

Optimal combination of density forecasts*

Stephen G. Hall

Department of Economics, Leicester University, Leicester, UK
and National Institute of Economic and Social Research, London, UK

James Mitchell

National Institute of Economic and Social Research, London, UK

This Revision: December 2005

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. These optimal weights are those weights that minimize the ‘distance’, as measured by the Kullback-Leibler information criterion, between the forecasted and true but unknown density. We explain how this minimization both can and should be achieved. Comparisons with the optimal combination of point forecasts are made. An application to simple time-series density forecasts and two widely used published density forecasts for U.K. inflation, namely the Bank of England and NIESR “fan” charts, illustrates that combination can but need not always help.

JEL Classification: C53; E37

Keywords: Density forecasts; Uncertainty; Combining forecasts; Evaluating forecasts; Inflation forecasting

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

*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 Ken Wallis, Martin Weale, an Associate Editor and two anonymous referees for helpful comments. Mitchell gratefully acknowledges financial support from the ESRC (Award Reference: RES-000-22-0610).

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 forecast combination. We propose a simple data-driven approach to combine density forecasts directly using optimal weights. These optimal weights are those weights that minimize the ‘distance’, as measured by the Kullback-Leibler information criterion [*KLIC*], between the combined forecast density and true (but unknown) density.

While Clements (2005) and Granger et al. (1989) have considered respectively the combination of event and quantile forecasts, that inevitably involve a loss of information compared with consideration of the ‘whole’ density, the combination of density forecasts has been relatively neglected. Indeed Clements (2003) identifies this as “an area waiting investigation” (p.2).

However the finite mixture distribution offers a well understood and much exploited means of combining density forecasts; see Wallis (2005). For example, the Survey of Professional Forecasters [SPF], previously the ASA-NBER survey, has essentially used it since 1968 to publish a combined density forecast of inflation, amongst other things. Since respondents to the SPF supply density forecasts in the form of histograms the average or combined density forecast is defined as the mean density forecast across respondents.¹ Despite this long history, to-date little attention has been paid to how the weights on the competing density forecasts in the finite mixture should be determined. But as experience of combining point forecasts has taught us, irrespective of its performance in practice, use of equal weights is only one of many options. For example, one popular alternative to equal weights in the point forecast literature, the so-called regression approach, is to tune the weights to reflect the historical performance of the competing forecasts; e.g. see Granger & Ramanathan (1984) [GR].

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 subsequent realizations of the variable. 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 realizations of the variable 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

¹The SPF survey has been analysed *inter alia* by Zarnowitz & Lambros (1987), Diebold et al. (1999), Giordani & Söderlind (2003) and Clements (2005).

“accurate” density forecast, in a statistical sense.² This is defined as that set of weights in the finite mixture that minimize the *KLIC* distance between the combined density forecast and the true but unknown density of the variable to be forecast. Practically, and conveniently, this minimization can be achieved using the logarithmic scoring rule. Scoring rules evaluate the quality of density forecasts by assigning a numerical score based on the forecast and the subsequent realization of the variable. The use of scoring rules is attractive as it circumvents the need to specify/estimate either the unknown true density of the variable to be forecast or the density of the probability integral transforms of the realization of the variable with respect to the forecast densities.

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 choose the combining weights optimally. Comparisons with the optimal combination of point forecasts are made. 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 with a simple time-series density forecast. Section 5 concludes.

2 Combination of Density Forecasts

Consider N forecasts made by a group of forecasters i ($i = 1, \dots, N$) of a variable y_t at time t ($t = 1, \dots, T$), assumed to be real-valued. These N forecasts, denoted g_{it} , are density forecasts, assumed continuous.³ 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:

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

²It 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).

³A continuous parametric density function can always be fitted to discrete density forecasts. For example, Giordani & Söderlind (2003) fit normal distributions to the SPF histograms.

where w_i are a set of non-negative weights that sum to unity.⁴ 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). Recently Wallis (2005) has also motivated $p_t(y_t)$ as an appropriate statistical model for a combined density forecast. Further descriptive properties of mixture distributions are summarised in Everitt & Hand (1981).

Hall & Mitchell (2004) considered an alternative approach to density forecast combination, again deriving from management science, that combines density forecasts explicitly taking into account their dependence. Following Morris (1974, 1977) and Winkler (1981) they adopt a Bayesian approach where competing densities are combined by a “decision maker” who views them as data that are used to update a prior distribution. Hall and Mitchell also distinguish between combining competing forecasts of various moments of the forecast density and directly combining the individual densities themselves, as with the finite mixture density. We do not consider this approach further.

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.

