Evaluating density forecasts: a comment

Tsyplakov, Alexander

Novosibirsk State University, Economics Department

30 May 2011

Evaluating Density Forecasts: A Comment

Alexander Tsyplakov
Department of Economics, Novosibirsk State University

May 30, 2011

This is a comment on Mitchell and Wallis (2011) which in turn is a critical reaction to Gneiting et al. (2007). The comment discusses the notion of forecast calibration, the advantage of using scoring rules, the "sharpness" principle and a general approach to testing calibration. The aim is to show how a more general and explicitly stated framework for evaluation of probabilistic forecasts can provide further insights.

1 What is "forecast calibration"?

Both Gneiting et al. (2007) (hereafter GBR) and Mitchell and Wallis (2011) (hereafter MW) examine various important aspects of calibration, but they do not propose a consistent framework and do not give sufficiently universal definitions of "calibration" and "ideal forecast". This leads to vague argumentation subject to confusion. So I start by clearing up the notion of calibration.

The general idea is that a well-calibrated probabilistic forecast must coincide with a suitable conditional distribution. The classical definition of calibration for dichotomous 0/1 outcomes requires that if π is a forecast of the probability of $X = 1$ then $P(X = 1|\pi) = \pi$. Predicted frequencies must coincide with actual frequencies conditionally on the information contained in the forecast itself.[1] More generally, let F be a probabilistic forecast of X stated in the form of a cumulative distribution function. It is calibrated if

$$F(x) = G(x|F),$$

where $G(x|F)$ is the conditional cumulative distribution function of X given F. I call this type of forecast calibration *auto-calibration*.

Now suppose that forecasting is based on some available information Ω. Then F is a function of Ω: $F = F(\cdot;\Omega)$. Conditioning on Ω allows to strengthen the definition of calibration. The forecast $F = F(\cdot;\Omega)$ is *calibrated with respect to* Ω if

$$F(x) = G(x|\Omega).$$

Such a forecast can be called the ideal one (of all the forecasts based on Ω).

Obviously, if a forecast is calibrated with respect to Ω then it is auto-calibrated while the converse is not true in general. The conditional cumulative distribution function $G(x|\Omega)$ need not coincide with $G(x|F)$, because a non-ideal forecast utilize available information only partially.

[1] E.g. Lichtenstein et al. (1982), p. 307: "Formally, a judge is calibrated if, over the long run, for all propositions assigned a given probability, the proportion true equals the probability assigned."

Note that these definitions of calibration are made in terms of theoretical moments. GBR give their definitions of calibration in terms of empirical moments (for a potentially infinite sequence of forecasts).

2 Can independence and uniformity of PIT values be called complete calibration?

One point of MW's criticism is that GBR are concerned with forecasting of white noise instead of more relevant forecasting of dependent time series. However, what is really important is that GBR fail to emphasize the conditional nature of calibration.[2] MW define "complete calibration" as independence and uniformity of the probability integral transform (PIT) values of a time series. Indeed, it is very true that the available history of a time series (including PIT values) is an important information for judging calibration of forecasts of this series, but in general the view is unnecessarily narrow. The notion of forecast calibration is not about time series properties, it is about conditioning.

We first need to connect the definition of calibration given above with PIT. If the conditional distribution of X given Ω is absolutely continuous then a probabilistic forecast $F = F(\cdot;\Omega)$ is calibrated with respect to Ω if and only if

$$P|\Omega \sim U[0,1],$$

where $P = F(X|\Omega)$ is PIT of X based on the forecast F.

