Received: 7 January 2013 | Revised: 24 April 2018 | Accepted: 3 May 2019

DOI: 10.1002/for.2604

RESEARCH ARTICLE

A likelihood ratio and Markov chain-based method to evaluate density forecasting

Yushu Li[1,2] | Jonas Andersson[3]

[1] Department of Mathematics, University of Bergen, Bergen, Norway

[2] Department of Economics and Statistics, Linnaeus University, Småland, Sweden

[3] Department of Business and Management Science, Norwegian School of Economics, Bergen, Norway

Correspondence
Yushu Li, Department of Mathematics, University of Bergen, Norway.
Email: yushu.li@uib.no

Funding information
Finance Market Fund, Norwegian Research Council, Grant/Award Number: 274569

Abstract

In this paper, we propose a likelihood ratio-based method to evaluate density forecasts, which can jointly evaluate the unconditional forecasted distribution and dependence of the outcomes. Unlike the well-known Berkowitz test, the proposed method does not require a parametric specification of time dynamics. We compare our method with the method proposed by several other tests and show that our methodology has very high power against both dependence and incorrect forecasting distributions. Moreover, the loss of power, caused by the nonparametric nature of the specification of the dynamics, is shown to be small compared to the Berkowitz test, even when the parametric form of dynamics is correctly specified in the latter method.

KEYWORDS

density forecasting, likelihood ratio test, Markov chain

1 | INTRODUCTION

An evaluation of the quality of forecasts can have different purposes. It could be to determine whether point forecasts are, on average, hitting the actual outcome not yet observed. It could be, for example, in a risk management context, to investigate whether interval forecasts have the coverage probability the model used would imply. The evaluation of point forecasts is typically done by comparing different forecasting models and investigating whether one has a significantly larger expected loss function. This loss function could be mean squared error (MSE), mean absolute error (MAE) or, in cases where available, economic loss incurred by using a forecast compared to having the actual values. Examples on papers dealing with the evaluation of point forecasts are Wallis (1995), Diebold and Lopez (1996), and Gneiting (2011). Interval forecasts are evaluated by the relative frequency of an interval to cover the actual outcome (Chatfield, 1993; Granger, White, & Kamstra, 1989). An often cited paper on the evaluation of interval forecasting is Christoffersen (1998), which proposed a theory to evaluate the interval forecast. This evaluation procedure is based on the likelihood ratio test and, owing to the additivity of the likelihood ratio test, the method can jointly test the unconditional coverage and independence by testing the correct conditional coverage. This test and its extensions (Berkowitz, Christoffersen, & Pelletier, 2011; Clements & Taylor, 2003; Dumitrescu, Hurlin, & Madkour, 2013; Engle & Manganelli, 2004) are most widely used to evaluate an interval forecast, especially in the value-at-risk (VaR) analysis, which can be viewed as a one-sided interval forecast.

Finally, an even more detailed forecast is the density forecast, which estimates the probability density of a future value of the process, conditional on the observations used in the forecast. Point and interval forecasts can then be seen as a by-product of this as they are, for example, the mean and quantiles in this conditional density. Tay and Wallis (2000) carried out a survey of density forecasting. They pointed out the necessity of an accurate forecast of the probability density in applications such as macroeconomics—for example, of inflation and output growth—and in finance—for example, of portfolio returns, risk management, and volatility. The literature on evaluating the uncertainty of the density forecast is limited and mainly based on the idea of the probability integral transform (PIT) or its extension (Berkowitz, 2001; Diebold, Gunther, & Tay, 1998; Diebold, Hahn, & Tay, 1999; Tay & Wallis, 2000). Among these few papers, the forecasting evaluation framework proposed by Berkowitz (2001) is the most widely applied, because of its comparatively good small-sample power performance. Wallis (2003) proposed using Pearson chi-squared based statistics, which can evaluate the goodness of fit and independence at the same time. This paper will extend the likelihood ratio-based method of Christoffersen (1998) in order to evaluate density forecasting. Owing to the additivity of the likelihood ratio test, our method can jointly test the unconditional distribution and independence. Moreover, our test is a nonparametric test and no parametric model is needed for the independence test. We will compare our new method with the evaluation framework proposed by Berkowitz and the Kolmogorov–Smirnov (KS) test.

The paper is divided into the following sections: Section 2 introduces the likelihood interval forecast, Section 3 describes our evaluation method for density forecasting, and Section 4 compares the new method with previous ones by means of a Monte Carlo experiment. A conclusion closes the paper.

2 | LIKELIHOOD RATIO AND MARKOV CHAIN-BASED INTERVAL FORECAST

For the ex post realization $Y = (y_1, y_2, \dots, y_T)$, the ex ante interval forecast made at time $t-1$ is $C_{t\mid t-1}(p) = [L_{t\mid t-1}(p), U_{t\mid t-1}(p)]$, where p is the probability of coverage. Define the indicator variable $\{I_t\}_{t=1}^{T}$ as

$$I_t = \begin{cases} 1, & y_t \in C_{t|t-1}(p) \\ 0, & y_t \notin C_{t|t-1}(p). \end{cases}$$

That is, $I_t = 1$ when the ex post realization lies inside $C_{t\mid t-1}(p)$ and $I_t = 0$ otherwise. Christoffersen (1998) constructed a test framework to evaluate whether $C_{t\mid t-1}(p) = [L_{t\mid t-1}(p), U_{t\mid t-1}(p)]$ is an "efficient" interval forecast with respect to the past information $\Psi_{t-1} = \{I_t, I_{t-1}, \dots\}$ by testing whether $E(I_t \mid \Psi_{t-1}) = p$. The evaluation framework includes three tests:

