On density forecast evaluation

Diks, C.

Publication date
2008

Published in
Aenorm

On Density Forecast Evaluation

Traditionally, probability integral transforms (PITs) have been popular means for evaluating density forecasts. For an ideal density forecast, the PITs should be uniformly distributed on the unit interval and independent. However, this is only a necessary condition, and not a sufficient one, as shown by some simple examples. I discuss an alternative approach to density forecast evaluation, via the Kullback-Leibler information criterion (KLIC), and illustrate it with a small simulation study.

Cees Diks

studied theoretical physics at Utrecht University and obtained a PhD (1996) in mathematics from Leiden University for his research on nonlinear time series analysis. In 1998, after a two-year postdoctoral research fellowship at the University of Kent at Canterbury (UK), he joined the Center for Nonlinear Dynamics in Economics and Finance (CeNDEF) group of the Faculty of Economics and Econometrics, where he is now an associate professor.

Comparing density forecasts

If we would like to predict a currently unknown random variable, such as tomorrow's closing value of the AEX index, based on currently available information, there are many different ways to do this. One way would be to predict the mean of the future random variable, given the information available to us (a point predictor). Another way would be to give an interval in which we think it is likely (say, at a confidence level of 95%) that the unknown future variable will fall (an interval predictor). Here we will be concerned with the most detailed type of forecast, namely estimates of the entire conditional distribution of the future random variable, given the available information. Typically this distributional forecast is given in terms of a probability density function, hence the name density forecast. Naturally, we would prefer an accurate forecast over an inaccurate one, so it is natural to develop methods that compare the accuracy of competing density forecasts.

To fix the notation, consider a univariate time series process $\{Y_t\}$, $t \in \mathbb{Z}$. For simplicity I consider one-step-ahead forecasts only. The density forecast for Y_{t+1}, made at time t, is based on the information F_t available at time t. Besides the past observations Y_s, $s \le t$, the information set may also contain exogenous variables known at time t. A one-step-ahead density forecast is a mapping from the information set to a probability density function (pdf) for Y_{t+1}, i.e. given F_t, it gives a density forecast for Y_{t+1} in the form of a pdf, denoted here by $f_{t,t+1}(y)$.

Ideally the pdf $f_{t,t+1}(y)$ is a good approximation to the true conditional pdf, $g_{t,t+1}(y)$ of Y_{t+1} given F_t. Clearly it is of interest to asses the quality of density forecasts. If we have a density forecast we are often interested in its quality relative to the true conditional density. This leads to the area of *goodness-of-fit tests*. The issue of how to select one of several alternative density forecasts that at our disposal leads to *forecast selection procedures*.

There is a branch of econometric literature that builds on density forecast evaluation along the lines proposed by Diebold *et al*. (1998). The idea is to transform the predictive densities into a sequence of PITs, defined as

$$U_{t+1} = F_{t,t+1}(Y_{t+1}),$$

where $F_{t,t+1}(y) = \int_{-\infty}^{y} f_{t,t+1}(s)ds$, the cumulative distribution function (CDF) of the density forecast made at time t. In the ideal case that the den-

Forecast	density forecast for Y_{t+1} when $\mu_t \sim N(0, 1)$, determined exogenously, is part of F_t
I	$N(\mu_t, 1)$
II	$N(0, 2)$
III	$(N(\mu_t, 1) + N(\mu_t + \tau_t, 1))/2$, where $\tau_i = \pm 1$, each with probability 1/2.
IV	$N(\mu_t + \delta_t, \sigma_t^2)$, where $(\delta_t, \sigma_t^2) = (1/2,1)$, (-1/2,1) or (0, 1.69), each with probability 1/3.

Table 1: Competing density forecasts for Y_{t+1}, which is $N(\mu_t, 1)$ distributed, where $\mu_t \sim N(0, 1)$ is part of the information set F_t.

