FORECAST UNCERTAINTY, ITS REPRESENTATION AND EVALUATION

Tutorial lectures, IMS Singapore, 3-6 May 2004
Revised January 2007*

Kenneth F. Wallis, University of Warwick

*These lectures formed part of the program on Econometric Forecasting and High-Frequency Data Analysis at the Institute for Mathematical Sciences, National University of Singapore, jointly organised by the School of Economics and Social Sciences, Singapore Management University, May 2004; the kind hospitality of both institutions is gratefully acknowledged. The lectures were written up for publication in August 2004. Subsequently some of the material formed the basis of lectures given, under the same general title, as the Sir Richard Stone lectures in London in October 2005. Taking advantage of the passage of time, some of the added material has been updated and incorporated into the present text.

Contents

1. Introduction

Forecasts of future economic outcomes are subject to uncertainty. It is increasingly accepted that forecasters who publish forecasts for the use of the general public should accompany their point forecasts with an indication of the associated uncertainty. These lectures first describe the various available methods of communicating information about forecast uncertainty. It is equally important that forecasters' statements about the underlying uncertainty should be reliable. The lectures go on to consider the various available statistical techniques for assessing the reliability of statements about forecast uncertainty.

The lectures draw on and extend material covered in previous survey articles such as Wallis (1995) and, most notably, Tay and Wallis (2000) on density forecasting. While Tay and Wallis discussed applications in macroeconomics and finance, the present lectures are oriented towards macroeconomics, while other lecturers in this program deal with financial econometrics. Relevant research articles are referenced in full, but background material in statistics, econometrics, and associated mathematical methods is not; readers needing to refer to the general literature are asked to consult their favourite textbooks.

This introduction first motivates the lectures by considering the "why" question – why say anything about forecast uncertainty? – and then presents an overview of the issues to be addressed in the two main sections, based on an introductory theoretical illustration.

1.1 *Motivation*

Why not just give a forecast as a single number, for example, inflation next year will be 2.8%? But what if someone else's inflation forecast is 3.1%, is this an important difference, or is it negligible in comparison to the underlying uncertainty? At the simplest level, to acknowledge the uncertainty that is always present in economic forecasting, and that "we all know" that inflation next year is unlikely to be exactly 2.8%, contributes to better-informed discussion about economic policy and prospects. The central banks of many countries now operate an inflation-targeting monetary policy regime, in which forecasts of inflation play an important part, since monetary policy has a delayed effect on inflation. Uncertainty has a crucial role in policy decisions, and considerations of transparency and its impact on the credibility of policy have led many banks to discuss the "risks to the forecast" in their forecast publications. Some have gone further, as described in detail below, and publish a density forecast of inflation, that is, an estimate of the probability distribution of the possible future values of inflation. This represents a complete description of the uncertainty associated with a forecast.

The decision theory framework provides a more formal justification for the publication of density forecasts as well as point forecasts. The decision theory formulation begins with a loss function $L(d,y)$ that describes the consequences of taking decision d today if the future state variable has the value y. If the future were known, then the optimal decision would be the one that makes L as small as possible. But if the future outcome is uncertain, then the loss is a random variable, and a common criterion is to choose the decision that minimises the expected loss. To calculate the expected value of $L(d,y)$ for a range of values of d, in order to find the minimum, the complete probability distribution of y is needed in general. The special case that justifies restricting attention to a point forecast is the case in which L is a quadratic function of y. In this case the certainty equivalence theorem states that the value of d that minimises expected loss $E\left(L(d,y)\right)$ is the same as the value that minimises $L\left(d,E(y)\right)$, whatever the distribution of y might be. So in this case only a point forecast, specifically the conditional expectation of the unknown future state variable, is required. In practice, however, macroeconomic forecasters have little knowledge of the identity of the users of forecasts, not to mention their loss functions, and the assumption that these are all quadratic is unrealistic. In many situations the possibility of an unlimited loss is also unrealistic, and bounded loss functions are more reasonable. These are informally referred to as “a miss is as good as a mile” or, quoting Bray and Goodhart (2002), “ you might as well be hung for a sheep as a lamb”. In more general frameworks such as these, decision-makers require the complete distribution of y.

