Second International Conference in Memory of Carlo Giannini

Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness

Kenneth F. Wallis

Emeritus Professor of Econometrics, University of Warwick
http://www.warwick.ac.uk/go/kfwallis

Joint work with James Mitchell (NIESR, London)

Banca d'Italia *19-20 January 2010*

Summary

This paper reviews current density forecast evaluation procedures, in the light of a recent proposal that these be augmented by an assessment of “sharpness”.

This proposal is motivated by an example in which some standard evaluation procedures based on probability integral transforms cannot distinguish between the ideal forecast and several competing forecasts. From a time-series forecasting perspective, however, this example has some unrealistic features, and so does not give a strong case that existing calibration procedures are inadequate in practice.

We present a more realistic example in which several competing forecasts may nevertheless satisfy probabilistic calibration. We show how relevant statistical methods, including information-based methods, provide the required discrimination between competing forecasts and the ideal forecast. We propose an extension to these methods to test density forecast efficiency. We conclude that there is no need for a subsidiary criterion of sharpness, which in practice may be misleading.

Contents of the paper

1. Introduction
2. The statistical framework
3. Gneiting, Balabdaoui and Raftery (GBR)’s example
4. Forecasting an autoregressive process
5. Conclusion

The statistical framework

2.1. Calibration

Dawid's (1984) prequential principle: assessments should be based on the forecast-observation pairs only.

Consider forecasts given as predictive CDFs F_t of outcomes $X_t,\ t = 1, 2, \ldots$.

The standard assessment tool is the sequence of probability integral transforms (PITs) $p_t = F_t(x_t),\ t = 1, 2, \ldots$, of observed outcomes in the forecast CDFs.

If F_t coincides with G_t, the correct or "ideal" forecast, the PITs are iid*U*[0,1].

In practice, does such a sequence "look like" a random sample from *U*[0,1]? Note that to check this out we don't need to know G_t, actually or hypothetically.

We refer to the two-component condition – indep and *U* – as ***complete calibration***.

GBR refer to the condition of uniform PITs as ***probabilistic calibration***.

Attention to the ***information set*** on which a forecast is based is usually needed.

Let Ω_t denote the set of all relevant information available at the forecast origin.

Then the "ideal" forecast or correct conditional distribution is written $G_t\left(x_t \middle| \Omega_t\right)$:

in economics this is often called the "rational" or "fully efficient" forecast.

A practical forecast $F_t\left(x_t \middle| W_t\right)$ has different information, functional form, ...

Denote the correct distribution conditional on W_t as $G_t^*\left(x_t \middle| W_t\right)$.

Then if $F_t\left(x_t \middle| W_t\right)$ coincides with $G_t^*\left(x_t \middle| W_t\right)$ it has uniform PITs: F_t satisfies

probabilistic calibration, but not necessarily complete calibration.

Analogously, a point forecast may have zero-mean errors and so be unbiased, but

it may not necessarily be efficient in a minimum MSE sense.

2.2. Statistical tests of calibration

Diagnostic checks can be based on the PITs or their inverse normal transforms $z_t = \Phi^{-1}(p_t)$ (Smith, 1985).

If the PITs are iid*U*[0,1] then the z_ts are iid*N*[0,1]; again we don't need to know G_t.

Goodness-of-fit tests commonly used include chi-squared tests and, for uniformity, Kolmogorov-Smirnov and Anderson-Darling tests, and for normality, the Doornik-Hansen test.

(These are based on random sampling assumptions.)

Tests of independence include, for the p_ts, the Ljung-Box test, and for the z_ts, the likelihood ratio test of Berkowitz (2001).

2.3. Scoring rules, distance measures and sharpness

For forecast density f_{jt} the logarithmic score is $\log S_j(x_t) = \log f_{jt}(x_t)$.

For two forecasts, the difference in log scores is the log Bayes factor.

If one of the forecasts is the "ideal" g_t, the expected difference in log scores is the Kullback-Leibler information criterion or distance measure

$$\text{KLIC}_t = E_g\left\{\log g_t(x_t) - \log f_{jt}(x_t)\right\} = E_g\left\{d_t(x_t)\right\}.$$

KLIC-based tests for density forecast evaluation replace E by a sample average and x by p or z.