Further characteristics of the combined density $p_t(y_t)$ can be drawn out by defining m_{it} and v_{it} as the mean and variance of forecast i ’s distribution at time t : $m_{it} = \int_{-\infty}^{\infty} y_t g_{it}(y_t) dy_t$

and $v_{it} = \int_{-\infty}^{\infty} (y_t - m_{it})^2 g_{it}(y_t) dy_t$; ($i = 1, \dots, N$). Then the mean and variance of (1) are given by:⁵

$$E[p_t(y_t)] = m_t^* = \sum_{i=1}^N w_{it} m_{it}, \quad (2)$$

$$\text{Var}[p_t(y_t)] = \sum_{i=1}^N w_{it} v_{it} + \sum_{i=1}^N w_{it} \{m_{it} - m_t^*\}^2. \quad (3)$$

⁴The restriction that the weights are positive might be relaxed; for discussion and references see Genest & Zidek (1986). In the finite mixture distribution the weights, the mixing proportions, are positive by construction; see Everitt & Hand (1981).

⁵Related expressions decomposing the aggregate density (1), based on the ‘law of conditional variances’, are seen in Giordani & Söderlind (2003). This law states that for the random variables y_t and i : $V(y_t) = E[V(y_t|i)] + V[E(y_t|i)]$. For criticism see Wallis (2005).

(3) indicates that the variance of the combined distribution equals average individual uncertainty (“within” model variance) plus disagreement (“between” model variance).⁶ This result stands in contrast to that obtained when combining point forecasts where combination using “optimal” (variance or RMSE minimising) weights means the RMSE of the combined forecast must be equal to or less than that of the smallest individual forecast; see Bates & Granger (1969) and for related discussion in a regression context Granger & Ramanathan (1984). Density forecast combination will in general increase the combined variance. However, this increase in uncertainty need not be deleterious; when evaluated the combined density forecast may perform better than the individual density forecasts.

The key practical issue is to determine w_i . Most simply, equal weights, $w_i = 1/N$, have been advocated; e.g. see Hendry & Clements (2004) and for related discussion Wallis (2005). 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 (BMA) 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 in-sample statistical fit of the model used to generate the forecasts. Alternatively, also within a BMA framework, Mitchell & Hall (2005) suggest so-called ‘KLIC weights’. These weight the competing density forecasts according to the probability that they are the most accurate. However, these weights are computed using the probability integral transforms of the realization of the variable with respect to the forecast densities and do not require estimation of a statistical model. Likewise 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

We propose to combine density forecasts optimally by identifying that set of weights that deliver the most “accurate” density forecast, in a statistical sense. They are the set of weights in the finite mixture, (1), that minimize the *KLIC* ‘distance’ between the combined density forecast and the true, but unknown, density $f_t(y_t)$ which delivers the realizations of the variable $\{y_t\}_{t=1}^T$. Note that $\{y_t\}_{t=1}^T$ may, or may not, be covariance-stationary.

Specifically, the *KLIC* distance between the true density $f_t(y_t)$ and the combined

⁶For further discussion of the relationship, if any, between dispersion/disagreement and individual uncertainty see Bomberger (1996).

density forecast $p_t(y_t)$ ($t = 1, \dots, T$) is defined as:

$$KLIC_t = \int f_t(y_t) \ln \left\{ \frac{f_t(y_t)}{p_t(y_t)} \right\} dy_t \text{ or} \quad (4)$$

$$KLIC_t = \mathbb{E} [\ln f_t(y_t) - \ln p_t(y_t)]. \quad (5)$$

The smaller this distance the closer the density forecast to the true density. $KLIC_t = 0$ if and only if $f_t(y_t) = p_t(y_t)$. For a related discussion see White (1982) and Vuong (1989).

Under some regularity conditions $\mathbb{E} [\ln f_t(y_t) - \ln p_t(y_t)]$ can be consistently estimated by \overline{KLIC} , the average of the sample information on $f_t(y_t)$ and $p_t(y_t)$ ($t = 1, \dots, T$):

$$\overline{KLIC} = \frac{1}{T} \sum_{t=1}^T [\ln f_t(y_t) - \ln p_t(y_t)]. \quad (6)$$

Definition 1 *The optimal combined density forecast is $p_t^*(y_t) = \sum_{i=1}^N w_i^* g_{it}(y_t)$, where the optimal weight vector \mathbf{w}^* , $\mathbf{w}^* = (w_1^*, \dots, w_N^*)$, minimizes the $KLIC$ distance between the combined and true density, (6). This minimization is achieved as follows:*

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \frac{1}{T} \sum_{t=1}^T \ln p_t(y_t), \quad (7)$$

where $\frac{1}{T} \sum_{t=1}^T \ln p_t(y_t)$ is the average logarithmic score of the combined density forecast over the sample $t = 1, \dots, T$.

