"Optimal" combination of density forecasts

Stephen G. Hall and James Mitchell*
Imperial College, London and NIESR

November 5, 2004

Abstract

This paper brings together two important but hitherto largely unrelated areas of the forecasting literature, density forecasting and forecast combination. It proposes a simple data-driven approach to direct combination of density forecasts using "optimal" weights.

JEL Classification: C53; E37
Keywords: Density forecasts; combination

1 Introduction

Measures of uncertainty surrounding a "central tendency" (the point forecast) can enhance its usefulness; e.g. see Garratt et al. (2003). So called "density" forecasts are being used increasingly since they provide commentators with a full impression of the uncertainty associated with a forecast; see Tay & Wallis (2000) for a review. More formally, density forecasts of inflation, say, provide an estimate of the probability distribution of its possible future values.

It is well established that combining competing individual point forecasts of the same event can deliver more accurate forecasts, in the sense of a lower root mean squared error (RMSE); e.g. see Stock & Watson (2004). The success of combination follows from the fact that individual forecasts may be based on misspecified models, poor estimation or non-stationarities; e.g. see Hendry & Clements (2004).

This paper takes the natural next step of considering density forecast combination, to-date a relatively unexplored area. This brings together two important but hitherto largely unrelated areas of the forecasting literature in economics, density forecasting and

*Address for correspondence: James Mitchell, National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London, SW1P 3HE, U.K. Tel: +44 (0) 207 654 1926. Fax: +44 (0) 207 654 1900. E-Mail: j.mitchell@niesr.ac.uk. Thanks to Ray Barrell, Robert Metz, Kostas Mouratidis, Rebecca Riley and Martin Weale for helpful comments. Mitchell gratefully acknowledges financial support from the ESRC (Award Reference: RES-000-22-0610).

forecast combination.[1] We propose a simple data-driven approach to combine density forecasts directly using "optimal" weights.

How we measure the accuracy of forecasts is central to how we choose to combine them "optimally". Point forecasts are traditionally evaluated on the basis of their RMSE relative to the outturn. Then point forecasts can be optimally combined to achieve the most "accurate" combined forecast, in the sense of minimum RMSE; this amounts to choosing the optimal weights via OLS estimation of the outturn on the competing point forecasts. Our methodology for optimally combining density forecasts extends this logic and is motivated by the desire to obtain the most "accurate" density forecast, in a statistical sense. It can be contrasted with economic approaches to evaluation, that evaluate forecasts in terms of their implied economic value; see Granger & Pesaran (2000) and Clements (2004).

The plan of this paper is as follows. Section 2 discusses some characteristics of combined density forecasts, and Section 3 proposes a simple approach to optimally choose the combining weights. Section 4 then provides an application to UK inflation. One-year ahead density forecasts of UK inflation are now published each quarter both by the Bank of England in its "fan" chart and the National Institute of Economic and Social Research (NIESR) in its quarterly forecast, and have been for the last ten years. The fan chart is central to the setting of monetary policy by the Monetary Policy Committee at the Bank of England. We examine whether in practice improved density forecasts for inflation might have been obtained if one had optimally combined these competing forecasts. Section 5 concludes.

2 Combination of Density Forecasts

Consider N forecasts made by forecaster i ($i = 1, ..., N$) of a variable y_t at time t ($t = 1, ..., T$), assumed to be real-valued. These N forecasts, denoted g_{it}, are density forecasts, assumed continuous. While the benefits of combining information about point forecasts are well appreciated in economics, less attention has been paid to the aggregation of probability distributions. However, this has received considerable attention within many management science and risk analysis journals; for reviews see Genest & Zidek (1986) and Clemen & Winkler (1999). One popular approach is to aggregate these N density forecasts directly: the "linear opinion pool" takes a weighted linear combination of the forecasters' probabilities. Then the combined density is defined as the finite mixture:

$$p_t(y_t) = \sum_{i=1}^{N} w_i g_{it}(y_t), \tag{1}$$

[1] Related work has considered the combination of event, interval and quantile forecasts; see Clements (2002) and Granger et al. (1989). These inevitably involve a loss of information compared with consideration of the 'whole' density; e.g. only as the number of quantiles examined reaches infinity is no information about the density lost. Garratt et al. (2003) consider the combination of probability forecasts based on Bayesian model averaging.

where w_i are a set of non-negative weights that sum to unity. This combined density satisfies certain properties such as the "unanimity" property (if all forecasters agree on a probability then the combined probability agrees also); for further discussion see Genest & Zidek (1986) and Clemen & Winkler (1999). Further descriptive properties of mixture distributions are summarised in Everitt & Hand (1981).