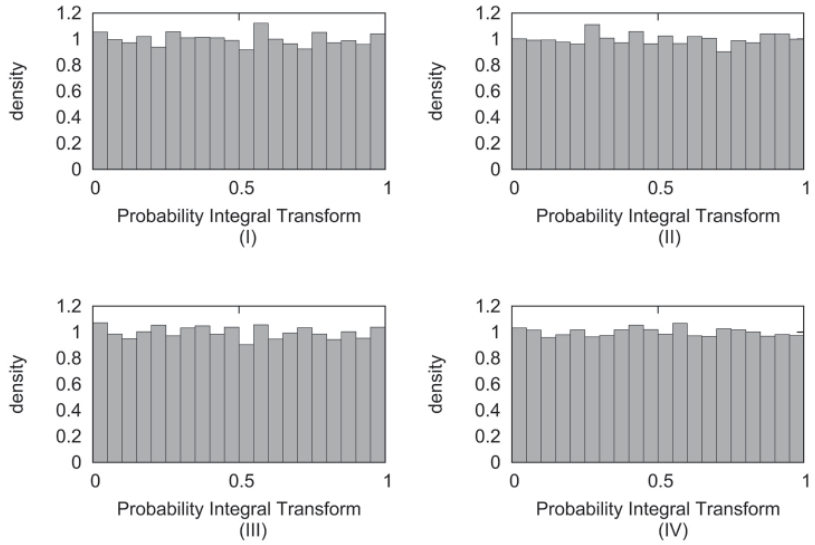

Figure 1: Uniform histograms PIT for density forecast I up to IV.

sity forecast is correct, a standard result in probability theory states that the sequence of PITs consists of independent uniform random variables on the interval [0, 1]. This observation has led to a methodology where one assesses the accuracy of density forecasts by investigating the sequence of PITs. Formal tests for uniformity and serial independence applied to observed PIT sequences are then used to judge the accuracy of density forecasts.

As illustrated by an example by Hamill (2001), however, the uniformity and independence of the PIT sequence is only a necessary, and not a sufficient, condition. In other words, there are density forecasts that are not ideal, but still have a sequence of PITs that are uniformly distributed on [0, 1] and serially independent. Gneiting *et al*. (2007) give a number of additional examples. Following these authors I here would like to consider the following example. The data Y_{t+1} are independently drawn from $N(\mu_t, 1)$, where μ_t is an exogenous $N(0, 1)$ variable (generated by nature), which is part of the information F_t, available at at time t. Table 1 shows four competing density forecasts. Density forecast I corresponds exactly to the conditional density of Y_{t+1} given μ_t, hence it is the ideal forecast. Forecast II simply only uses the marginal density of $\{Y_t\}$, and could be called the historical forecast. It is like predicting the weather for++ tomorrow on the basis of the weather on the same calendar day in the past 100 years. Forecast III involves an irrelevant variable τ_t, and hence is called an unfocused density forecast. Forecast IV uses three incorrect density forecasts, which each are chosen with equal probability.

This example was constructed to illustrate that PITs can be independent (they are so by construction here) and have a textbook uniform distribution, even if they are far from ideal. Figure 1 shows the estimated density of the respective PITs for each of these forecasts, based on 10,000 simulated values of Y_t. The histograms are practically flat. In fact, for forecasts I – III, it can be shown analytically that the PITs are distributed uniformly on the unit interval. It can be shown that the distribution of the PITs for forecast IV, which corresponds with the example given by Hamill (2001), has small deviations from uniformity, but this is not visible in Figure 1 due to the accuracy of the histogram.

"It is like predicting the weather for tomorrow on the basis of the weather on the same calendar day in the past 100 years"

My aim is to show that an approach using scoring rules automatically resolve the issues associated with PITs, at least when we want to compare competing pairs of density forecasts. This is illustrated numerically using the fore-

casts from the above example. An additional theoretical advantage to a score-based approach is that it is easily extended to a multivariate setting.

A popular scoring rule for density forecasts is based on the Kullback-Leibler information criterion (KLIC). Let Y have density $g(y)$ The KLIC for a density $f(y)$ relative to the true density $g(y)$ is defined as

$$\mathrm{KLIC}(f,g) = \mathbb{E}\left(\ln\left(\frac{g(Y)}{f(Y)}\right)\right) = \int_{-\infty}^{\infty} g(y)\ln\left(\frac{g(Y)}{f(Y)}\right)dy$$

The KLIC is a divergence measure between the densities f and g, which means that it is nonnegative, and zero only if f and g coincide. Of course, in practice the true density g is not known, but if we subtract two KLICs, the unknown density g drops out: $\mathrm{KLIC}(f^1, g) - \mathrm{KLIC}(f^2, g) = \mathbb{E}(\ln f^1(Y) - \ln f^2(Y))$. This motivates the use of the logarithmic scoring rule for comparing competing density forecasts.

Tests for equal predictive ability

Suppose we would like to measure, in the KLIC sense, which of two competing density forecasts is closer to the ideal density forecast, then we can use the testing methodology proposed by Giacomini and White (2006). The idea is to

$$T_n = \sqrt{n}\frac{\bar{d}}{\sigma_n},$$

introduce a scoring rule for the competing forecasts, and then test the hypothesis that they are performing equally well. The reason for using the logarithmic scoring rule here is that, as argued above, the ideal forecast should, on average, receive a larger logarithmic score than any deviating forecast.