Now consider a sequence $F_t(x)$ of probabilistic forecasts of a univariate time series X_t based on its history $\Omega_{t-1} = X_1,\ldots,X_{t-1}$, $t = 1,2,\ldots$. (For $t = 1$ the forecast is unconditional). Define the corresponding PIT values as follows:

$$P_t = F_t(X_t).$$

This provides for each t a one-to-one transformation between $P_1,\ldots,P_t$ and $\Omega_t = X_1,\ldots,X_t$ whenever all the distributions are absolutely continuous. The series of forecasts is well-calibrated if and only if

$$P_t|\Omega_{t-1} \sim U[0,1]$$

or equivalently

$$P_t|P_1,\ldots,P_{t-1} \sim U[0,1].$$

MW concentrate on an equivalent property. The series of forecasts is well-calibrated if and only if all P_t are distributed as $U[0,1]$ and are jointly independent (because under assumption of calibration what we have here is the Rosenblatt's transformation $P_t = G(x_t|\Omega_{t-1})$; see Rosenblatt (1952)). The problem with PIT-based approach is that it is insufficiently universal; it does not generalize to other forecasting situations. Using MW's own words, it is important "to pay attention to the information set on which a forecast is based, its content and its timing".

For example, independence and uniformity of a sequence of PIT values is not sufficient for calibration in the case when forecasts are based on information other than the history of X_t itself. In particular, consider predicting X_t from $X_1,\ldots,X_{t-1}$ and the history of some other series $Z_1,\ldots,Z_{t-1}$. Then calibration with respect to $\Omega_{t-1} = (X_1,\ldots,X_{t-1},Z_1,\ldots,Z_{t-1})$ is equivalent to

$$P_t|\Omega_{t-1} \sim U[0,1],$$

[2] Clements and Taylor (2003), p. 446: "Evaluating probability forecasts by calibration ignores the conditional aspect".

where $P_t = F_t(X_t)$. For a sequence of forecasts F_t these conditions imply uniformity and independence of P_t, but in general the later is a weaker condition. The fact is fully recognized by MW, but they do not adjust their definition of "complete calibration" accordingly.

Another example is multi-step forecasting. An h-step-ahead forecast of X_t would be based only on $X_1, \dots X_{t-h}$. In such a situation the PIT series would in general be dependent.

3 Testing vs. scoring rules

Suppose that we need to choose between several probabilistic forecasts. When can we state that one forecast is better than another? This is a situation of decision making under risk. Let $u(y,a)$ be a utility function depending on an outcome x and an action a. (Equivalently, we can consider a loss function). Suppose $a(F)$ is the best action given a probabilistic forecast F, that is,

$$a(F) \in \arg\max_a \mathsf{E}[u(X,a)] \text{ where } X \sim F.$$

One can say that the forecast F_1 is (non-strictly) better than forecast F_2 if it provides at least as large expected utility:

$$\mathsf{E}u(X,a(F_1)) \geq \mathsf{E}u(X,a(F_2)).$$

This is closely connected with the notion of a proper scoring rule (see Gneiting and Raftery (2007) for a comprehensive review). Define a scoring rule S as the utility of outcome x under action $a(F)$:

$$S(F,x) = u(x,a(F)).$$

This utility-based scoring rule would be a proper one since

$$S(F_1,F_1) = \mathsf{E}[u(X,a(F_1)] \geq \mathsf{E}[u(X,a(F_2)] = S(F_2,F_1) \quad \text{for } X \sim F_1.$$

Here

$$S(F_2,F_1) = \mathsf{E}[S(F_2,X)] \quad \text{for } X \sim F_1$$

is the expected score of forecast F_2 under the assumption that X is distributed as F_1.

Because of the link with utility maximization it is logical to base the theory of evaluating probabilistic forecasts on proper scoring rules. Then the choice between competing probabilistic forecasts is made on the basis of their expected scores. Expected scores can be estimated by the corresponding empirical average scores. Thus, all one has to do in order to choose from a set of competing probabilistic forecasts is to select a forecast with the highest average score according to a suitable proper scoring rule. For a fixed set of forecasting models there is no convincing reason to apply statistical testing in order to choose one model and reject others.