1. *The unconditional coverage test statistic* $\mathrm{LR_{ud}}$ to test whether the expected value of the indicator sequence $\{I_t\}_{t=1}^{T}$ is equal to the coverage rate. This test ignores the dependence of I_t and the null hypothesis is $E(I_t) = p$, while the alternative hypothesis is $E(I_t) = \pi \neq p$. Define n_0 as $n_0 = \mathrm{sum}[I_t = 0]$ and $n_1 = \mathrm{sum}[I_t = 1]$. The likelihoods under the null and alternative hypotheses are $L_p = (1-p)^{n_0} p^{n_1}$ and $L_{\widehat{\pi}} = (1-\widehat{\pi})^{n_0} \widehat{\pi}^{n1}$ respectively, where the relative hit frequency $\widehat{\pi} = n_1/(n_0 + n_1)$ is the maximum likelihood estimate (MLE) of π. Then the likelihood ratio-based test statistic $\mathrm{LR_{ud}} = -2\log\left(L_p/L_{\widehat{\pi}}\right) \sim \chi^2(1)$ under the null hypothesis. Christoffersen (1998) reported that the pure unconditional coverage test will have very low power and is inefficient when $\{I_t\}_{t=1}^{T}$ is clustered in a time-dependent fashion. He therefore introduced an independence test and a joint test for independence and unconditional coverage.
2. *The independence test statistic* $\mathrm{LR_{ind}}$ to test whether I_t is independent over the whole period. Independence means that there are no clusters of violation in certain time periods and lack of violations in others. The likelihood ratio-based test statistic $\mathrm{LR_{ind}}$ is constructed by using a first-order Markov chain with two states. We will provide a detailed illustration of $\mathrm{LR_{ind}}$ in Section 3, where we construct our density forecasting evaluation method, which is based on a k-states Markov chain.
3. *Conditional coverage test statistic* $\mathrm{LR_{cd}}$ to test whether the forecasting interval has correct conditional coverage in the form $E(I_t \mid \Psi_{t-1}) = p$. As the test of unconditional coverage and independence will not affect each other, this conditional coverage test is the combination of the unconditional coverage test and the independence test. Owing to the additivity of the likelihood ratio test statistics (Bera & McKenzie, 1985), we have $\mathrm{LR_{cd}} = \mathrm{LR_{ud}} + \mathrm{LR_{ind}}$, which can jointly test the randomness and correct coverage, while the test of individual subcomponents can still be retained.

The likelihood ratio test by Christoffersen (1998) has been followed by several developments in the literature (Berkowitz et al., 2011; Clements & Taylor, 2003; Dumitrescu et al., 2013; Engle & Manganelli, 2004) in terms of both theoretical extensions and applications.

3 | LIKELIHOOD RATIO AND MARKOV CHAIN-BASED DENSITY FORECAST EVALUATIONS

Methods to evaluate the density forecast (Berkowitz, 2001; Diebold et al., 1998, 1999; Tay & Wallis, 2000) are, to a large extent, built on the seminal paper of Diebold et al. (1998) using the probability integral transform (PIT*)*. The main idea is that when the ex ante forecasted distribution $\{s_t(y_t)\}_{t=1}^{T}$ is correct, then for the ex post realization $Y = (y_1, y_2, \ \ldots \ , y_T)$, we have that $x_t = \int_{-\infty}^{y_t} s_y(u)du \sim \text{i.i.d.}U(0,\,1)$. Deviation from i.i.d. means that the ex ante forecast fails to capture the underlying time dynamics of the data-generating process (DGP). Deviation from U(0,1) implies that the used model yields an incorrect forecast distribution. Berkowitz (2001) used the PIT to formulate a formal test of density forecasts. It is constructed by transforming the PIT through the inverse distribution function of the standard normal distribution and thus under the null hypothesis of a correctly specified forecasting model, obtaining normally distributed variables. A parametric model for the dependence is then formulated for these normally distributed variables. The parameters of this model are then tested for independence of time by means of a likelihood ratio test. A simultaneous test of independence and distributional shape is also constructed. The idea of combining the goodness of fit and independence tests was given by Wallis (2003), where the interval evaluation method of Christoffersen (1998) is formulated in the framework of a contingency table-based Pearson chi-squared test. While density forecasts are mentioned in Wallis (2003), they concentrate on interval forecasting evaluation based on contingency tables for small samples. To summarize, there are two main strands of this literature. The first, mainly due to Christoffersen, is nonparametric as it does not require any specification of the time dependence of the forecast distribution. It deals with interval forecasts. The other, mainly due to Berkowitz, requires the specification of time dependence but can be used to test the density forecasts and not only intervals. As far as we know, there exist no nonparametric methods for the evaluation of density forecasts. Our proposed method will fill this gap and extend the likelihood ratio evaluation method by Christoffersen for interval forecast to density forecast. The method is still constructed in three steps: a test for goodness of fit, a test for independence, and a joint test for goodness of fit and independence.

1. *Unconditional density test statistic* LR$_{\text{ud}}$: Consider the ex post outcome $Y = (y_1, y_2, \ \ldots \ , y_T)$, which is generated by the distribution $f(y_t)$ and the ex ante forecasted density $s(y_t)$. The range of y_t is $[I_0, I_n]$ with $I_0 < y_t < I_n$. We divide $[I_0, I_n]$ into k mutually exclusive states as $\left[\underbrace{I_0,\, I_1}_{1}, \ldots, \underbrace{I_{k-1},\, I_n}_{k}\right]$ and let the number of y_t which lie in state i be n_i. Note that the interval forecasting is a special case where $k = 2$ and the test statistic LR$_{\text{ud}}$ is actually based on the likelihood from a binomial distribution. To evaluate whether $s(y_t)$ yields the correct description of the unconditional probabilities of future values is equivalent to testing $f(y_t) = s(y_t)$. Under the null hypothesis $f(y_t) = s(y_t)$, $N = (n_1, n_2, \ \ldots \ , n_k)$ follows a multinomial distribution $multinom(T, p_1 \ldots p_k)$ with event probability $p_i = \int_{I_{i-1}}^{I_i} s_y(u)du$. Thus the likelihood function under the null hypothesis is

$$L(p) = \frac{T!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k},$$

where $p_i = \int_{I_{i-1}}^{I_i} s_y(u)du$.