Inspection of (1) reveals that taking a weighted linear combination of the forecasters' densities can generate a combined density with characteristics quite distinct from those of the forecasters. For example, if all the forecasters' densities are normal, but with different means and variances, then the combined density will be mixture normal. Mixture normal distributions can have heavier tails than normal distributions, and can therefore potentially accommodate skewness and kurtosis. If the true (population) density is non-normal we can begin to appreciate why combining individual density forecasts, that are normal, may mitigate misspecification of the individual densities. Equally, if the true distribution is normal combining using (1) will, in general, get the distribution wrong; for further discussion see Hall & Mitchell (2004).

The key practical issue is how to determine w_i. Granger & Jeon (2004) suggest a thick-modelling approach, based on trimming to eliminate the $k\%$ worst performing forecasts and then taking a simple average of the remaining forecasts. Bayesian model averaging has been suggested also; e.g. see Garratt et al. (2003). This provides a means of weighting alternative model based density forecasts according to their respective posterior probabilities. These probabilities are often proxied by some measure of the relative statistical in-sample fit of the model. Most simply, equal weights, $w_i = 1/N$, have been advocated; e.g. see Hendry & Clements (2004).

In contrast to Bayesian model averaging, the simple data-driven approach to density combination suggested in this paper, which is designed to seek out the "optimal" values of w_i, is not predicated on estimation of a statistical model; it is operational both with model-based and subjective (e.g. survey based) density forecasts.

3 "Optimal" combination of density forecasts: a suggestion

While point forecasts are traditionally evaluated on the basis of RMSE, density forecasts can be evaluated statistically *ex post* using the probability integral transform; see Diebold et al. (1998). They popularised the idea of evaluating a sample of density forecasts based on the idea that a density forecast can be considered "optimal" if the model for the density is correctly specified.

A sequence of estimated density forecasts, $\{p_t(y_t)\}_{t=1}^T$, for the realisations of the process $\{y_t\}_{t=1}^T$, coincides with the true densities $\{f_t(y_t)\}_{t=1}^T$ when the sequence of probability integral transforms, z_t, is independently and identically distributed ($i.i.d.$) with a uniform distribution, U(0,1), where: $z_t = \int_{-\infty}^{y_t} p_t(u)du$, $(t = 1, ..., T)$.

Density forecasts are optimal and capture all aspects of the distribution of y_t only

when the $\{z_t\}$ are both *i.i.d.* and U(0,1). Various statistical tests have been employed to evaluate density forecasts.[2] Let $s(z_t)$ denote a generic test statistic for H_0: optimality. For s and a given choice of size, say 5%, the statistic must have an associated critical region $\{s(z_t) > c\}$, where $P_{H_0}[s(z_t) > c] \leq 5\%$. We reject H_0 when $s(z_t) > c$.

Then define the "optimal" combination weight vector, $\widehat{\mathbf{w}}$, where $\mathbf{w} = (w_1, .., w_N)$, as that $\mathbf{w}$ that minimizes the test statistic $s(z_t)$:

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} s(z_t). \tag{2}$$

Minimizing the test statistic over $\mathbf{w}$ delivers a test statistic with size less than or equal to that associated with $\mathbf{w} \neq \widehat{\mathbf{w}}$ which in turn is less than or equal to the nominal size; for related discussion in terms of testing for common features see Engle & Kozicki (1993).

Unfortunately there is no clear consensus to-date about the most appropriate test for *i.i.d.* uniformity. Various tests have been used in empirical studies; generally, these have included separate tests for the distribution (tests for uniformity or, via a transformation, normality - see Berkowitz (2001)) and dependence. Testing is complicated by the fact that the impact of dependence on the tests for uniformity/normality is unknown, as is the impact of non-uniformity/normality on tests for dependence.