1.2 *Overview*

A theoretical illustration

We consider the simple univariate model with which statistical prediction theory usually begins, namely the Wold moving average representation of a stationary, non-deterministic series:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots , \quad \sum_{j=0}^{\infty} \theta_j^2 < \infty \quad \left(\theta_0 = 1\right)$$

$$E\left(\varepsilon_t\right) = 0, \quad \operatorname{var}\left(\varepsilon_t\right) = \sigma_\varepsilon^2, \quad E\left(\varepsilon_t \varepsilon_s\right) = 0, \quad \text{all } t, s \neq t.$$

To construct a forecast *h* steps ahead, consider this representation at time $t+h$:

$$y_{t+h} = \varepsilon_{t+h} + \theta_1 \varepsilon_{t+h-1} + \ldots + \theta_{h-1} \varepsilon_{t+1} + \theta_h \varepsilon_t + \theta_{h+1} \varepsilon_{t-1} + \ldots \ .$$

The optimal **point forecast** with respect to a squared error loss function, the “minimum mean squared error” (mmse) forecast, is the conditional expectation $E\left(y_{t+h} \mid \Omega_t\right)$, where Ω_t denotes the relevant information set. In the present case this simply comprises available data on the *y*-process at the forecast origin, *t*, hence the mmse *h*-step-ahead forecast is

$$\hat{y}_{t+h} = \theta_h \varepsilon_t + \theta_{h+1} \varepsilon_{t-1} + \ldots ,$$

with forecast error $e_{t+h} = y_{t+h} - \hat{y}_{t+h}$ given as

$$e_{t+h} = \varepsilon_{t+h} + \theta_1 \varepsilon_{t+h-1} + \ldots + \theta_{h-1} \varepsilon_{t+1} \ .$$

The forecast error has mean zero and variance σ_h^2, where

$$\sigma_h^2 = E\left(e_{t+h}^2\right) = \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \theta_j^2 \ .$$

The forecast root mean squared error is defined as $RMSE_h = \sigma_h$. The forecast error is a moving average process and so in general exhibits autocorrelation at all lags up to $h-1$: only the one-step-ahead forecast has a non-autocorrelated error. Finally note that the optimal forecast and its error are uncorrelated:

$$E\left(e_{t+h} \hat{y}_{t+h}\right) = 0 .$$

An **interval forecast** is commonly constructed as the point forecast plus or minus one or two standard errors, $\hat{y}_{t+h} \pm \sigma_h$, for example. To attach a probability to this statement we need a distributional assumption, and a normal distribution for the random shocks is commonly assumed:

$$\varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right).$$

Then the future outcome also has a normal distribution, and the above interval has probability 0.68 of containing it. The **density forecast** of the future outcome is this same distribution, namely

$$y_{t+h} \sim N\left(\hat{y}_{t+h}, \sigma_h^2\right).$$

Example

A familiar example in econometrics texts is the stationary first-order autoregression, abbreviated to AR(1):

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad |\phi| < 1.$$

Then in the moving average representation we have $\theta_j = \phi^j$ and the *h*-step-ahead point forecast is

$$\hat{y}_{t+h} = \phi^h y_t.$$

The forecast error variance is

$$\sigma_h^2 = \sigma_\varepsilon^2 \frac{1-\phi^{2h}}{1-\phi^2}.$$

As *h* increases this approaches the unconditional variance of *y*, namely $\sigma_\varepsilon^2 \big/ \left(1-\phi^2\right)$. Interval and density forecasts are obtained by using these quantities in the preceding expressions.

Generalisations

This illustration uses the simplest univariate linear model and treats its parameters as if their values are known. To have practical relevance these constraints need to be relaxed. Thus in Section 2, where methods of measuring and reporting forecast uncertainty are discussed, multivariate models and non-linear models appear, along with conditioning variables and non-normal distributions, and the effects of parameter estimation error and uncertainty about the model are considered.