Definition 1 explains that the optimal weights are found by numerically searching for that set of weights which maximize the average logarithmic score of the combined density forecast with respect to the realization of the variable. Even when $p_t(\cdot)$ does not contain $f_t(\cdot)$ maximizing the average logarithmic score minimizes \overline{KLIC} ; for related discussion in terms of quasi-maximum likelihood estimation see White (1982). Note that, as we discuss further below, minimizing \overline{KLIC} by maximizing the logarithmic score is convenient as it avoids having to postulate and estimate $f_t(y_t)$, which is unknown. The logarithmic scoring rule is intuitively appealing as it gives a high score to a density forecast that provides a high probability to the value y_t that materializes. We note that alternative distance measures to $KLIC$, such as the integrated squared difference [see Li & Tkacz (2004)], might be considered although they do not appear so convenient for the optimal combination of density forecasts. This is because they have to specify/estimate $f_t(y_t)$; Li & Tkacz (2004), for example, estimate $f_t(y_t)$ using a nonparametric kernel estimator.

The optimally combined density forecast cannot provide worse forecasts (in-sample $t = 1, \dots, T$), as evaluated by the average logarithmic score, than the best individual forecast. Since $KLIC_{it} = \mathbb{E} [\ln f_t(y_t) - \ln g_{it}(y_t)] \geq 0$, $\mathbb{E}(\ln p_t(y_t)) \geq \mathbb{E}(\ln g_{it}(y_t))$ implying $KLIC_t \leq KLIC_{it}$ ($i = 1, \dots, N$; $t = 1, \dots, T$); see also Raftery et al. (1997). Since the best model $g_{it}(y_t)$ according to the KLIC is the model with the highest posterior probability [see Fernandez-Villaverde & Rubio-Ramirez (2004)], we might think of the optimally combined density forecast $p_t^*(y_t)$ as the model, albeit combined, with the highest posterior probability.

3.1 A comparison with goodness-of-fit tests

In practice forecasters often use alternatives to scoring rules to evaluate density forecasts statistically. A popular alternative is to evaluate them based on the probability integral transforms (pits) of the realization of the variable with respect to the forecast densities; see Diebold et al. (1998).⁷ Diebold et al. showed that a density forecast can be considered optimal, importantly irrespective of the user's loss function, if the model for the density is correctly specified. The density forecast is 'correct' when the pits are uniform and for one-step ahead density forecasts also independently and identically distributed (i.i.d.). In practice evaluation therefore requires application of a goodness-of-fit test - a statistical test for uniformity or, via a transformation, normality. Unfortunately there is no clear consensus to-date about the most appropriate test.

Various goodness-of-fit tests have been used in empirical studies; generally, these have included separate tests for the distribution and independence. Tests used include the Kolmogorov-Smirnov, Anderson-Darling and Doornik-Hansen distributional tests; Ljung-Box and LM tests for independence; Hong, Thompson and Berkowitz Likelihood Ratio tests for the distribution and independence. For empirical examples and references see Clements & Smith (2000), Clements (2004) and Hall & Mitchell (2004).

Those goodness-of-fit tests directly related to the KLIC, such as the Berkowitz (2001) likelihood ratio (LR) test, do offer an alternative to scoring rules to combine density forecasts optimally. As we explain in Section 3.1.1, again one can numerically search for that set of weights in the combination (1) which minimize the KLIC.

But as Wallis (2005) identifies, confronted by apparently competing methods of determining the track record of rival density forecasts (different distance measures, scoring rules and goodness-of-fit statistics), it is important when choosing the combination weights to offer guidance to practitioners about the most appropriate method. Accordingly, we suggest that tests based on the pits are less suited than the logarithmic scoring rule to choosing how to weight density forecasts optimally in a combination. This is because they require the practitioner to specify/estimate an unknown density such as $f_t(y_t)$, something conveniently avoided when using the logarithmic scoring rule. In addition the logarithmic score has a direct link with the KLIC, something apparently not offered by all goodness of fit tests. For these goodness-of-fit tests, such as the Anderson-Darling test, again one could numerically search for that set of weights that minimize the test statistic. We summarise this in Definition 2* in the Appendix. In the absence of a proof of a direct link with the KLIC, of the sort supplied by Bao et al. (2004) for the LR test, we do not claim the weights derived from Definition 2* are optimal in the way the weights are in both Definition 1 and Definition 2 (below).

⁷An alternative approach is based on the integrated squared difference between the density forecast and a nonparametric estimate of $f_t(y_t)$; see Li & Tkacz (2004). For a critical review of the various tests see Corradi & Swanson (2005).