In the application below we focus on the Anderson and Darling (AD) goodness-of-fit test. This tests if a sample of data come from a population with a uniform distribution. Noceti et al. (2003) found the AD test to have more power to detect misspecification in the mean, variance, skewness and/or kurtosis of the forecasts than related distributional tests. The AD test is not designed to be robust to dependence. While we could use some joint test for *i.i.d.* uniformity, it is nevertheless instructive to find those weights that are optimal distributionally.[3] This is particularly so in this application where we might expect serial dependence in $\{z_t\}$ since the forecast horizon is longer than the periodicity of the data. In such a case it is not obvious that we wish to eradicate dependence completely. Comparisons of $\widehat{\mathbf{w}}$ across different statistical tests could be informative in drawing out different aspects of 'optimality'.

4 An application to UK inflation

We focus on quarterly forecasts of one-year ahead RPIX inflation (RPI excluding mortgage payments), the principal monetary policy target over the sample period. The year ahead forecasts correspond to a five quarter ahead horizon.

As discussed by Hendry & Clements (2004), in any application the reasons for success or failure of combination can be multi-faceted. This application is intended to illustrate

[2]Graphical means of exploratory data analysis are often used too; see Diebold et al. (1998).

[3]Hong (2002) has proposed a joint test. This is theoretically attractive as being a joint test one can control the size of the test, something that cannot easily be done using separate tests for uniformity/normality and independence. Thompson (2002) suggests a portmanteau test of unformity and independence.

the use of the proposed method of combination, rather than explain why combination may, or may not, help.

The Bank of England has published one-year ahead inflation density forecasts each quarter from 1993q1. Up until 1995q4 the density forecast is (implicitly) assumed normal. From 1996q1 the Bank has published the so-called "fan" chart, that allows for skewness. The fan chart is based analytically on the two-piece normal distribution; see Wallis (2004).[4]

NIESR density forecasts are published each quarter in the *National Institute Economic Review*. Since 1992q3 NIESR has, in a sense implicitly, published probability forecasts for inflation, in that the *Review* contained a table indicating the historical accuracy of their forecasts based on the mean absolute error.[5] Since 1996q1 NIESR has explicitly published probability forecasts for inflation. Normality is assumed, because earlier work that analysed the historical errors could not reject it. The variance of the density forecast is then set equal to the variance of the historical forecast error.[6] The *Review* focuses on forecasting inflation in the fourth quarter of the current year and the fourth quarter of the next year; therefore only the q4 publication offers a one-year head forecast. While we can extract from back-issues of the *Review* one-year ahead point forecasts for the other quarters, published uncertainty estimates are only available for q4. Therefore, we make an assumption in order to infer uncertainty estimates for the other quarters. We simply assume the density forecast is normal with standard deviation equal across the four quarters in a year. This assumption is sensible if we believe NIESR only re-calibrated their forecast variances once a year.

As is increasingly common in forecasting 'competitions', and following Clements (2004) in his evaluation of Bank density forecasts, we also consider a benchmark density forecast. It is assumed Gaussian with mean equal to actual inflation five quarters previously (so that it is known in real-time) and variance equal to that estimated from the available sample for actual inflation. Using actual inflation data up to 2004q2, we therefore have a sample of 42 density forecasts to compare with the subsequent outturn for RPIX inflation from 1994q1-2004q2.

4.1 In-sample and recursive out-of-sample results

We compare the performance of Bank of England, NIESR, benchmark and combined density forecasts, see (1), both in-sample and using recursive out-of-sample experiments. In-sample we compute the optimal combining weights on the three forecasts using all of the 42 time-series observations. Let w_1 denote the weight on the Bank of England density and w_2 the weight on the NIESR density, implying a weight of $(1 - w_1 - w_2)$ on

[4]The density forecasts from 1993q1-1997q2 are available at: http://www.bankofengland.co.uk/inflationreport/historicalforecastdata.xls. From 1997q3 they are available at http://www.bankofengland.co.uk/inflationreport/rpixinternet.xls.