To test equal predictive accuracy of two forecasts f_{jt} and f_{kt}, their KLIC difference does not involve g_t, and an LR test is based on the sample average of $\log f_{jt}(x_t) - \log f_{kt}(x_t)$. This is a simple example of a Giacomini-White test.

We propose an efficiency test based on the regression of the density forecast error $d_t(x_t)$ on elements of W_t.

Some simple relations for normal density forecasts

The expected log score of the correct conditional density is a simple function of its forecast variance (sharpness/concentration/precision):

$$E_g\left\{\log g(x)\right\} = -\tfrac{1}{2}\log\left(2\pi\sigma_g^2\right) - \tfrac{1}{2} \ .$$

With a competing forecast $f(x)$ we obtain the KLIC as

$$E_g\left\{\log g(x) - \log f(x)\right\} = -\tfrac{1}{2} - \tfrac{1}{2}\log\left(\sigma_g^2 \,/\, \sigma_f^2\right) + \tfrac{1}{2}\frac{\sigma_g^2}{\sigma_f^2} + \frac{\left(\mu_g - \mu_f\right)^2}{2\sigma_f^2} \ .$$

This expression has a minimum at zero. A positive KLIC may be the result of departures in mean and/or variance in either direction, and a KLIC-based test is not constructive. PIT histograms can indicate the direction of such departures.

Gneiting, Balabdaoui and Raftery's example

Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society* B, vol.69 (2007), pp.243-268.

Table 1. Scenario for the simulation study†

Forecaster	*F_t when nature picks $G_t = \mathcal{N}(\mu_t, 1)$ where $\mu_t \sim \mathcal{N}(0, 1)$*
Ideal	$\mathcal{N}(\mu_t, 1)$
Climatological	$\mathcal{N}(0, 2)$
Unfocused	$\frac{1}{2}\{\mathcal{N}(\mu_t, 1) + \mathcal{N}(\mu_t + \tau_t, 1)\}$ where $\tau_t = \pm 1$ with probability $\frac{1}{2}$ each
Hamill's	$\mathcal{N}(\mu_t + \delta_t, \sigma_t^2)$ where $(\delta_t, \sigma_t^2) = \left(\frac{1}{2}, 1\right), \left(-\frac{1}{2}, 1\right)$ or $\left(0, \frac{169}{100}\right)$ with probability $\frac{1}{3}$ each

†At times $t = 1, 2, \ldots, 10\,000$, nature picks a distribution G_t, and the forecaster chooses a probabilistic forecast F_t. The observations are independent random numbers x_t with distribution G_t. We write $\mathcal{N}(\mu, \sigma^2)$ for the normal distribution with mean μ and variance σ^2. The sequences $(\mu_t)_{t=1,2,\ldots}$, $(\tau_t)_{t=1,2,\ldots}$ and $(\delta_t, \sigma_t^2)_{t=1,2,\ldots}$ are independent identically distributed and independent of each other.