3.1.1 Berkowitz's LR test

Consider the Berkowitz (2001) LR test. As explained by Bao et al. (2004), the LR test in fact provides estimates of \overline{KLIC} . This is seen by invoking Berkowitz (2001) Proposition 2, and noting the following equivalence:

$$\ln f_t(y_t) - \ln p_t(y_t) = \ln h_t(z_t) = \ln q_t(z_t^*) - \ln \phi(z_t^*), \quad (8)$$

where $z_t = \int_{-\infty}^{y_t} p_t(u)du$ are the pits, $z_t^* = \Phi^{-1}z_t$, $h_t(\cdot)$ and $q_t(\cdot)$ are respectively the unknown densities of z_t and z_t^* , $\phi(\cdot)$ is the standard normal density and Φ is the c.d.f. of the standard normal. Conditional on specification of $q_t(z_t^*)$ estimates of \overline{KLIC} can then be derived.

To explain further consider Berkowitz's three degrees-of-freedom LR test for the joint null hypothesis of a zero mean, unit variance and independence of z_t^* against z_{t-1}^* following a first-order autoregressive process: $z_t^* = \mu + \rho z_{t-1}^* + \varepsilon_t$, where $\text{Var}(\varepsilon_t) = \sigma^2$. This corresponds to assuming:

$$q_t(z_t^*) = \phi \left[(z_t^* - \mu - \rho z_{t-1}^*) / \sigma \right] / \sigma. \quad (9)$$

The LR test is then:

$$LR = 2 \sum_{t=1}^T [\ln q_t(z_t^*) - \ln \phi(z_t^*)]. \quad (10)$$

so that $\overline{KLIC} = LR/2T$. For further discussion see Mitchell & Hall (2005).

In theory, therefore, density forecasts might also be optimally combined by minimizing LR rather than maximizing the logarithmic score. This is summarized in Definition 2.

Definition 2 *The optimal combination weight vector, \mathbf{w}^* , is that set of weights which minimize the LR test statistic for testing the null hypothesis that $\mu = 0$, $\rho = 0$ and $\sigma^2 = 1$ in the model:*

$$\Phi^{-1}z_t = \mu + \rho\Phi^{-1}z_{t-1} + \varepsilon_t \quad (11)$$

where $\varepsilon_t \sim N(0, \sigma^2)$. Let LR_{\min} denote the value of LR associated with \mathbf{w}^* .

3.1.2 Recommendations

But in practice, as alluded to above, we suggest that practitioners follow the scoring rules approach to the optimal combination of density forecasts, summarized in Definition 1. This is because KLIC minimization using tests based on the pits is only as good as the underlying goodness-of-fit test. $q_t(\cdot)$ should be specified to nest i.i.d. normality, which it equals when $f_t(y_t) = p_t(y_t)$ implying $LR = \overline{KLIC} = 0$. But there is the danger that the true density for z_t^* is not nested by the chosen specification for $q_t(\cdot)$. In such a case the weights chosen via Definition 2 will not minimize the 'true' KLIC distance between the combined and true but unknown density.

For example, consider $q_t(\cdot)$ as specified in (9). This means that (10) has power to detect non-normality only through the first two moments. This explains why some authors

when evaluating individual density forecasts, such as Clements & Smith (2000) and Hall & Mitchell (2004), have supplemented the Berkowitz LR test with a nonparametric normality test, such as the Doornik-Hansen test. Consequently $\{\Phi^{-1}z_t\}_{t=1}^T$ can exhibit skewness and/or kurtosis, but this would not be detected by the LR test. We leave it for future research to determine how this might lead to Definition 2, but perhaps not Definition 1, choosing the ‘wrong’ weights. It will also be useful to determine whether the semi nonparametric density function for $q_t(\cdot)$, suggested by Bao et al. (2004), because of its less restrictive form might mitigate this problem. However, in the empirical application below we do compare weights from both Definitions 1 and 2.

Finally we gather together the following remarks.

Remark 3 *The LR test statistic can be used both to estimate the combination weights and test whether the combined density forecast is correctly specified. Consider the null hypothesis that the combined density forecast is correctly specified $H_0: \overline{KLIC} = LR = 0$. Under the null hypothesis $LR_{\min} \sim \chi_3^2$.⁸*

Remark 4 *In contrast the average logarithmic score, see Definition 1, cannot be used to test the null hypothesis $H_0: \overline{KLIC} = 0$ for the optimally combined density forecast. This can be achieved, see (8), only if the practitioner specifies $f_t(\cdot)$ or $q_t(\cdot)$.*

Remark 5 *However, one can compare competing density forecasts based on their logarithmic scores without having to specify $f_t(\cdot)$ or $q_t(\cdot)$. A Diebold-Mariano type test of equal predictive accuracy of two density forecasts can be constructed following Giacomini (2002) and Mitchell & Hall (2005).*

3.2 Comparison with optimal combination of point forecasts

Let us draw out some further characteristics of the proposed density forecast combination method by relating it to the combination of point forecasts.