Forecast evaluation

Given a time series of forecasts and the corresponding outcomes or realisations $y_t,\ t=1,...,n,$ we have a range of techniques available for the statistical assessment of the quality of the forecasts. For point forecasts these have a long history; for a review, see Wallis (1995, §3). The first question is whether there is any systematic bias in the forecasts, and this is usually answered by testing the null hypothesis that the forecast errors have zero mean, for which a *t*-test is appropriate. Whether the forecasts have minimum mean squared error cannot be tested, because we do not know what the minimum achievable mse is, but other properties of optimal forecasts can be tested. The absence of correlation between errors and forecasts, for example, is often tested in the context of a realisation-forecast regression, and the non-autocorrelation of forecast errors at lags greater than or equal to *h* is also testable. Information can often be gained by comparing different forecasts of the same variable, perhaps in the context of an extended realisation-forecast regression, which is related to the question of the construction of combined forecasts.

Tests of interval and density forecasts, a more recent development, are discussed in Section 3. The first question is one of correct coverage: is the proportion of outcomes falling in the forecast intervals equal to the announced probability; are the quantiles of the forecast densities occupied in the correct proportions? There is also a question of independence, analogous to the non-autocorrelation of the errors of point forecasts. The discussion includes applications to two series of density forecasts of inflation, namely those of the US Survey of Professional Forecasters (managed by the Federal Reserve Bank of Philadelphia, see http://www.phil.frb.org/econ/spf/index.html) and the Bank of England Monetary Policy Committee (as published in the Bank's quarterly *Inflation Report*). Finally some recent extensions to comparisons and combinations of density forecasts are considered, which again echo the point forecasting literature.

Section 4 contains concluding comments.

2. Measuring and reporting forecast uncertainty

We first consider methods of calculating measures of expected forecast dispersion, both model-based and empirical, and then turn to methods of reporting and communicating forecast uncertainty. The final section considers some related issues that arise in survey-based forecasts. For a fully-developed taxonomy of the sources of forecast uncertainty see Clements and Hendry (1998).

2.1 *Model-based measures of forecast uncertainty*

For some models formulae for the forecast error variance are available, and two examples are considered. In other models simulation methods are employed.

The linear regression model

The first setting in which parameter estimation error enters that one finds in econometrics textbooks is the classical linear regression model. The model is

$$y = X\beta + u \ , \quad u \sim N\left(0, \sigma_u^2 I_n\right) .$$

The least squares estimate of the coefficient vector, and its covariance matrix, are

$$b = \left(X'X\right)^{-1} X'y \ , \quad \mathrm{var}(b) = \sigma_u^2 \left(X'X\right)^{-1} .$$

A point forecast conditional on regressor values $c' = [1 \; x_{2f} \;\; x_{3f} \; \ldots \; x_{kf}]$ is $\hat{y}_f = c'b$, and the forecast error has two components

$$e_f = y_f - \hat{y}_f = u_f - c'(b - \beta) .$$

Similarly the forecast error variance has two components

$$\mathrm{var}(e_f) = \sigma_u^2 \left(1 + c'\left(X'X\right)^{-1} c\right).$$

The second component is the contribution of parameter estimation error, which goes to zero as the sample size, *n*, increases (under standard regression assumptions that ensure the convergence of the second moment matrix $X'X/n$). To make this expression operational the unknown error variance σ_u^2 is replaced by an estimate s^2 based on the sum of squared residuals, which results in a shift from the normal to Student's *t*-distribution, and interval and density forecasts are based on the distributional result that

$$\frac{y_f - \hat{y}_f}{s\sqrt{1 + c'\left(X'X\right)^{-1} c}} \sim t_{n-k}.$$

It should be emphasised that this result refers to a forecast that is conditional on given values of the explanatory variables. In practical forecasting situations the future values of deterministic variables such as trends and seasonal dummy variables are known, and perhaps some economic variables such as tax rates can be treated as fixed in short-term forecasting, but in general the future values of the economic variables on the right-hand side of the regression equation need forecasting too. The relevant setting is then one of a multiple-equation model rather than the above single-equation model. For a range of linear multiple-equation models generalisations of the above expressions can be found in the literature. However the essential ingredients of forecast error – future random shocks and parameter estimation error – remain the same.