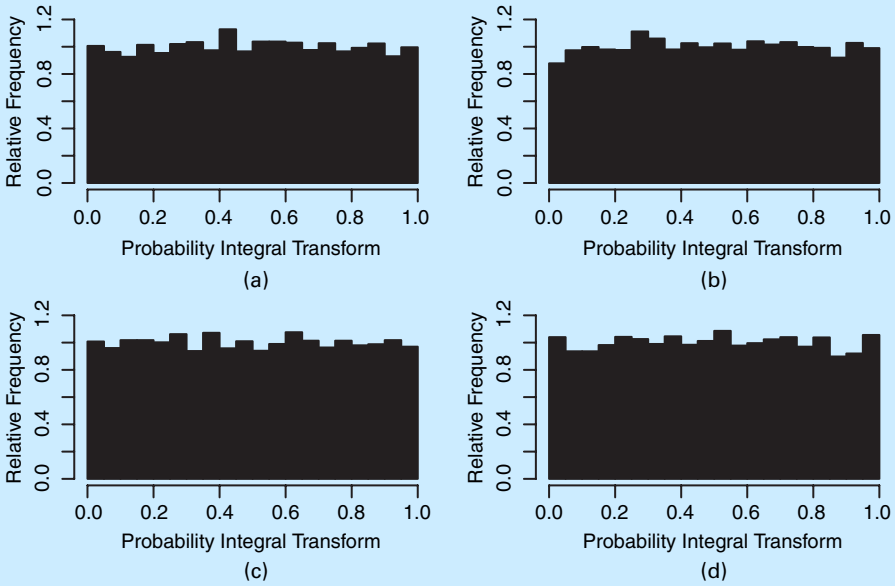


Fig. 1. PIT histograms for (a) the ideal forecaster, (b) the climatological forecaster, (c) the unfocused forecaster and (d) Hamill's forecaster

- all four PIT histograms are “essentially uniform” – a “disconcerting result”
- they cannot distinguish the ideal forecast from its competitors
- GBR propose maximising sharpness of the predictive distributions, subject to calibration
- to assess this they compare the average width of 50% and 90% prediction intervals and mean log scores across the four forecasts (ranked, not tested).

The first indistinguishable competitor is the unconditional forecaster (panel b). Its distribution is correct, but in typical time-series forecasting problems time dependence results in autocorrelation of the point forecast errors and density forecast PITs of an unconditional forecast, denying complete calibration.

However GBR are forecasting white noise (cf. Granger, 1983), which is scarcely a representative example in time-series forecasting.

Panels c and d are based on model mixtures or switching models:
the forecast issued in each period is one of two (c) or three (d) possible forecasts, (none of which have the correct distribution), chosen at random.

Contrast the forecast combination literature, since Bates and Granger (1969): multiple (point) forecasts are available in every period, and can be combined in various ways; Wallis (2005) considers density forecast combination.

Equally-weighted combinations of the above forecasts have non-uniform PITs.

Do our formal tests solve their “disconcerting result”?

- goodness-of-fit and autocorrelation tests: **no**
- KLIC-based tests: **yes**

Forecasting an autoregressive process

We consider the Gaussian second-order autoregressive DGP:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t, \;\; \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\rho_1 = \phi_1 / \left(1 - \phi_2\right), \;\; \rho_2 = \phi_1 \rho_1 + \phi_2, \;\; \sigma_\varepsilon^2 = \left(1 - \phi_1 \rho_1 - \phi_2 \rho_2\right) \sigma_y^2$$

Given observations y_{t-1} and y_{t-2} and knowledge of parameter values, the true conditional distribution or "ideal" forecast is

$$G_t = N\left(\phi_1 y_{t-1} + \phi_2 y_{t-2}, \sigma_\varepsilon^2\right).$$

Five competing density forecasts are constructed, as follows.

Climatological forecaster: $N\left(0,\sigma_y^2\right)$

AR1 forecaster: $N\left(\rho_1 y_{t-1},\sigma_1^2\right), \qquad \sigma_1^2=\left(1-\rho_1^2\right)\sigma_y^2$

AR2 (same, with data delay): $N\left(\rho_2 y_{t-2},\sigma_2^2\right), \qquad \sigma_2^2=\left(1-\rho_2^2\right)\sigma_y^2$

Combined forecast: $\frac{1}{2}N\left(\rho_1 y_{t-1},\sigma_1^2\right)+\frac{1}{2}N\left(\rho_2 y_{t-2},\sigma_2^2\right)$

Unfocused (GBR) forecaster: $0.5\left\{G_t+N\left(\phi_1 y_{t-1}+\phi_2 y_{t-2}+\tau_t,\sigma_\varepsilon^2\right)\right\},$

where τ_t is either 1 or –1, each with probability one-half.

The first three of these conditional distributions are correct with respect to their specific information sets, so we expect to find that they satisfy probabilistic calibration. The unfocused forecaster's biases are expected to be offsetting.