The likelihood function under the alternative hypothesis is $L(\hat{p}) = [T!/(n_1! \ldots n_k!)]\hat{p}_1^{n_1} \ldots \hat{p}_k^{n_k}$, where $\hat{p}_i = n_i/T$ is the MLE of the event probability over the whole parameter space. The likelihood ratio test (LRT) statistic is $\text{LR}_{\text{ud}} = -2\log[L(p)/L(\hat{p})]$ and $\text{LR}_{\text{ud}} \sim \chi^2(k\,-\,1)$ under the null hypothesis. Just as the unconditional coverage test statistic LR$_{\text{uc}}$ in interval forecast, LR$_{\text{ud}}$ can only discover the biasedness of the forecasted distribution with the null hypothesis being $s(y_t) = f(y_t)$, and it can be viewed as a pure goodness-of-fit test.

2. *Independence test statistic* LR$_{\text{ind}}$: Wallis (2003) reported that the test for independence in the interval forecast could be extended to the density forecast, without analyzing this. The following will provide a detailed illustration of how to do this. The independence is tested against a k-state first-order Markov chain. Let $\pi_{ij} = \ \Pr\,(y_t \in \text{state } j | y_{t\,-\,1} \in \text{state } i)$. Then, the Markov chain is specified with the transition probability matrix

$$\Pi = \begin{bmatrix} \pi_{11} \ldots \pi_{1k} \\ \pi_{i,j} \\ \pi_{k1} \ldots \pi_{kk} \end{bmatrix}.$$

Let n_{ij} denote that the number of events where a state i is followed by a state j as $n_{ij} = Nr(y_t; y_t \in j \ \& \ y_{t\,-\,1} \in i)$. Then, the likelihood function under the alternative hypothesis for the whole process is

$$L(\Pi)=\left(\pi_{11}^{n_{11}}....\pi_{1k}^{n_{1k}}\right)...\left(\pi_{i1}^{n_{i1}}....\pi_{ik}^{n_{ik}}\right)...\left(\pi_{k1}^{n_{k1}}....\pi_{kk}^{n_{kk}}\right)=\prod_{i=1}^{k}\prod_{j=1}^{k}\pi_{ij}^{n_{ij}},$$

with $\widehat{\pi}_{ij}=n_{ij}/\sum_{j=1}^{k}n_{ij}$ being the MLE of π_{ij}. Under the null hypothesis of independence, the present outcome will not be influenced by past information. Thus, when the outcome y_t is in state j, the previous outcome y_{t-1} has the same probability of lying in any state and this can be denoted by $\pi_{1j}=\pi_{2j}\ ...\ =\pi_{kj}=\pi_{.j}$. Thus we have

$$\left(\pi_{11}^{n_{11}}....\pi_{1k}^{n_{1k}}\right)...\left(\pi_{i1}^{n_{i1}}....\pi_{ik}^{n_{ik}}\right)...\left(\pi_{k1}^{n_{k1}}....\pi_{kk}^{n_{kk}}\right)=\prod_{j=1}^{k}\pi_{.j}^{n_{.j}},$$

where $n_{.j}=\sum_{i=1}^{k}n_{ij}$. As $\pi_{.j}$ is actually the probability that an outcome lies in state j and $n_{.j}$ is the number of outcomes that lies in state j, the MLE of $\pi_{.j}$ is $\widehat{\pi}_{.j}=n_j/T$ with $n_j=n_{.j}$. Therefore, the likelihood function under the null hypothesis is $L\left(\widehat{\Pi}_0\right)=\prod_{j=1}^{k}\left(n_j/T\right)^{n_j}$ and the unrestricted likelihood function is $L\left(\widehat{\Pi}_1\right)=\prod_{i=1}^{k}\prod_{j=1}^{k}\left(n_{ij}/\sum_{j=1}^{k}n_{ij}\right)^{n_{ij}}$. The LRT for independence is then

$$\mathrm{LR}_{\mathrm{ind}}=-2\log\frac{L\left(\widehat{\Pi}_0\right)}{L\left(\widehat{\Pi}_1\right)}\sim\chi^2\left[(k-1)^2\right].$$

We note that $L\left(\widehat{\Pi}_0\right)\propto L(\widehat{p})$ and this relationship will simplify the joint test statistics in the following paragraph.

3. *Conditional density test statistic* $\mathrm{LR}_{\mathrm{cd}}$: To test whether the conditional forecasted density distribution based on the past information $s(y_t)\mid\Omega_{t-1}$ provides correct conditional probabilities for events associated with future actual outcomes. As the conditional coverage test statistic $\mathrm{LR}_{\mathrm{cd}}$ in the situation of interval forecasting, this test can be viewed as a combination of a goodness-of-fit test and a test for independence; we test whether $s(y_t)=f(y_t)$ and whether $\{y_t\}_{t=1}^{T}$ is independent. The test statistic can be constructed based on the additivity of the LRT: The test statistic to test a joint hypothesis is the sum of the test statistics which test the components of the null hypothesis separately. Then the test statistic $\mathrm{LR}_{\mathrm{cd}}$, which can jointly test the independence and goodness of fit, is $\mathrm{LR}_{\mathrm{cd}}=\mathrm{LR}_{\mathrm{ud}}+\mathrm{LR}_{\mathrm{ind}}$, where

$$\mathrm{LR}_{\mathrm{ud}}=-2\log\frac{L_p}{L_{\widehat{\pi}}}=-2\log\left(\frac{\frac{T!}{n_1!...n_k!}p_1^{n_1}...p_k^{n_k}}{\frac{T!}{n_1!...n_k!}\widehat{p}_1^{n_1}...\widehat{p}_k^{n_k}}\right)$$
$$=-2\left[\log\left(p_1^{n_1}...p_k^{n_k}\right)-\log\left(\widehat{p}_1^{n_1}...\widehat{p}_k^{n_k}\right)\right]$$