MW argue that non-ideal forecasts can be distinguished from the ideal one by means of testing forecasting models against each other. However, proper scoring rules are no less suitable for the task of detecting the ideal forecast. It can be shown that the well-calibrated forecast is characterized by the maximal expected score for any proper scoring rule.[3] Indeed, suppose that $G = G(\cdot|\Omega)$ is such a forecast and F is some other forecast based on Ω. Then

$$\mathsf{E}[S(G,X)|\Omega] = S(G,G) \geq S(F,G) = \mathsf{E}[S(F,X)|\Omega]$$

[3]Diebold et al. (1998), p. 866: "…If a forecast coincides with the true data generating process, then it will be preferred by all forecast users, regardless of loss function." See also Granger and Pesaran (2000).

and

$$\mathsf{E}[S(G,X)] \geq \mathsf{E}[S(F,X)].$$

This gives a reason for calling a forecast coinciding with the conditional distribution $G(\cdot|\Omega)$ ideal.

The usefulness of using scoring rules for detecting the ideal forecast is illustrated by Table IV in MW's paper. They use the logarithmic scoring rule $S(F,x) = \log f(x)$, where $f(x) = F'(x)$ is the probability density function corresponding to c.d.f. F. The logarithmic rule is known to be (strictly) proper. See also Table 5 in GBR's paper which compare the values of the logarithmic score and the continuous ranked probability score (another useful strictly proper scoring rule) of competing forecasts.

Scoring rules are straightforward and easy to deal with. Testing of one model against another is a less trivial operation while rarely providing much beyond the comparison of average scores. (An example of an understandable use is testing whether the average score of the best model is significantly greater than the average score of a competing model).

I do not want to say that statistical testing is redundant in the context of probabilistic forecasting. Most important, various kinds of diagnostic tests can help to develop a forecasting model which is well enough calibrated by showing directions in which models can be improved in order to increase expected score. However, discovering mis-calibration by diagnostic testing is not always a straightforward task. The following example is somewhat artificial and is not related to real forecasting problems, but it is suggesting.

Consider a series $X_t \sim N(0,1)$, $t = 1,2,\ldots$. To obtain X_t from X_{t-1} (1) transform X_t to $U[0,1]$, (2) transform $U[0,1]$ to $U[0,1]$ by a highly nonlinear "tangling" transformation, (3) transform $U[0,1]$ back to $N(0,1)$, and finally (4) add an independent normal error and scale to unit variance. For example,

$$\mu_t = \Phi^{-1}(\{K|2\Phi(X_{t-1})-1|\})\sqrt{1-\lambda},$$
$$X_t|X_1,\ldots,X_{t-1} \sim N(\mu_t,\lambda).$$

Here $\Phi(\cdot)$ is the standard normal c.d.f., $\{\cdot\}$ is the fractional part function, K is a large enough integer and $\lambda \in (0,1)$. One can see that X_t is a white noise series and is exactly distributed as $N(0,1)$ for each t whenever $X_1 \sim N(0,1)$ (though it is dependent and thus could not be strictly called a Gaussian white noise series). For large values of K ordinary PIT-based tests would not detect any dependence. Consequently, on the basis of PIT one would choose $N(0,1)$ as a good forecasting distribution. However, $N(\mu_t,\lambda)$ is a dramatically better forecasting distribution for a small λ.

It can be concluded from the example that sometimes the direction of mis-calibration is not obvious. Popular PIT-based diagnostic tests can fail to detect departures from perfect calibration. At the same time comparison of forecasts by their average scores with proper scoring rules is feasible and works as expected irrespective of the kind of mis-calibration. (A reservation, however, should be made here. If there exists an alternative forecasting model then it can be utilized for the task of testing calibration. An idea of such a test is discussed below.)

Actually MW's paper is not an exception in its attention to statistical testing of forecasting models. A "test-test-test" bias is typical for econometric forecasting literature in general. Cf. Corradi and Swanson (2006) which is a survey of density forecast evaluation almost entirely devoted to various test procedures.