Estimation error in multi-step forecasts

To consider the contribution of parameter estimation error in multi-step forecasting with a dynamic model we return to the AR(1) example discussed in Section 1.2. Now, however, the point forecast is based on an estimated parameter:

$$\hat{y}_{t+h} = \hat{\phi}^h y_t \,.$$

The forecast error again has two components

$$y_{t+h} - \hat{y}_{t+h} = e_{t+h} + \left(\phi^h - \hat{\phi}^h\right) y_t \,,$$

where the first term is the cumulated random error defined above, namely

$$e_{t+h} = \varepsilon_{t+h} + \phi\varepsilon_{t+h-1} + \ldots + \phi^{h-1}\varepsilon_{t+1} \,.$$

To calculate the variance of the second component we first neglect any correlation between the forecast initial condition y_t and the estimation sample on which $\hat{\phi}$ is based, so that the variance of the product is the product of the variances of the factors. Using the result that the variance of the least squares estimate of ϕ is $\left(1-\phi^2\right)/n$, and taking a first-order approximation to the non-linear function, we then obtain

$$\operatorname{var}\left(\hat{\phi}^h - \phi^h\right) \approx \frac{\left(h\phi^{h-1}\right)^2\left(1-\phi^2\right)}{n} \,.$$

The variance of y_t is $\sigma_\varepsilon^2/\left(1-\phi^2\right)$, hence the forecast error variance is

$$E\left(y_{t+h} - \hat{y}_{t+h}\right)^2 \approx \sigma_\varepsilon^2\left(\frac{1-\phi^{2h}}{1-\phi^2} + \frac{\left(h\phi^{h-1}\right)^2}{n}\right).$$

The second contribution causes possible non-monotonicity of the forecast error variance as *h* increases, but goes to zero as *h* becomes large. As above, this expression is made operational by replacing unknown parameters by their estimates, and the *t*-distribution provides a better approximation for inference than the normal distribution. And again, generalisations can be found in the literature for more complicated dynamic linear models.

Stochastic simulation in non-linear models

Practical econometric models are typically non-linear in variables. They combine log-linear regression equations with linear accounting identities. They include quantities measured in both real and nominal terms together with the corresponding price variables, hence products and ratios of variables appear. More complicated functions such as the constant elasticity of substitution (CES) production function can also be found. In these circumstances an analytic expression for a forecast does not exist, and numerical methods are used to solve the model.

A convenient formal representation of a general non-linear system of equations, in its structural form, is

$$f\left(y_t, z_t, \alpha\right) = u_t \,,$$

where *f* is a vector of functions having as many elements as the vector of endogenous variables y_t, and z_t, α and u_t are vectors of predetermined variables, parameters and random disturbances respectively. This is more general than is necessary, because models are mostly linear in parameters, but no convenient simplification is available. It is assumed that a unique solution for the endogenous variables exists. Whereas multiple solutions might exist from a mathematical point of view, typically only one of them makes sense in the economic context. The solution has no explicit analytic form, but it can be written implicitly as

$$y_t = g\left(u_t, z_t, \alpha\right),$$

which is analogous to the reduced form in the linear case.

Taking period *t* to be the forecast period of interest, the "deterministic" forecast $\hat{y}_t$ is obtained, for given values of predetermined variables and parameters, as the numerical solution to the structural form, with the disturbance terms on the right-hand side set equal to their expected values of zero. The forecast can be written implicitly as

$$\hat{y}_t = g\left(0, z_t, \alpha\right),$$

and is approximated numerically to a specified degree of accuracy.