and

$$\mathrm{LR}_{\mathrm{ind}}=-2\log\frac{L\left(\widehat{\Pi}_0\right)}{L\left(\widehat{\Pi}_1\right)}=-2\log\left[\frac{\prod_{j=1}^{k}\left(\frac{n_j}{T}\right)^{n_j}}{\prod_{i=1}^{k}\prod_{j=1}^{k}\left(\frac{n_{ij}}{\sum_{j=1}^{k}n_{ij}}\right)^{n_{ij}}}\right]$$
$$=-2\left\{\log\left[\prod_{j=1}^{k}\left(\frac{n_j}{T}\right)^{n_j}\right]-\log\left[\prod_{i=1}^{k}\prod_{j=1}^{k}\left(\frac{n_{ij}}{\sum_{j=1}^{k}n_{ij}}\right)^{n_{ij}}\right]\right\},$$

with $\widehat{p}_j=n_j/T$. Then $\mathrm{LR}_{\mathrm{cd}}=\mathrm{LR}_{\mathrm{ud}}+\mathrm{LR}_{\mathrm{ind}}$ can be simplified as

$$\mathrm{LR}_{\mathrm{cd}}=-2\left\{\log\left(p_1^{n_1}...p_k^{n_k}\right)-\log\left[\prod_{i=1}^{k}\prod_{j=1}^{k}\left(\frac{n_{ij}}{\sum_{j=1}^{k}n_{ij}}\right)^{n_{ij}}\right]\right\}$$
$$\sim\chi^2[k(k-1)],$$

where $p_i=\int_{I_{i-1}}^{I_i}s_y(u)du$. Compared with $\mathrm{LR}_{\mathrm{ud}}$, which only has power against biased unconditional forecasted densities and ignores the internal dependence of $\{y_t\}_{t=1}^{T}$, $\mathrm{LR}_{\mathrm{cd}}$ has power against both misspecified density forecasting and internal dependence of the data series. Therefore, instead of only testing the unbiasedness of the forecasted distribution, $\mathrm{LR}_{\mathrm{cd}}$ can discover time dependence such as autocorrelation or conditional heteroscedasticity in the forecast errors.

The $\mathrm{LR}_{\mathrm{cd}}$ test can then be applied to evaluate the efficiency of density forecasts. Under the null hypothesis $s(y_t)\mid\Omega_{t-1}=f(y_t)$, or $s(y_t)=f(y_t)$ and $\{y_t\}_{t=1}^{T}$ is independent, we have

$$\mathrm{LR}_{\mathrm{cd}}=\mathrm{LR}_{\mathrm{ud}}+\mathrm{LR}_{\mathrm{ind}}\sim\chi^2[k(k-1)].$$

To investigate the performance of the test statistics $\mathrm{LR}_{\mathrm{ud}}$, $\mathrm{LR}_{\mathrm{ind}}$ and $\mathrm{LR}_{\mathrm{cd}}$, a Monte Carlo study is carried out in the next section. The benchmark we use is the evaluation framework proposed by Berkowitz (2001) but we also compare against the Kolmogorov–Smirnov (KS) test. Diebold et al. (1998) reported that when the ex ante forecasted distribution, $\{s_t(y_t)\}_{t=1}^{T}$, is produced by a correctly specified model, then

$$x_t=\int_{-\infty}^{y_t}s_y(u)du\sim\text{i.i.d.}U(0,\,1).$$

However, Berkowitz (2001) showed that the test based on $\{x_t\}_{t=1}^{T}$ displayed low power in sample sizes smaller than 1,000. Instead, Berkowitz proposed a test based on transformation of x_t, $z_t = \Phi^{-1}(x_t)$ where Φ is the standard normal cumulative distribution function. Under the null hypothesis $s(y_t) \mid \Omega_{t-1} = f(y_t)$, z_t = i. i. d. $N(0,1)$. Berkowitz further developed the test within the likelihood ratio framework, which can test both independence and density distribution. However, in this test we need to specify a parametric model for $\{z_t\}_{t=1}^{T}$ under the alternative hypothesis. For example, to test the null against a first-order autoregressive model, an AR(1) model $z_t - \mu = \rho(z_{t-1} - \mu) + \varepsilon_t$ can be used. The null hypothesis $s(y_t) \mid \Omega_{t-1} = f(y_t)$ is then $\mu = 0$, $\rho = 0$ and $\sigma^2 = 1$. Let $L(\mu, \sigma^2, \rho)$ denote the likelihood ratio function of (1) and $\widehat{\mu}$, $\widehat{\sigma}^2$, $\widehat{\rho}$are the estimated values for μ,σ^2,ρ. The likelihood ratio test of independence across the observations is then

$$\mathrm{Ber}_{\mathrm{ind}} = -2\left[L\left(\widehat{\mu}, \widehat{\sigma}^2, 0\right) - L\left(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho}\right)\right].$$

Under the null that the observations are independent, $\mathrm{Ber}_{\mathrm{ind}} \sim \lambda^2(1)$. A joint likelihood ratio test to test both independent and correct density forecasting is then

$$\mathrm{Ber} = -2\left[L(0, 1, 0) - L\left(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho}\right)\right].$$

Under the null hypothesis $s(y_t) \mid \Omega_{t-1} = f(y_t)$, $\mathrm{Ber} \sim \lambda^2(3)$. As an alternative parametric model is needed in Berkowitz (2001), the test can be viewed as a semiparametric test. Instead, all the tests proposed in our paper, $\mathrm{LR_{ud}}$,$\mathrm{LR_{ind}}$, and $\mathrm{LR_{cd}}$, are nonparametric.