4 Maximizing sharpness subject to calibration

MW are critical of the conjecture that the problem of finding a good forecast can be viewed as the problem of maximizing sharpness subject to calibration which was stated by GBR. However, it can be shown that the conjecture is actually true.

If F is an auto-calibrated forecast (calibrated with respect to itself), that is, $F = G(\cdot|F)$ and S is some scoring rule then

$$\mathsf{E}[S(F,X)|F] = S(F,F).$$

Taking unconditional expectations of both sides gives

$$\mathsf{E}[S(F,X)] = \mathsf{E}[S(F,F)].$$

For a proper scoring rule $S(F,F)$ can be viewed as a measure of sharpness[4] of a probabilistic forecast F. Hence the expected score for an auto-calibrated forecast equals its expected sharpness. This means that auto-calibrated forecasts can be compared on the basis of the levels of their expected sharpness. The ideal forecast is the sharpest of all auto-calibrated forecasts, because it is characterized by the maximal expected score.

This fact is seen from the classical partitioning of the Brier score for dichotomous outcomes into the sum of sharpness and calibration ("validity") terms developed in Sanders (1963). For an auto-calibrated forecast the calibration term is zero. Thus, maximizing the expected sharpness among auto-calibrated forecasts is equivalent to maximizing the expected score of a proper scoring rule. Bröcker (2009) extended the decomposition to the case of an arbitrary discrete distribution and an arbitrary proper scoring rule. It can further be shown that the same decomposition holds even more generally and applies also to continuous distributions.

Given some proper scoring rule S define the divergence[5] between distributions F_1 and F_2 as

$$d(F_1,F_2) = S(F_1,F_2) - S(F_2,F_2)$$

and denote $G_F = G(\cdot|F)$. Then for the expected score we have $\mathsf{E}[S(F,X)] = \mathsf{E}[\mathsf{E}[S(F,X)|F]] = \mathsf{E}[S(F,G_F)]$ which can be decomposed as follows:

$$\mathsf{E}[S(F,X)] = \mathsf{E}[S(G_F,G_F)] - \mathsf{E}[d(F,G_F)].$$

The first term can be interpreted as the expected sharpness of G_F, which is the "recalibrated" version of forecast F, while the second term relates to the divergence between F and G_F (that is, it is a measure of mis-calibration of F with respect to information contained in itself).

Although the sharpness conjecture proves to be correct, MW are right in pointing out that the criterion of sharpness per se seems somewhat excessive. First, obviously a proper scoring rule already combines sharpness and calibration in a balanced manner. Second, the principle of maximizing sharpness subject to calibration is difficult to apply in practice, because achieving perfect auto-calibration of a forecast can be a difficult (and even infeasible) task. However, the sharpness principle provides a useful insight into the essence of probabilistic forecasting.

[4]For a proper scoring rule $S(F,F)$ is a convex function of F. Thus, according to DeGroot (1962) a concave function $-S(F,F)$ can be viewed as a measure of uncertainty of probability distribution F. For the logarithmic scoring rule $-S(F,F)$ is the familiar Shannon's entropy measure.

[5]The divergence $d(F_1,F_2)$ is non-negative as long as S is proper and it is zero when the two distributions coincide.

5 A broader approach to calibration testing

MW rightly note that calibration can be evaluated by testing orthogonality conditions analogous to those known from point forecasting. It is important that their idea can be further generalized. This broader approach allows to design various kinds of diagnostic tests for forecast calibration. Most of the tests and criteria of calibration/efficiency developed in the literature can be shown to fall within this approach.

Consider a function $m = m(x, \Omega)$ depending on an outcome x and forecasting information Ω. Define

$$\mu = \mu(\Omega) = \mathsf{E}[m(X, \Omega)] \qquad \text{for } X \sim F = F(\cdot; \Omega).$$

Conditional calibration of F with respect to Ω is equivalent to the following unconditional orthogonality:

$$\mathsf{E}[(m(X, \Omega) - \mu(\Omega))h(\Omega)] = 0$$

for any m and any function h. That is, μ is an unbiased point forecast for m and the forecast error is uncorrelated with any function h of the available information Ω.