The optimal weights \mathbf{w}^* which minimize the *KLIC* distance between the combined and true density are comparable to the optimal weights used to combine point forecasts. They are comparable in the sense that like when optimal weights are used to combine competing point forecast, accuracy, in this case measured by RMSE, cannot get worse. Similarly accuracy, as defined in Definitions 1 and 2, cannot get worse when optimal weights are used to combine competing density forecasts. As Wallis (2005) explains, optimal weights mean it can be helpful to combine a good forecast with an inferior forecast. But, as with the optimal combination of point forecasts, there is no guarantee that optimal weights \mathbf{w}^* deliver better density forecasts out-of-sample.

The estimated weights from Definition 1 and 2 can be 1,0 ($w_1^* = 1$ and $w_i^* = 0$ for $\forall i \neq 1$) but this does not necessarily, in contrast to optimal combination of point forecasts

⁸Since the LR test statistic is used first to estimate the combination weights, see Definition 2, and then test the optimally combined density forecast care should be exercised in using the critical values from the χ^2 distribution to test H_0 . The test may not have correct asymptotic size.

when $RMSE = 0$, imply the forecast with a weight of 1 is optimal ($\overline{KLIC} = 0$). It just implies it is better than the other density forecast. Only when the ‘true’ model is in the set of N models under consideration will the $KLIC$ distance also be zero.

However, when the optimal weights are not 1,0 this does imply, as in the case of point forecasts combined following GR, that combination will deliver improved density forecasts (in-sample) - a higher average logarithmic score under Definition 1 and a lower LR test statistic under Definition 2.

The analogous case with point forecast combination appears to be testing for “conditional efficiency” (*encompassing*) of forecast 1 relative to forecast 2 (a zero coefficient on forecast 2 in the GR regression) but not simultaneously “Mincer-Zarnowitz [MZ] efficiency” (a unit coefficient on forecast 1 in the GR regression).⁹ With density forecasts the counterpart of MZ efficiency is met only when $\overline{KLIC} = 0$ for the combined (1,0), or best individual, density forecast. So to establish efficiency, in both senses, of the combined density forecast it is important to supplement examination of the weights with a statistical test for $H_0: \overline{KLIC} = LR = 0$. Density forecast combination via the linear opinion pool requires the weights to sum to unity. Future work is required to consider how we might simultaneously examine the density analogues of conditional and MZ efficiency.

In addition it would be helpful to move from inspection of combination weights to tests of their statistical significance by accounting for their uncertainty. Practitioners could then undertake encompassing tests. Interpreting \mathbf{w}^* from Definition 1 as a quasi-maximum likelihood estimator is an option.¹⁰ Under regularity conditions, consistent estimates of the standard errors associated with the optimal weights might be obtained from the inverse of the Hessian matrix, namely the inverse of a matrix whose ij -th element is given by:

$$\mathbb{E} \left[\frac{\partial \ln p_t(y_t)}{\partial w_i^*} \frac{\partial \ln p_t(y_t)}{\partial w_j^*} \right]. \quad (12)$$

In practice maximum likelihood estimation of unknown parameters in a finite mixture distribution (in our case the mixing proportions w_i) can be difficult even when N is small and it is assumed $g_{it}(\cdot) = g_i(\cdot)$ is normal; see Everitt & Hand (1981).

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

⁹Related discussion for forecast probabilities of events is provided by Clements (2005).

¹⁰An alternative is to consider a bootstrap procedure.

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). The density forecasts are available from the Bank of England’s web-site.

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.¹¹ 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.¹² 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 follow Mitchell (2005) in his summary of National Institute density forecasts and 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. The quarterly time-series of NIESR density forecasts used in this paper are available from Mitchell (2005).

Alongside these published forecasts we also consider a simple time-series 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 realizations of RPIX inflation from 1994q1-2004q2.

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

We compare the performance of Bank of England, NIESR, time-series and combined density forecasts 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 the time-series 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,

¹¹Assuming normality, a 58% confidence interval around the point forecasts corresponds to the point estimate plus/minus the mean absolute error.

¹²Past forecast errors are commonly used as a practical way of forecasting future errors; e.g. see Wallis (1989), pp. 55-56.

one could have pooled the Bank of England, NIESR and time-series 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 average logarithmic score ($\log S$), for different combinations of weights on the three rival forecasts.

Figure 1 shows that the optimal weights in-sample according to Definition 1, those weights that maximize the average logarithmic score at a value of -0.698 , are $w_1^* = 0.41$, $w_2^* = 0.00$, implying a weight of 0.59 on the time-series forecast. This is an 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 from Figure 1 that the average logarithmic score equals -0.826 for $w_1 = 1$, -1.554 for $w_2 = 1$ and -0.822 for $w_1 = 0, w_2 = 0$. So of the three individual density forecasts, NIESR forecasts perform worse.¹³ Simply using equal weights across the three competing forecasts delivers a logarithmic score of -0.894 . Assuming equal weights between the Bank of England and NIESR, and placing no weight on the time-series density, yields an average logarithmic score of -1.112 . Looking at the main diagonal on Figure 1 we see that when placing a zero weight on the time-series 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.