Forecast uncertainty likewise cannot be described analytically. Instead, stochastic simulation methods are used to estimate the forecast densities. First *R* vectors of pseudo-random numbers $u_{tr},\ r = 1,...,R,$ are generated with the same properties as those assumed or estimated for the model disturbances: typically a normal distribution with covariance matrix estimated from the model residuals. Then for each replication the model is solved for the corresponding values of the endogenous variables y_{tr}, say, where

$$f\left(y_{tr}, z_t, \alpha\right) = u_{tr},\ r = 1,...,R$$

to the desired degree of accuracy, or again implicitly

$$y_{tr} = g\left(u_{tr}, z_t, \alpha\right).$$

In large models attention is usually focused on a small number of key macroeconomic indicators. For the relevant elements of the *y*-vector the empirical distributions of the y_{tr} values then represent their density forecasts. These are presented as histograms, possibly smoothed using techniques discussed by Silverman (1986), for example.

The early applications of stochastic simulation methods focused on the mean of the empirical distribution in order to assess the possible bias in the deterministic forecast. The non-linearity of g is the source of the lack of equality in the following statement,

$$E\left(y_t \middle| z_t, \alpha\right) = E\left(g\left(u_t, z_t, \alpha\right)\right) \neq g\left(E\left(u_t\right), z_t, \alpha\right) = \hat{y}_t \, ,$$

and the bias is estimated as the difference between the deterministic forecast and the simulation sample mean

$$\bar{y}_t = \frac{1}{R}\sum_{r=1}^{R} y_{tr} \; .$$

Subsequently attention moved on to second moments and event probability estimates. With an economic event, such as a recession, defined in terms of a model outcome, such as two consecutive quarters of declining real GDP, then the relative frequency of this outcome in R replications of a multi-step forecast is an estimate of the probability of the event. Developments also include study of the effect of parameter estimation error, by pseudo-random sampling from the distribution of $\hat{\alpha}$ as well as that of u_t . For a fuller discussion, and references, see Wallis (1995, §4).

Loss functions

It is convenient to note the impact of different loss functions at this juncture, in the light of the foregoing discussion of competing point forecasts. The conditional expectation is the optimal forecast with respect to a squared error loss function, as noted above, but in routine forecasting exercises with econometric models one very rarely finds the mean stochastic simulation estimate being used. Its computational burden becomes less of a concern with each new generation of computer, but an alternative loss function justifies the continued use of the deterministic forecast.

In the symmetric linear or absolute error loss function, the optimal forecast is the median of the conditional distribution. (Note that this applies only to forecasts of a single variable, strictly speaking, since there is no standard definition of the median of a multivariate distribution. Commonly, however, this is interpreted as the set of medians of the marginal univariate distributions.) Random disturbances are usually assumed to have a symmetric distribution, so that the mean and median are both zero, hence the deterministic forecast $\hat{y}_t$ is equal to the median of the conditional distribution of y_t provided that the transformation $g(\cdot)$ preserves the median. That is, provided that

$$\mathrm{med}\left(g\left(u_t, z_t, \alpha\right)\right) = g\left(\mathrm{med}\left(u_t\right), z_t, \alpha\right).$$

This condition is satisfied if the transformation is bijective, which is the case for the most common example in practical models, namely the exponential function, whose use arises from the specification of log-linear equations with additive disturbance terms. Under these conditions the deterministic forecast is the minimum absolute error forecast. There is simulation evidence that the median of the distribution of stochastic simulations in practical models either coincides with the deterministic forecast or is very close to it (Hall, 1986).

The third measure of location familiar in statistics is the mode, and in the context of a density forecast the mode represents the most likely outcome. Some forecasters focus on the mode as their preferred point forecast believing that the concept of the most likely outcome is most easily understood by forecast users. It is the optimal forecast under a step or "all-or-nothing" loss function, hence in a decision context in which the loss is bounded or "a miss is as good as a mile", the mode is the best choice of point forecast. Again this applies in a univariate, not multivariate setting: in practice the mode is difficult to compute in the multivariate case (Calzolari and Panattoni, 1990), and it is not preserved under transformation. For random variables X and Y with asymmetric distributions, it is not in general true that the mode of $X+Y$ is equal to the mode of X plus the mode of Y, for example.