4 | MONTE CARLO STUDY

The null hypotheses in the Monte Carlo study are that the forecasted density distributions $s(y_t)$ for $\{y_t\}_{t=1}^{T}$ are, respectively, the independent normal distribution, the independent $t(6)$ distribution, and the independent truncated Cauchy distribution. When $s(\cdot)$ is the independent normal distribution, the mean and standard errors are estimated from simulated data. When $s(\cdot)$ is the independent truncated Cauchy distribution, the upper and lower bounds are set as maximum and minimum of the generated data. We have chosen these three distributions because, based on the density function graph, they look similar to each other and we therefore need formal tests. The DGP for $\{y_t\}_{t=1}^{T}$ will be designed to check how the tests will perform from both size and power perspectives, and it can be divided into three cases:

Case 1. y_t~i. i. d. $N(0,1)$, y_t~i. i. d. $t(6)$, y_t~i. i. d. tCauch(−10,10)

Case 2. $y_t = n_t\sqrt{h_t}$; $h_t = 0.15 + 0.15y_{t-1}^2 + 0.70h_{t-1}$, $n_t \tilde{}\,\mathrm{i.i.d.}N(0, 1)$, $n_t \tilde{}\,\mathrm{i.i.d.}t(6)$

Case 3. $y_t = n_t\sqrt{h_t}$; $h_t = 0.15 + 0.70y_{t-1}^2 + 0.15h_{t-1}$, $n_t \tilde{}\,\mathrm{i.i.d.}N(0, 1)$, $n_t \tilde{}\,\mathrm{i.i.d.}t(6)$

The DGP in Case 1, with no time dependence, is used to investigate size for the three test statistics in Section 3 and also power for $\mathrm{LR_{ud}}$ and $\mathrm{LR_{cd}}$. The DGP's in Cases 2 and 3 are used to investigate the power of the three tests as there exists GARCH(1,1)-type dependence in these processes. Because of its relevance in risk management, the GARCH(1,1) model is a common model used in previous research to evaluate interval forecasting (Christoffersen, 1998; Clements & Taylor, 2003) and density forecasting (Bao, Lee, & Saltoglu, 2007; Diebold et al., 1998). We set the sample size N to 100, 250, 500 and 1,000. Following Sturges' rule (Sturges, 1926), which is used to decide the ideal bin width in constructing histogram, the number of states k is initially chosen as the integer value of 1+$\log_2(T)$ for finite sample size T, and the interval length for each state is as identical. In the case that the sample size and number of states k converge to infinity then the test would converge to a true density forecast test. If there exist empty bins based on the initial division, we combine the nearby bins until each bin contains observations. However, based on Sturges rule, when T = 250, 500 and 1,000, the integer values of 1+$\log_2(T)$ are 9, 10, and 11, respectively, and we seldom come across the situation that a bin contains 0 observations.

The number of Monte Carlo replications is 10,000. We first investigate the size properties for all the tests: $\mathrm{LR_{ud}}$, $\mathrm{LR_{ind}}$, and $\mathrm{LR_{cd}}$, the Berkowitz (2001) tests, Ber and $\mathrm{Ber_{ind}}$, and the Kolmogorov–Smirnov (KS) test. The results are presented in Table 1.

In Table 1, the forecasted distribution $s(\cdot)$ is the same as the true distribution $f(\cdot)$ of the DGP. For the Monte Carlo simulation we use 10,000 replications, implying that the approximate 95% confidence interval for the estimated size at a nominal 5% significant level is

$$0.05 \pm 1.96^* \sqrt{\frac{0.05(1-0.05)}{10,000}} = (0.0457, 0.0543).$$

Table 1 shows that when the DGP is i. i. d. $t(6)$ or i. i. d. tCauch(−10,10), the sizes are mostly unbiased or nearly unbiased. When the DGP is i. i. d. $N(0,1)$, $\mathrm{LR_{ud}}$ and $\mathrm{LR_{ind}}$ have somewhat too large rejection rates for small samples. This size distortion decreases when sample size increases. On the other hand, when the DGP is i. i. d. $N(0,1)$, the size of KS and Ber tests is smaller than nominal size, and this size distortion will not be improved when sample size increases.

We next investigate the properties of the tests of goodness of fit when the data are generated from a DGP with

TABLE 1 Size of the tests when $s(\cdot) = f(\cdot)$

DGP	i.i.d. N(0,1)				i.i.d. t(6)				i.i.d. tCauch(−10,10)			
N	100	250	500	1,000	100	250	500	1,000	100	250	500	1,000
KS	0.000	0.000	0.000	0.033	0.041	0.050	0.053	0.033	0.037	0.043	0.052	0.058
LR_{ud}	0.015	0.012	0.020	0.048	0.055	0.044	0.049	0.048	0.021	0.058	0.041	0.045
LR_{cd}	0.072	0.070	0.068	0.052	0.058	0.054	0.033	0.045	0.020	0.040	0.052	0.048
Ber	0.005	0.007	0.007	0.006	0.054	0.053	0.055	0.055	0.058	0.061	0.048	0.053
LR_{ind}	0.090	0.087	0.080	0.061	0.060	0.057	0.042	0.045	0.021	0.042	0.050	0.060
Ber_{ind}	0.040	0.047	0.033	0.050	0.042	0.050	0.045	0.052	0.041	0.043	0.053	0.049

no time dependence. The series $\{y_t\}_{t=1}^{T}$ are still generated from Case 1, but the forecasted distributions $s(\cdot)$ and the true distribution $f(\cdot)$ are varied. Moreover, for the Ber test, we will specify two alternative models when constructing the test statistic: one where z_t follows an AR(1) model in the equation $z_t - \mu = \rho(z_{t-1} - \mu) + \varepsilon_t$, and the other when it follows a GARCH model, $z_t = n_t\sqrt{h_t};\ h_t = c + a z_{t-1}^2 + b h_{t-1}$. When the alternative model is an AR(1) model, we denote the Berkowitz (2001) independent test by Ber_{ind} and the joint test by Ber. When the alternative model is a GARCH model, we denote the Berkowitz (2001) independent test by BerG_{ind} and the joint test by BerG. Tables 2, 3, and 4 present the power properties for the test statistics KS, LR_{ud}, Ber, and BerG, and the size properties for the test statistics LR_{ind}, Ber_{ind}, and BerG_{ind}.