In a related application combining just the Bank of England and NIESR density forecasts Mitchell & Hall (2005) found, within a BMA framework, their KLIC weights led to the combined density forecast performing worse than the best individual forecast. The KLIC weight on the Bank of England was 0.572 , and 0.428 on NIESR. From Figure 1 we can see that these weights result in an average logarithmic score for the combined density forecast of -1.07 , worse than the Bank of England’s score. In contrast the optimal weights ensure the combined density forecast performs at least as well as the best individual forecast. When combining just the Bank of England and NIESR density forecasts we do indeed find that the optimally combined density forecast performs only as well as the best individual forecast, but crucially no worse.

Further to Section 3.1, although we recommend use of the scoring rule for the optimal combination of density forecasts, for completeness we did consider combination using the Berkowitz LR test (Definition 2). Since we are looking at five-step ahead quarterly forecasts we expect autocorrelation in z_t^* and therefore consider a two rather than three degrees-of-freedom test which allows for possible dependence. That is, we set $q_t(z_t^*) =$

¹³These results are consistent with earlier studies that tend to find that the Bank of England (year-ahead) density forecasts fail tests for independence (constituting no violation of forecast optimality) but perform better against the distributional ones, at least over the 1997q3- period; e.g. see Clements (2004).

$\phi[(z_t^* - \mu)/\sigma]/\sigma$. Numerical optimization implied an optimal weight of 0.19 on the Bank of England and 0.81 on the time-series model. Although the weights are sensitive to how the KLIC is minimized, NIESR continues to receive a weight of zero. This is also the case under Definition 2*, when the Anderson-Darling test statistic is minimized; the weight on the Bank of England is 0.27 and the weight on the time-series forecast is 0.73. It is also of note that the implied p -value for the LR test on the optimally combined density forecast was 0.656. This suggests that this forecast is well-calibrated.

Table 1 presents the out-of-sample results. It compares the value of the average logarithmic score using optimal weights from Definition 1, recursively computed, across the three rival models with equal weights and weighting schemes that focus on the Bank of England, NIESR or time-series densities alone. Table 1 shows that using optimal weights also delivers gains out-of-sample, albeit slight, relative to the best individual density forecast (the Bank of England).

Table 1: Performance, as judged by the average logarithmic score ($\log S$), of the combined density forecast using various weighting schemes in recursive out-of-sample experiments

weights	$\log S$
optimal	-0.611
Bank: $w_1 = 1$	-0.612
NIESR: $w_2 = 1$	-1.420
time-series: $w_1 = 0; w_2 = 0$	-0.722
equal	-0.750

4.2 Lessons for the future production of density forecasts

Despite increased interest few organizations currently publish density forecasts in real-time. It is therefore helpful to draw two lessons for the future production of density forecasts. The first pertains to the construction of individual density forecasts, while the second concerns combination.

1. 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 when regimes change. NIESR was in fact well aware of this. For example, to 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 targeting. 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". But until 2002 NIESR continued to publish uncertainty forecasts based on forecast errors back to 1982. However Mitchell (2005) has found that a break in the unconditional variance of NIESR's forecast errors around 1993-94 could have been detected via recursive analysis of these forecast errors towards the end of 1996. It is therefore important to monitor historical forecast errors regularly, using statistical tests for structural breaks at an unknown point, to help select a period of history which is informative about the future. Stochastic simulation has been discussed as an alternative to historical errors to measure the uncertainty associated with the inflation rate; see Blake (1996). It is explained that this is expected to deliver a better measure of uncertainty if a new policy regime (such as a new target for inflation) is 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 than analysis of the historical errors.

2. Density forecast combination using optimal weights can help. Table 2 shows that the accuracy of NIESR's real-time forecasts would have been improved if they had taken an optimal combination of their forecast with those of the Bank of England and the time-series forecast. Indeed they would have realized their forecast received a weight of zero.

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 competing density forecasts can but need not deliver empirical gains. Encouragingly from the perspective of potential practitioners, the optimal weights introduced in this paper ensure that the combined density forecast does not perform worse and can perform better, on an in-sample basis, than the best individual density forecast. But in the empirical application we saw that the combined density forecast need not beat the best individual density forecast. This was seen when combining the Bank of England and NIESR density forecasts; the optimal combination of these two densities was 1 and 0, that is no combination. This result is consistent with our prior that combination with an inferior forecast need not help. But it will not make matters worse, at least in-sample, if optimal weights are used. Instead if one gave the NIESR density a positive weight combination would indeed produce a less accurate density forecast than the Bank of England. Clearly as with the combination of point forecasts the weights used matter.