The log-likelihood score associated with a forecast density $f_{t,t+1}(x)$ is

$$S_t = \ln f_{t,t+1}(X_{t+1})$$

For two competing density forecasts, say f_1 and f_2, we define the score difference as

$$d_t = \ln(f_1/f_2) = \ln f^1_{t,t+1}(X_{t+1}) - \ln f^2_{t,t+1}(X_{t+1})$$

Consider the null hypothesis is that the models receive the same average score:

$$H_0 : E[d_t] = 0.$$

The null hypothesis can be tested using the standardized sample mean of the scores where is σ_n^2 a heteroskedasticity and covariance robust estimator of the asymptotic variance of $\bar{d}$

Forecast	I	II	III	IV
$E[S_t]$	-1.4189	-1.7665	-15.304	-1.5298

Table 2: Average log-likelihood score for the four models.

(a HAC estimator). Under the null hypothesis of equal predictive ability T_n is asymptotically standard normally distributed. If one of the competing models is outperforming the other, then a two-tailed test ideally rejects, and the sign of the test statistic indicates which of the two models is receiving statistically significantly higher scores, on average. This is how such a test would be used in practice. In the simulations below I report size-power plots for the one-sided test, in order to keep track of the rejection rates corresponding with each of the tails of the distribution.

A small simulation study

In this section I investigate the behaviour of the test for equal predictive ability by applying it repeatedly to the example forecasts given in Table 1. The actual data are generated in correspondence with the ideal forecast: $Y_{t+1} \sim \mu_t$, where $\mu_t \sim$ N(0, 1) is drawn by nature, and known at time t.

For later reference I calculated numerical values of the true average scores for the four competing models. The results are shown in Table 2. From this table we can observe against which alternative the test should reject the null hypothesis, ideally. For instance, if we compare forecast I with forecast II, the test should reject the null of equal predictive ability in favour of forecast I, since forecast I actually has a higher average score than forecast II. In general, the test should ideally reject the null against the

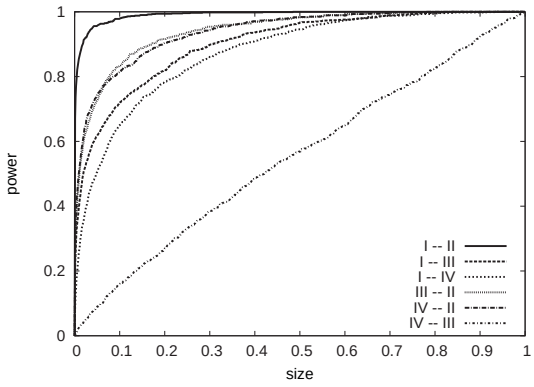

Figure 2: Size-power plots for the test of equal expected log-scores for each of the six pairs of processes. The lines are labeled by the two density forecasts that were being compared. The power plotted is that of rejecting the null of equal predictive ability against the model being mentioned first receiving better scores, on average. Sample size n = 50.

alternative that the model with higher average score in Table 2 performs better. To make the interpretation of the size-power graphs easier, for each pair of density forecasts tested against each other, I report the power in terms of rejections of the null hypothesis against this 'correct' alternative.

Figure 2 shows the size-power plots of all pairs of density forecasts, based on 1000 replications,for a sample size $n = 50$. The horizontal axis represents the nominal size, while the curves plotted correspond with the power (fraction of rejections) observed among the 1000 replications, for each particular nominal size. For all pairs of competing density forecasts involving the ideal density forecast (I) the test clearly has power against the null hypothesis, in favour of the alternative that forecast I performs better. Likewise, the test has power against the 'best' density forecasts involving the other pairs, except for the pair III–IV, for which the power almost coincides with the nominal size (i.e. a size-power plot along the diagonal). The reason is that the average logarithmic scores are so close for these two forecasts (see Table 2) that much larger sample sizes are required to detect the difference in performance. Additional simulations (not shown) indicate that the sample size should be of the order of 50,000 to obtain considerable power for this pair of density forecasts.

Summary

There are various ways to evaluate density forecasts. Among the most popular methods are those based on the sequence of probability integral (PIT) transforms, and score-based methods. In this paper I have illustrated some of the limitations of PIT-based methods by means of examples showing that for density forecasts that are far from ideal, the sequences of PITs can have the same properties as those for the ideal density forecast. Subsequently I have shown that score-based methods can distinguish between the predictive ability of these competing density forecasts.

References

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

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, **74**, 1545–1578.

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

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