For the power in Table 2, the Ber test generally has the highest power in the goodness-of-fit test, while the following are then BerG, LR_{ud}, LR_{cd}, and KS. Almost all the power will approach one when sample size increases to larger than 500. However, as can be seen in Table 3, when $s(\cdot)$ is an independent normal distribution, Ber and BerG have almost no power even when the sample size increases. This is explained in Dowd (2004), who shows that a deviation from normality of the transformed data, which is what happens in our case, makes it difficult for the test to detect deviations from the null hypothesis. The power of LR_{ud}, LR_{cd}, and KS approaches one as sample size increases, while LR_{ud} and LR_{cd} have higher power than KS for all the sample sizes. The sizes of the independence tests LR_{ind} and Ber_{ind} are quite close to the nominal size of 5%, but BerG_{ind} is seriously oversized

TABLE 3 Power of the goodness-of-fit test and size of independence test (The underlined entries) when s is i. i. d. Normal

DGP	i. i. d. t(6)				i. i. d. tCauch(−10,10)			
N	100	250	500	1,000	100	250	500	1,000
KS	0.011	0.037	0.115	0.423	0.613	0.996	1.000	1.000
LR_{ud}	0.094	0.306	0.674	0.946	0.913	1.000	1.000	1.000
LR_{cd}	0.085	0.120	0.265	0.643	0.470	0.979	1.000	1.000
Ber	0.006	0.014	0.029	0.054	0.001	0.001	0.004	0.005
BerG	0.018	0.027	0.051	0.093	0.04	0.060	0.072	0.085
LR_{ind}	0.070	0.049	0.033	0.031	0.015	0.030	0.056	0.074
Ber_{ind}	0.050	0.060	0.037	0.041	0.046	0.034	0.045	0.051
BerG_{ind}	0.067	0.063	0.083	0.092	0.116	0.154	0.181	0.205

TABLE 2 Power of the goodness-of-fit test and size of independence test (The underlined entries) when s is i.i.d. t(6) and i. i. d. tCauchy

	s is i.i.d. t(6)								s is i. i. d. tCauchy							
DGP	i. i. d. N(0,1)				i. i. d. tCauch(−10,10)				i. i. d. N(0,1)				i. i. d. t(6)			
N	100	250	500	1,000	100	250	500	1,000	100	250	500	1,000	100	250	500	1,000
KS	0.047	0.059	0.089	0.210	0.305	0.829	1.000	1.000	0.053	0.096	0.509	0.982	0.053	0.204	0.791	0.998
LR_{ud}	0.098	0.304	0.741	0.994	0.983	1.000	1.000	0.084	0.097	0.577	0.981	1.000	0.122	0.741	0.999	1.000
LR_{cd}	0.150	0.199	0.411	0.814	0.867	0.998	1.000	1.000	0.134	0.348	0.837	0.999	0.094	0.362	0.916	1.000
Ber	0.273	0.670	0.973	1.000	0.999	1.000	1.000	1.000	0.070	0.680	1.000	1.000	0.071	0.814	1.000	1.000
BerG	0.142	0.579	0.951	1.000	0.997	1.000	1.000	1.000	0.025	0.604	0.986	1.000	0.026	0.751	0.999	1.000
LR_{ind}	0.011	0.100	0.094	0.084	0.020	0.024	0.045	0.063	0.133	0.112	0.089	0.090	0.081	0.050	0.039	0.035
Ber_{ind}	0.052	0.054	0.049	0.060	0.051	0.053	0.047	0.061	0.056	0.069	0.045	0.048	0.051	0.051	0.049	0.054
BerG_{ind}	0.006	0.014	0.034	0.014	0.022	0.033	0.034	0.048	0.020	0.021	0.015	0.013	0.029	0.027	0.023	0.014

TABLE 4 Power of the tests when *s* is i.i.d. *t*(6) and DGP is from case 2 and case 3

	Case 2								**Case 3**							
DGP	**i.i.d. *t*(6)**				**i.i.d. *N*(0,1)**				**i.i.d. *t*(6)**				**i.i.d. *N*(0,1)**			
N	**100**	**250**	**500**	**1,000**	**100**	**250**	**500**	**1,000**	**100**	**250**	**500**	**1,000**	**100**	**250**	**500**	**1,000**
KS	0.165	0.277	0.427	0.704	0.100	0.180	0.300	0.535	0.318	0.550	0.756	0.972	0.703	0.978	1.000	1.000
LR$_{cd}$	0.333	0.579	0.840	0.974	0.270	0.448	0.732	0.977	0.803	0.980	0.998	1.000	0.903	0.995	0.987	1.000
Ber	0.427	0.648	0.828	0.979	0.504	0.721	0.900	0.993	0.633	0.623	0.664	0.711	0.861	0.938	0.992	0.999
BerG	0.617	0.937	0.998	1.000	0.639	0.966	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000
LR$_{ind}$	0.183	0.337	0.607	0.881	0.169	0.221	0.373	0.616	0.609	0.880	0.978	0.992	0.509	0.892	0.999	0.997
Ber$_{ind}$	0.097	0.103	0.117	0.118	0.073	0.084	0.089	0.083	0.225	0.261	0.285	0.302	0.190	0.225	0.277	0.275
BerG$_{ind}$	0.558	0.941	0.998	0.999	0.297	0.738	0.962	1.000	0.977	1.000	1.000	1.000	0.934	1.000	1.000	1.000

when the DGP is i. i. d. tCauch(−10,10). A possible cause of the failure of BerG$_{ind}$ is that the thick tails of the data are captured by the GARCH(1,1) model fitted under the alternative hypothesis—a model with thicker tails than the standard normal assumed under the null hypothesis.