But one important difference with the optimal combination of point forecasts is that 1,0 weights do not imply that the combined density forecast is itself correctly specified. They just imply that one forecast is better than the other. However, consistent with previous findings suggesting that combination of point forecasts can help, when optimally combining the Bank of England and time-series density forecasts we found that

combination is better than using either of the forecasts individually.

Our encouraging findings therefore suggest that density forecast combination is a productive area for further research, both theoretical and applied. It is important to try and build up both an increased understanding and an empirical consensus about the circumstances in which density forecast combination works. Practitioners can then use density forecast combination methods, such as the one proposed in this paper, with increased confidence.

Appendix

Definition 2* Let $s(z_t)$ denote a goodness-of-fit test statistic for $H_0 : \overline{KLIC} = 0$. 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:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} s(z_t). \quad (\text{A.1})$$

Minimizing the test statistic over w delivers a test statistic with size less than or equal to that associated with $w \neq w^*$ 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).

References

- Bao, Y., Lee, T.-H. & Saltoglu, B. (2004), A test for density forecast comparison with applications to risk management. Department of Economics, UC Riverside.
- Bates, J. M. & Granger, C. W. J. (1969), ‘The combination of forecasts’, *Operational Research Quarterly* **20**, 451–468.
- 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.
- Bomberger, W. (1996), ‘Disagreement as a measure of uncertainty’, *Journal of Money, Credit and Banking* **28**, 381–392.
- Clemen, R. & Winkler, R. (1999), ‘Combining probability distributions from experts in risk analysis’, *Risk Analysis* **19**, 187–203.
- Clements, M. P. (2003), ‘Editorial: Some possible directions for future research’, *International Journal of Forecasting* **19**, 1–3.
- Clements, M. P. (2004), ‘Evaluating the Bank of England density forecasts of inflation’, *Economic Journal* **114**, 844–866.
- Clements, M. P. (2005), ‘Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts’, *Empirical Economics*. Forthcoming.
- Clements, M. P. & Smith, J. (2000), ‘Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment’, *Journal of Forecasting* **19**, 255–276.

- Corradi, V. & Swanson, N. R. (2005), Predictive density evaluation, *in* G. Elliott, C. W. J. Granger & A. Timmermann, eds, ‘Handbook of Economic Forecasting’, North-Holland, North Holland. Forthcoming.
- Diebold, F. X., Gunther, A. & Tay, K. (1998), ‘Evaluating density forecasts with application to financial risk management’, *International Economic Review* **39**, 863–883.
- Diebold, F. X., Tay, A. S. & Wallis, K. F. (1999), Evaluating density forecasts of inflation: the Survey of Professional Forecasters, *in* R. Engle & H. White, eds, ‘Cointegration, causality and forecasting: a festschrift in honour of Clive W. J. Granger’, Oxford University Press.
- 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.
- Fernandez-Villaverde, J. & Rubio-Ramirez, J. (2004), ‘Comparing dynamic equilibrium economies to data: A Bayesian approach’, *Journal of Econometrics* **123**, 153–187.
- Garratt, A., Lee, K., Pesaran, M. H. & Shin, Y. (2003), ‘Forecast uncertainties in macroeconomic 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.
- Giacomini, R. (2002), Comparing density forecasts via weighted likelihood ratio tests: asymptotic and bootstrap methods. UCSD Discussion Paper 2002-12.
- Giordani, P. & Söderlind, P. (2003), ‘Inflation forecast uncertainty’, *European Economic Review* **47**, 1037–1059.
- 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. & Ramanathan, R. (1984), ‘Improved methods of combining forecasts’, *Journal of Forecasting* **3**, 197–204.
- 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.
- Li, F. & Tkacz, G. (2004), ‘A consistent bootstrap test for conditional density functions with time-dependent data’, *Journal of Econometrics* . Forthcoming.
- Mitchell, J. (2005), ‘The National Institute density forecasts of inflation’, *National Institute Economic Review* **193**, 60–69.
- Mitchell, J. & Hall, S. G. (2005), ‘Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR ”fan” charts of inflation’, *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.
- Morris, P. (1974), ‘Decision analysis expert use’, *Management Science* **20**, 1233–1241.
- Morris, P. (1977), ‘Combining expert judgments: A Bayesian approach’, *Management Science* **23**, 679–693.
- Poulizac, D., Weale, M. & Young, G. (1996), ‘The performance of National Institute economic forecasts’, *National Institute Economic Review* **156**, 55–62.
- Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997), ‘Bayesian model averaging for linear regression models’, *Journal of the American Statistical Association* **92**, 179–191.
- 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.
- Vuong, Q. (1989), ‘Likelihood ratio tests for model selection and non-nested hypotheses’, *Econometrica* **57**, 257–306.
- 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.
- Wallis, K. F. (2005), ‘Combining density and interval forecasts: a modest proposal’, *Oxford Bulletin of Economics and Statistics* **67**, 983–994.
- White, H. (1982), ‘Maximum likelihood estimation of misspecified models’, *Econometrica* **50**, 1–25.
- Winkler, R. (1981), ‘Combining probability distributions from dependent information sources’, *Management Science* **27**, 479–488.
- Zarnowitz, V. & Lambros, L. (1987), ‘Consensus and uncertainty in economic prediction’, *Journal of Political Economy* **95**, 591–621.