Table 1. Simulation design

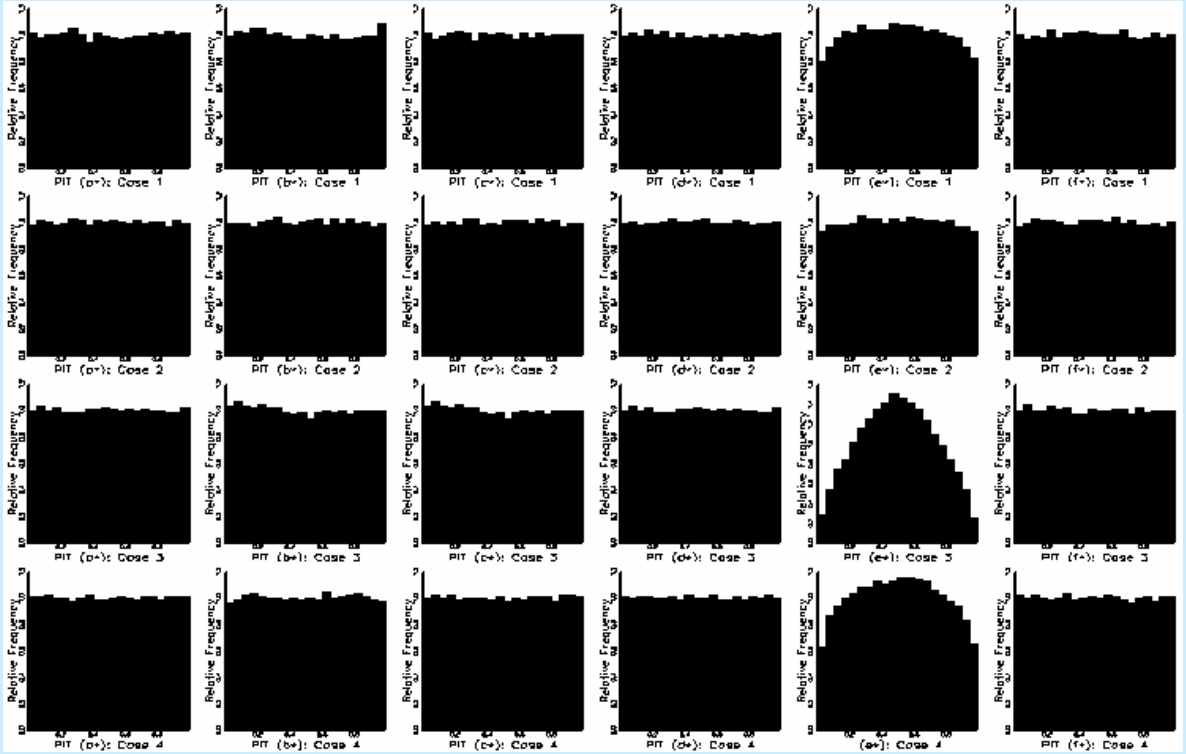
	Parameter		Autocorrelation	
	ϕ_1	ϕ_2	ρ_1	ρ_2
Case (1)	1.5	–0.6	0.94	0.80
Case (2)	0.15	0.2	0.19	0.23
Case (3)	0	0.95	0	0.95
Case (4)	–0.5	0.3	–0.71	0.66

Sample size 150; number of replications 500

Evaluation methods

- PIT histograms (Figure 2)
- goodness-of-fit tests; note performance under autocorrelation (Table 2)
  - Kolmogorov-Smirnov
  - Anderson-Darling
  - Doornik-Hansen test on z_ts
- performance of autocorrelation tests (Table 3)
  - Ljung-Box tests on PITs up to lag 4
  - Berkowitz 3 d.f. LR test on z_ts
- scoring rules and distance measures
  - performance of selection/ranking criteria (Table 4)
  - KLIC-based tests (Table 5)
- efficiency tests (Table 6), but note shortage of test regressors

Figure 2. PIT histograms



Rows: cases (1) to (4)