[5]Assuming normality, a 58% confidence interval around the point forecasts corresponds to the point estimate plus/minus the mean absolute error.

[6]Past forecast errors are commonly used as a practical way of forecasting future errors; e.g. see Wallis (1989), pp. 55-56.

the benchmark density. We restrict attention to positive values of w_i, and search for the optimal weights by considering all combinations of the weights in intervals of 0.01 in $[0, 1]$.

The out-of-sample analysis is designed to simulate whether in practice, in real-time, one could have pooled the Bank of England, NIESR and benchmark density forecasts to obtain 'better' forecasts. Accordingly, from 1997q3 recursively we re-estimate the optimal combining weights using data available up to period $(t-5)$. This acknowledges the fact that one has to wait five quarters to evaluate the performance of a given (year-ahead) forecast. These recursively computed optimal weights are then used to produce a series of combined density forecasts from 1997q4 to 2004q2. Our out-of-sample period corresponds to the period post Bank of England operational independence.

Figure 1 illustrates the in-sample performance of the combined density forecast, as judged by the value of the AD test statistic, for different combinations of weights on the three rival forecasts.

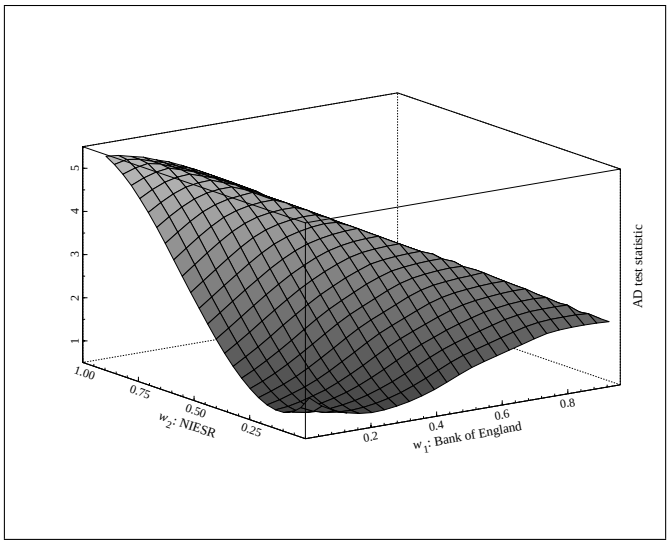

Figure 1: In-sample performance of the combined density forecast for various weights

Figure 1 shows that the optimal weights in-sample, those weights that minimize the AD test statistic at a value of 0.715, are $w_1 = 0.27$, $w_2 = 0.00$, implying a weight of 0.73 on the benchmark forecast. This is a clear improvement with respect to both focusing on one forecast exclusively and simply using equal weights across the three competing forecasts. Using one forecast alone we see that the AD statistic equals 1.722 for $w_1 = 1$, 3.350 for $w_2 = 1$ and 2.549 for $w_1 = 0, w_2 = 0$. So of the three individual density forecasts only those of the Bank of England appear well calibrated; the 95% critical value of the AD test is 2.5. Simply using equal weights across the three competing forecasts delivers an AD test statistic of 2.175. Assuming equal weights between the Bank of England and NIESR, and placing no weight on the benchmark density, yields an AD value of 3.813. Looking at the main diagonal on Figure 1 we see that when placing a zero weight on the benchmark density, the higher the weight on the Bank of England and the lower the weight on NIESR, the better the performance of the combined density. This finding is

consistent with knowledge that NIESR over-estimated the degree of uncertainty.