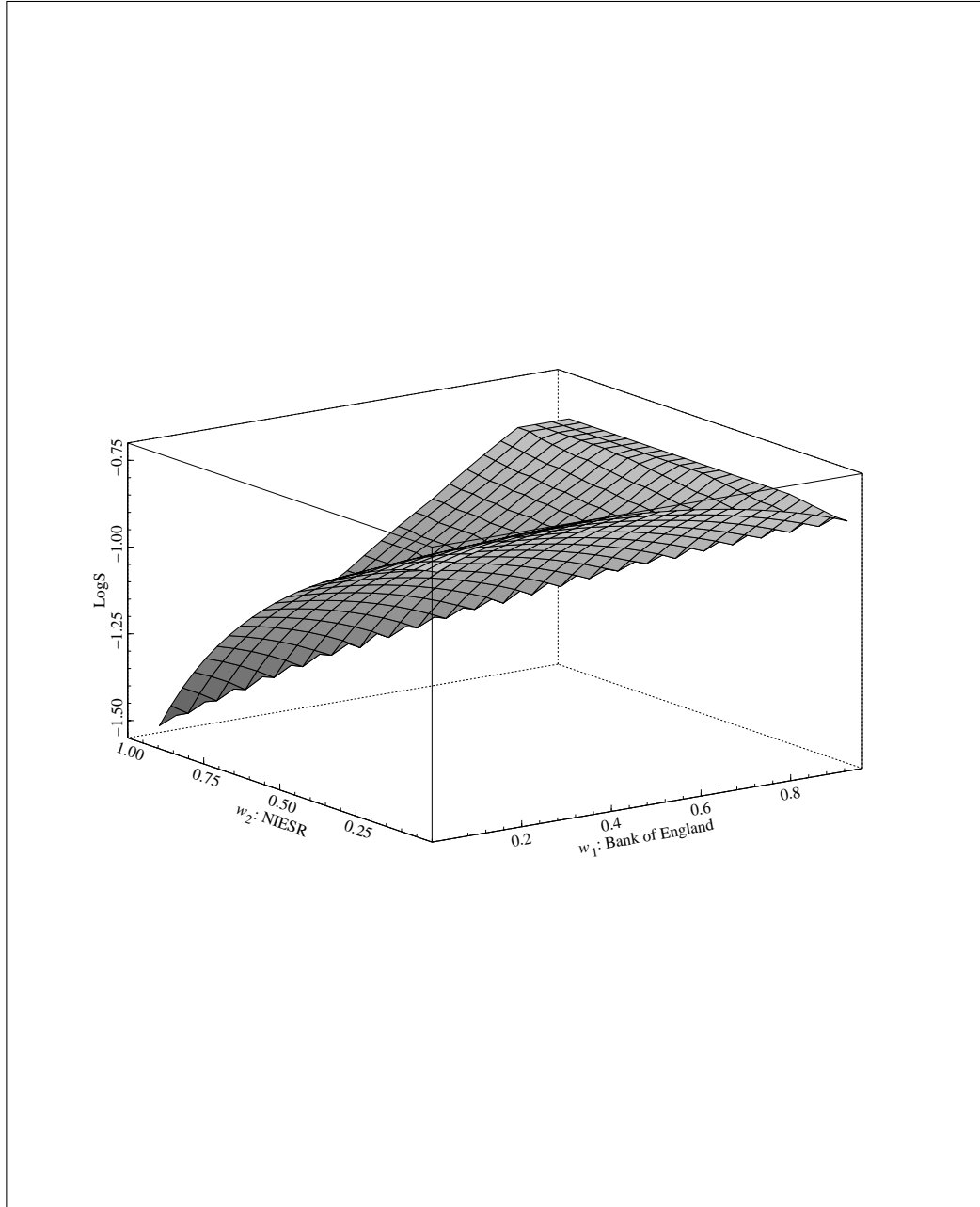


Figure 1: In-sample performance, as judged by the average logarithmic score ($\log S$), of the combined density forecast for various weights