Next, we investigate the power against both lack of fit and dependence. In order to do this the DGPs of Cases 2 and 3 are used. The power properties for test statistics KS, LR$_{cd}$, Ber, BerG, LR$_{ind}$, Ber$_{ind}$, and BerG$_{ind}$ are shown in Tables 4 and 5.

When $s(\cdot)$ is i. i. d. $t(6)$, Table 4 show that BerG has higher power than both Ber and LR$_{cd}$. This result is not surprising since, for BerG, the alternative model is correctly specified as a GARCH model, while LR$_{cd}$ is agnostic about the form of dependence. When $s(\cdot)$ is an independent normal distribution, Table 5 show that BerG has substantial power against GARCH-type dependence despite the incorrectly specified error distribution. The two extra estimated parameters are apparently not causing too much uncertainty.

We would also like to highlight the main advantage of our proposed test by showing the failure of Ber and Ber$_{ind}$ to detect dependence when it is not correctly parametrized. This is not a shortcoming of these tests but simply a consequence of the tradeoff between uncertainty and precision. Table 5 shows that Ber and Ber$_{ind}$ have very low power when the DGP is i. i. d. $N(0,1)$ for all the sample sizes and the alternative hypothesis is an AR(1) model. This shows that correctly specifying an alternative hypothesis in Berkowitz (2001) is crucial to guarantee a high power of the test. The BerG$_{ind}$ test has the highest power and Ber$_{ind}$ the lowest for all the cases when it comes to detecting GARCH-type forecast distributions. The nonparametric nature of the LR$_{ind}$ and LR$_{cd}$ tests is therefore naturally placed between the two Berkowitz-type tests and it is an empirical question whereabouts they are placed. For the cases in our Monte Carlo study, we find it fair to claim that they are working very well.

Based on Tables 1–5, we conclude that the tests proposed in our paper—LR$_{ud}$,LR$_{ind}$, and LR$_{cd}$—have good size and power properties. In applications, analogously with the test by Christoffersen (1998), the tests can be carried out in a natural sequence. The first step is to apply LR$_{cd}$ to jointly test the independence and goodness of

TABLE 5 Power of the tests when *s* is i. i. d. normal and DGP is from case 2 and case 3

	Case 2								**Case 3**							
DGP	**i.i.d. *t*(6)**				**i.i.d. *N*(0,1)**				**i.i.d. *t*(6)**				**i.i.d. *N*(0,1)**			
N	**100**	**250**	**500**	**1000**	**100**	**250**	**500**	**1000**	**100**	**250**	**500**	**1000**	**100**	**250**	**500**	**1000**
KS	0.069	0.308	0.689	0.967	0.001	0.002	0.002	0.005	0.485	0.924	0.999	1.000	0.111	0.420	0.728	0.975
LR$_{cd}$	0.212	0.619	0.910	0.997	0.118	0.193	0.374	0.590	0.808	0.930	0.999	1.000	0.523	0.950	0.999	1.000
Ber	0.023	0.139	0.318	0.590	0.016	0.021	0.017	0.018	0.327	0.688	0.948	1.000	0.112	0.309	0.544	0.821
BerG	0.551	0.922	1.000	1.000	0.239	0.645	0.939	1.000	0.959	1.000	1.000	1.000	0.894	1.000	1.000	1.000
LR$_{ind}$	0.156	0.319	0.606	0.890	0.155	0.220	0.392	0.589	0.627	0.926	0.969	0.992	0.478	0.891	0.985	0.999
Ber$_{ind}$	0.122	0.155	0.175	0.182	0.085	0.093	0.096	0.118	0.325	0.349	0.425	0.452	0.212	0.242	0.297	0.290
BerG$_{ind}$	0.718	0.966	1.000	1.000	0.441	0.831	0.989	1.000	0.984	1.000	1.000	1.000	0.941	1.000	1.000	1.000

TABLE 6 Size of the LR tests when $s(\cdot) = f(\cdot)$, sample size = 250 and size =500 (in bold)

DGP	i. i. d. N(0,1)				i. i. d. t(6)				i. i. d. tCauch(−10,10)			
Bin no.	7	9	15	20	7	9	15	20	7	9	15	20
LR$_{\text{ud}}$	0.013	0.016	0.016	0.011	0.040	0.044	0.039	0.062	0.049	0.052	0.049	0.038
	0.013	**0.017**	**0.015**	**0.023**	**0.043**	**0.043**	**0.034**	**0.034**	**0.058**	**0.048**	**0.0415**	**0.05**
LR$_{\text{ind}}$	0.098	0.101	0.058	0.006	0.080	0.060	0.016	0.001	0.060	0.040	0.001	0.000
	0.103	**0.090**	**0.068**	**0.023**	**0.058**	**0.046**	**0.024**	**0.001**	**0.099**	**0.050**	**0.001**	**0.000**
LR$_{\text{cd}}$	0.058	0.073	0.042	0.001	0.080	0.051	0.018	0.002	0.061	0.043	0.004	0.000
	0.072	**0.070**	**0.047**	**0.020**	**0.064**	**0.040**	**0.015**	**0.001**	**0.088**	**0.047**	**0.002**	**0.000**

TABLE 7 Power of the LR tests when s is i.i.d. t(6) and DGP is from case 2 and Case3, sample size = 250 and size =500 (in bold)