With the advantage of hindsight we can see that by considering historical forecast errors back until 1982, NIESR were basing their uncertainty forecasts on their track-record across two different inflation 'regimes', the recent regime (post 1992/3) characterized by lower volatility. From 2002 NIESR considered historical forecasting errors from 1993 only and the variance of their density dropped. This serves as a timely reminder to forecasters that just as with point forecasts, basing density forecasts on past experience can lead to misleading forecasts, something in fact well known to NIESR themselves as evidenced by the following quote from Poulizac et al. (1996) p. 62, "Both our inflation forecast and the reliability of this forecast must depend on the seriousness with which the government approaches inflation targetting. It is not clear that past experience is a good guide to this... and, in turn, [this] probably implies that the error variances [based on historical performance]... overstate the current uncertainty associated with the inflation rate".[7]

Table 1 presents the out-of-sample results. It compares the value of the AD test statistic using optimal weights, recursively computed, across the three rival models with equal weights and weighting schemes that focus on the Bank of England, NIESR or benchmark densities alone. Table 1 shows that using optimal weights also delivers gains out-of-sample.

Table 1: Performance of the combined density forecast using various weighting schemes in recursive out-of-sample experiments

weights	AD test statistic
optimal	0.526
Bank: $w_1 = 1$	0.675
NIESR: $w_2 = 1$	4.181
benchmark: $w_1 = 0$; $w_2 = 0$	1.762
equal	1.425

5 Conclusion

This paper proposes a simple means of optimally combining information across competing density forecasts. An application to UK inflation suggests that pooling information across density forecasts can deliver empirical gains. This is consistent with previous findings about point forecasts. Future work should consider alternative weighting schemes and examine how one can statistically test the significance of a given density forecast relative to a rival.

[7]NIESR, see Blake (1996), did consider how stochastic simulation could be used as an alternative to historical errors to measure the uncertainty associated with the inflation rate. It is explained that this is expected to deliver a better measure of uncertainty if a new policy regime (say a new target for inflation) has been adopted. Using a coherent policy structure with interest rate setting determined by a monetary policy rule, Blake found that stochastic simulation suggested a smaller inflation standard error.

References

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

Blake, A. (1996), 'Forecast error bounds by stochastic simulation', *National Institute Economic Review* **156**, 72–79.

Clemen, R. & Winkler, R. (1999), 'Combining probability distributions from experts in risk analysis', *Risk Analysis* **19**, 187–203.

Clements, M. P. (2002), 'An evaluation of the survey of professional forecasters probability distributions of expected inflation and output growth', *Warwick University Discussion Paper* .

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

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

Engle, R. F. & Kozicki, S. (1993), 'Testing for common features', *Journal of Business and Economic Statistics* **11**, 369–380.

Everitt, B. S. & Hand, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

Garratt, A., Lee, K., Pesaran, M. H. & Shin, Y. (2003), 'Forecast uncertainties in macroeconometric modelling: an application to the UK economy', *Journal of the American Statistical Association* **98**, 829–838.

Genest, C. & Zidek, J. (1986), 'Combining probability distributions: a critique and an annotated bibliography', *Statistical Science* **1**, 114–135.

Granger, C. W. J. & Jeon, Y. (2004), 'Thick modeling', *Economic Modelling* **21**, 323–343.

Granger, C. W. J. & Pesaran, M. H. (2000), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting* **19**, 537–560.

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

Hall, S. G. & Mitchell, J. (2004), Density forecast combination. National Institute of Economic and Social Research Discussion Paper No. 249.

Hendry, D. F. & Clements, M. P. (2004), 'Pooling of forecasts', *Econometrics Journal* **7**, 1–31.

Hong, Y. (2002), Evaluation of out-of-sample probability density forecasts. Cornell University Discussion Paper.

Noceti, P., Smith, J. & Hodges, S. (2003), 'An evaluation of tests of distributional forecasts', *Journal of Forecasting* **22**, 447–455.

Poulizac, D., Weale, M. & Young, G. (1996), 'The performance of National Institute economic forecasts', *National Institute Economic Review* **156**, 55–62.

Stock, J. & Watson, M. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**, 405–430.

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

Thompson, S. (2002), Evaluating the goodness of fit of conditional distributions, with an application to affine term structure models. manuscript Economics Department, Harvard University.

Wallis, K. F. (1989), 'Macroeconomic forecasting: a survey', *Economic Journal* **99**, 28–61.

Wallis, K. F. (2004), 'An assessment of Bank of England and National Institute inflation forecast uncertainties', *National Institute Economic Review* **189**, 64–71.