Let me provide just a few illustrations.

When

$$m = 1(F(x) \le p), \quad \mu = p,$$

we obtain an extended version of the "probabilistic calibration" of GBR. The orthogonality condition $\mathsf{E}[(m - p)h] = 0$ for each $p \in [0, 1]$ and any function h is another equivalent characterization of perfect calibration with respect to Ω. Note that unlike the formulation by GBR this one does not make a direct use of the ideal forecast $G(\cdot; \Omega)$ and thus is more suitable for construction of tests.

Similarly, an extended version of the "marginal calibration" of GBR is produced by setting

$$m = 1(x \le x_0), \quad \mu = F(x_0).$$

"Tests of efficiency" proposed by MW are also a special case of this approach with $m = \Phi^{-1}(F(x))$ and $\mu = 0$.

One more example parallels KLIC-based tests discussed by MW. The idea is to test calibration of one model against another one. Suppose that we want to test whether $F_1 = F_1(\cdot; \Omega)$ is well calibrated and $F_2 = F_2(\cdot; \Omega)$ is an alternative forecast. Let $m = S(F_2, X) - S(F_1, X)$ for some proper scoring rule S. For a forecast F_1 which is calibrated with respect to Ω

$$\mu = \mathsf{E}[m|\Omega] = S(F_2, F_1) - S(F_1, F_1).$$

So we can test calibration of F_1 by testing that

$$\mathsf{E}[(S(F_2, X) - S(F_1, X)) - (S(F_2, F_1) - S(F_1, F_1))] = 0.$$

The test based on this moment condition would have power against an alternative that F_2 is calibrated with respect to Ω, because then

$$\mathsf{E}[m - \mu] = \mathsf{E}[(S(F_2, F_2) - S(F_1, F_2)] + \mathsf{E}[(S(F_1, F_1) - S(F_2, F_1)] > 0,$$

if the scoring rule is strictly proper.

For the case of the logarithmic score and two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ the moment condition reduces to

$$\mathsf{E}\Big[\frac{1}{2\sigma_2^2}\left(\sigma_1^2 + (\mu_1 - \mu_2)^2 - (X - \mu_2)^2\right) - \frac{1}{2\sigma_1^2}\left(\sigma_1^2 - (X - \mu_1)^2\right)\Big] = 0.$$

A pair of reciprocal tests (F_1 against F_2 and F_2 against F_1) can help to judge possible gains from combining two forecasts.

References

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores, *Quarterly Journal of the Royal Meteorological Society* **135**(643): 1512–1519.

Clements, M. P. and Taylor, N. (2003). Evaluating interval forecasts of high-frequency financial data, *Journal of Applied Econometrics* **18**(4): 445–456.

Corradi, V. and Swanson, N. R. (2006). Predictive density evaluation, *in* C. W. J. Granger, G. Elliott and A. Timmermann (eds), *Handbook of Economic Forecasting*, Vol. 1, North-Holland, Amsterdam, chapter 5, pp. 197–286.

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments, *The Annals of Mathematical Statistics* **33**(2): 404–419.

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

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.

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

Granger, C. W. J. and Pesaran, M. H. (2000). A decision-theoretic approach to forecast evaluation, *in* W.-S. Chan, W. K. Li and H. Tong (eds), *Statistics and Finance: An Interface*, Imperial College Press.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980, *in* D. Kahneman, P. Slovic and A. Tversky (eds), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, UK, pp. 306–334.

Mitchell, J. and Wallis, K. F. (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness, *Journal of Applied Econometrics,* **n/a**: ??–?? forthcoming.

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

Sanders, F. (1963). On subjective probability forecasting, *Journal of Applied Meteorology* **2**: 191–201.