Columns: ideal, unconditional, AR1, AR2, combination, unfocused forecasters

Table 3. Tests of independence: error autocorrelns and rejection percents

	Case (1)			Case (2)			Case (3)*			Case (4)		
	$\rho_1(e)$	LB	Bk	$\rho_1(e)$	LB	Bk	$\rho_2(e)$	LB	Bk	$\rho_1(e)$	LB	Bk
Ideal	0	4.4	4.2	0	3.8	4.6	0	6.2	5.6	0	5.2	3.4
Climt	.94	100	100	.19	68	53	.95	100	99	–.71	100	100
AR(1)	.56	100	100	–.04	43	17	.95	100	99	.21	78	62
ARlag	.77	100	100	.15	24	30	0	6.2	5.6	–.35	99	97
Combo	.73	100	100	.06	16	14	.80	98	100	–.16	35	62
Unfocus	–.01	4.4	3.8	–.01	5.0	5.4	–.01	5.0	5.0	–.01	4.6	4.2

***in case (3)** $\rho_1(e) = 0$ for all forecasts except unfocus, where $\rho_1(e)$ is repeated

Table 2. Goodness-of-fit tests: rejection percentages at nominal 5% level

Forecast	Case (1)			Case (2)			Case (3)			Case (4)		
	KS	AD	DH	KS	AD	DH	KS	AD	DH	KS	AD	DH
Ideal	4.6	4.4	6.4	4.0	4.4	6.2	4.2	4.2	5.4	6.0	5.2	6.0
Climt	60	66	43	14	18	6.0	86	89	56	4.4	8.4	5.0
AR1	0.8	1.0	6.4	9.4	8.8	6.6	86	89	56	13	16	5.6
AR2	6.6	8.6	12	7.8	6.8	5.6	4.2	4.2	5.4	0.2	0	3.0
Combo	5.6	6.0	8.2	7.8	8.0	6.0	93	97	11	6.8	7.2	7.8
Unfocus	4.0	5.2	6.4	5.2	4.8	4.6	6.6	5.8	6.2	5.2	5.0	5.4

Monte Carlo standard error ≈ 1 under H_0

Table 4. Additional evaluation criteria

Forecast	**Case (1)**		**Case (2)**		**Case (3)**		**Case (4)**	
	KLIC	$-\log S$	KLIC	$-\log S$	KLIC	$-\log S$	KLIC	$-\log S$
Ideal	0	142	0	142	0	142	0	142
Climt	128	270	3.83	145	117	258	39.6	182
AR1	22	164	2.04	144	117	258	4.8	147
AR2	75	217	1.16	143	0	142	11.9	154
Combo	43	185	0.71	142	35	177	3.3	145
Unfocus	11	153	11.0	153	11	153	11.0	153

Table 5. KLIC-based tests against ideal forecaster: rejection percentages

Forecast	**Case (1)**	**Case (2)**	**Case (3)**	**Case (4)**
Climt	100	39	100	100
AR1	98	25	100	46
AR2	100	15	n.a.	93
Combo	100	10	100	55
Unfocus	87	87	87	90

Evaluation results

- PIT histograms: “essentially uniform”, except combo – “disconcerting”?
- g-o-f tests: similar conclusion; combo not very non-normal
  - error autocorrelation increases rejection rates above 5% ...
  - ... especially for climt (unconditional) forecaster
- indep tests: immediately distinguish ideal from all but unfocused forecast
- scoring rule: gives similar rankings across Cases (1) – (4) ...
  - ... except for combo, where in Cases (1), (3) optimal weights are far from the assumed equal weights
- KLIC-based test: again solves the “disconcerting result”
  - note that KLIC accounts for sharpness, at least in part
- efficiency test: power varies with amount of autocorrelation in the data

Conclusion

GBR's example is not a good guide to the adequacy or otherwise of existing evaluation methods because

- it has no time dimension, so the complete calibration criterion is irrelevant
- their competitor forecasts have no counterpart in the existing forecast combination literature
- but in any event, KLIC-based methods **can** discriminate between them.

Our example shows that, in a more realistic time-series forecasting setting,

- the usual complete calibration criterion (and, again, KLICs) can discriminate between the ideal forecast and other forecasts with incomplete information which nevertheless may have "essentially uniform" PIT histograms
- so there is no call for a subsidiary criterion of sharpness

In practical forecasting, a preference for the "sharpest" forecast may be a mistake.