32*(3) 788-803.

Lichtendahl Jr, K.C., Grushka-Cockayne, Y., & Pfeifer, P.E. (2013). The wisdom of competitive crowds. *Operations Research, 61*(6) 1383-1398.

Mincer, J. A., & Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance* (pp. 3-46). NBER.

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

Nieto, M.R., & Ruiz, E. (2016). Frontiers in VaR forecasting and backtesting. *International Journal of Forecasting, 32*(2) 475-501.

Nolde, N., & Ziegel, J.F. (2017). Elicitiability and backtesting: perspectives for banking regulation. *Annals of Applied Statistics*, *11*(4) 1833-1874.

Olszewski, W. (2015). Calibration and expert testing. In *Handbook of Game Theory with Economic Applications,* edited by H. P. Young and S. Zamir, *4* 949-984, Elsevier.

Ottaviani, M., & Sørensen, P.N. (2006). The strategy of professional forecasting. *Journal of Financial Economics, 81*(2) 441-466.

Pelletier, D., & Wei, W. (2015). The geometric-VaR backtesting method. *Journal of Financial Econometrics 14*(4) 725-745.

Proietti, T., Marczak, M., & Mazzi, G. (2017). Euromind-$\mathcal{D}$: A density estimate of monthly gross domestic product for the Euro Area. *Journal of Applied Econometrics*, *32*(3) 683-703.

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

Rossi, B., & Sekhposyan, T. (2014). Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set. *International Journal of Forecasting*, *30*(3) 662-682.

Rossi, B., & Sekhposyan, T. (2019). Alternative tests for correct specification of conditional predictive densities. *Journal of Econometrics*, *208*(2), 638-657.

Tsyplakov, A. (2011). Evaluating density forecasts: a comment. MPRA Paper 31184, University Library of Munich, Germany.

Tsyplakov, A. (2014). Theoretical guidelines for a partially informed forecast examiner. MPRA Paper 67333, University Library of Munich, Germany.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, 1067-1084.

West, K. D., & McCracken, M. W. (1998). Regression-based tests of predictive ability (No. t0226). National Bureau of Economic Research.