	Case 2								Case 3							
DGP	i.i.d. t(6)				i.i.d. N(0,1)				i.i.d. t(6)				i.i.d. N(0,1)			
Bin no.	7	9	15	20	7	9	15	20	7	9	15	20	7	9	15	20
LR$_{\text{cd}}$	0.575	0.563	0.564	0.554	0.448	0.459	0.476	0.442	0.820	0.923	0.935	0.939	0.956	0.982	0.992	0.982
	0.754	**0.788**	**0.792**	**0.777**	**0.672**	**0.749**	**0.739**	**0.756**	**0.804**	**0.980**	**0.998**	**0.998**	**0.950**	**0.998**	**1.000**	**1.000**
LR$_{\text{ind}}$	0.357	0.363	0.120	0.020	0.221	0.229	0.096	0.011	0.720	0.908	0.871	0.580	0.780	0.887	0.705	0.285
	0.554	**0.601**	**0.340**	**0.116**	**0.330**	**0.361**	**0.193**	**0.045**	**0.640**	**0.970**	**0.997**	**0.943**	**0.892**	**0.999**	**0.989**	**0.832**
LR$_{\text{cd}}$	0.607	0.602	0.375	0.171	0.444	0.458	0.242	0.036	0.880	0.984	0.982	0.835	0.973	0.996	1.000	0.870
	0.791	**0.845**	**0.674**	**0.447**	**0.719**	**0.740**	**0.540**	**0.197**	**0.811**	**0.995**	**1.000**	**1.000**	**0.992**	**0.999**	**1.000**	**1.000**

fit. If the null hypothesis is not rejected, we can conclude that $s(y_t)$ is the proper distribution and that the time dependence in the data has been captured by the forecasting model. However, if we reject the null hypothesis, we test whether the rejection is due to the dependence or to an incorrectly specified distribution by applying LR$_{\text{ud}}$ and LR$_{\text{ind}}$ separately.

As mentioned, the number of states k in Tables 1–5 is initially chosen by Sturges' rule (Sturges, 1926) with k is integer value $1+\log_2(T)$. To investigate whether this choice is reasonable, we performed the simulation study for the sample sizes $T = 250$ and $T = 500$ with k set to 7, the integer value of $1+\log_2(T)$, 15, and 20. As mentioned previously, for $T = 250$, the integer value of $1+\log_2(T)$is 9, whereas for $T = 500$ it is 10. First, the size performance was studied and the result is shown in Table 6 for $T = 250$ and $T = 500$.

Overall, our conclusion is that Sturges' rule works well. We also check how the power changes with k. Table 7 presents power against both lack of fit and dependence with DGP of Cases 2 and 3.

Based on the above size and power tables, we conclude that choosing k based on Sturges' rule yields the best performance in terms of unbiased size (except when $s(\cdot) = f(\cdot) =$ i. i. d. N(0,1) where $k = 15$ end up in more stable size of the test) and highest power. For the cases we study here, the most striking observation is that for the larger k values the tests are, in general, undersized.

5 | CONCLUSION

This paper proposes a test framework for the evaluation of density forecasts. It is an extension of the interval forecasting tests of Christoffersen (1998). We show that the proposed tests have high power against two types of time dependence, even though no parametric specification of this dependence is needed in the test. The power is compared to the parametric tests proposed by Berkowitz (2001) and shown to be competitive with them, even in situations when the parametric form in Berkowitz tests is correctly specified. When the dependence is incorrectly specified in Berkowitz tests, the proposed tests outperform them.

ACKNOWLEDGMENT

Yushu Li gratefully acknowledges funding from the Finance Market Fund, Norwegian Research Council (Project number 274569).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study, including the R code, are available from the corresponding author upon reasonable request.

ORCID

Yushu Li https://orcid.org/0000-0003-4105-9925
Jonas Andersson https://orcid.org/0000-0002-2899-6562

REFERENCES

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

Bera, A. K., & McKenzie, C. R. (1985). Alternative forms and properties of the score test. *Journal of Applied Statistics*, *13*, 13–25.

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

Berkowitz, J., Christoffersen, P., & Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science*, *57*(12), 2213–2227.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, *11*, 121–135.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*, 840–841.

Clements, M., & Taylor, N. (2003). Evaluating interval forecasts of high frequency financial data. *Journal of Applied Econometrics*, *18*, 445–456.

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

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

Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination. In G. S. Maddala & C. R. Rao (Eds.), *Handbook of statistics 14: Statistical methods in finance* (pp. 241–268). Amsterdam, Netherlands: North-Holland.

Dowd, K. (2004). A modified Berkowitz back-test. *Risk Magazine*, *17*(4), 86–87.

Dumitrescu, E. L., Hurlin, C., & Madkour, J. (2013). Testing interval forecasts: A GMM-based approach. *Journal of Forecasting*, *32*(2), 97–110.

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

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762.

Granger, C. W. J., White, H., & Kamstra, M. (1989). Interval forecasting: An analysis based upon ARCH quantile estimators. *Journal of Econometrics*, *40*, 87–96.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, *21*, 65–66.

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

Wallis, K. F. (1995). Large-scale macroeconometric modeling. In M. H. Pesaran, & M. R. Wickens (Eds.), *Handbook of applied econometrics* (pp. 312–355). Oxford, UK: Blackwell.

Wallis, K. F. (2003). Chi-square tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting*, *19*, 165–175.

AUTHOR BIOGRAPHIES

Yushu Li is Associate Professor of Statistics in the Department of Mathematics at the University of Bergen (UiB). Before working at UiB, Li worked for two years as an Assistant Professor at the Norwegian School of Economics (NHH) and before that worked as a full-time researcher in the economics department at Lund University. Her research interests are time series analysis, econometric modelling and wavelet methods and, more recently, statistical sparse learning methods.

Jonas Andersson is a professor at Norwegian School of Economics. He received his PhD in statistics at Uppsala University in 1999. His research interests centers around statistical modeling and its applications to business and economics.

How to cite this article: Li Y, Andersson J. A likelihood ratio and Markov chain-based method to evaluate density forecasting. *Journal of Forecasting*. 2019;1–9. https://doi.org/10.1002/for.2604