Model uncertainty

Model-based estimates of forecast uncertainty are clearly conditional on the chosen model. However the choice of an appropriate model is itself subject to uncertainty. Sometimes the model specification is chosen with reference to an *a priori* view of the way the world works, sometimes it is the result of a statistical model selection procedure. In both cases the possibility that an inappropriate model has been selected is yet another contribution to forecast uncertainty, but in neither case is a measure of this contribution available, since the true data generating process is unknown.

A final contribution to forecast uncertainty comes from the subjective adjustments to model-based forecasts that many forecasters make in practice, to take account of off-model information of various kinds: their effects are again not known with certainty, and measures of this contribution are again not available. In these circumstances some forecasters provide subjective assessments of uncertainty, whereas others turn to *ex post* assessments.

2.2 *Empirical measures of forecast uncertainty*

The historical track record of forecast errors incorporates all sources of error, including model error and the contribution of erroneous subjective adjustments. Past forecast performance thus provides a suitable foundation for measures of forecast uncertainty.

Let $y_t,\ t = 1,...,n$ be an observed time series and $\hat{y}_t,\ t = 1,...,n$ be a series of forecasts of y_t made at times $t - h$, where *h* is the forecast horizon. The forecast errors are then $e_t = y_t - \hat{y}_t,\ t = 1,...,n$. The two conventional summary measures of forecast performance are the sample root mean squared error,

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}\ ,$$

and the sample mean absolute error,

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|\ .$$

The choice between them should in principle be related to the relevant loss function – squared error loss or absolute error loss – although many forecasters report both.

In basing a measure of the uncertainty of future forecasts on past forecast performance we are, of course, facing an additional forecasting problem. Now it is addressed to measures of the dispersion of forecasts, but it is subject to the same difficulties of forecast failure due to structural breaks as point forecasts. Projecting forward from past performance assumes a stable underlying environment, and difficulties arise when this structure changes.

If changes can be anticipated, subjective adjustments might be made, just as is the case with point forecasts, but just as difficult. For example, the UK government publishes alongside its budget forecasts of key macroeconomic indicators the mean absolute error of the past ten years' forecasts. The discussion of the margin of error of past forecasts in the statement that accompanied the June 1979 budget, immediately following the election of Mrs Thatcher's first government, noted the "possibility that large changes in policy will affect the economy in ways which are not foreseen".

A more recent example is the introduction in several countries of a monetary policy regime of direct inflation targeting. It is claimed that this will reduce uncertainty, hence the "old regime" forecast track record may be an unreliable guide to the future uncertainty of inflation. Eventually a track record on the new regime will accumulate, but measures of uncertainty are needed in the meantime. One way to calibrate the variance of inflation in the new regime is to undertake a stochastic simulation study of the performance of a macroeconometric model augmented with a policy rule for the interest rate that targets inflation. Blake (1996) provides a good example, using the model of the UK economy maintained by the National Institute of Economic and Social Research. He finds that inflation uncertainty is indeed reduced in the new regime, although his estimates are of course conditional on the specification of the model and the policy rule. In particular the latter implies the vigorous use of interest rates to achieve the inflation target in the face of shocks, and the price to pay for a stable inflation rate may be higher interest rate variability.

2.3 *Reporting forecast uncertainty*

Interval forecasts and density forecasts are discussed in turn, including some technical considerations. The Bank of England's fan chart is a leading graphical representation, and other examples are discussed.

Forecast intervals

An interval forecast is commonly presented as a range centred on a point forecast, as noted in the Introduction, with associated probabilities calculated with reference to tables of the normal distribution. Then some typical interval forecasts and their coverage probabilities are

$\hat{y}_t \pm MAE$	57%	
$\hat{y}_t \pm RMSE$	68%	
$\hat{y}_t \pm 2RMSE$	95%	
$\hat{y}_t \pm 0.675RMSE$	50%	(the interquartile range).

In more complicated models other distributions are needed, as noted above. If parameter estimation errors are taken into account then Student's *t*-distribution is relevant, whereas in complex non-linear models the forecast distribution may have been estimated non-parametrically by stochastic simulation. In the latter case the distribution may not be symmetric, and a symmetric interval centred on a point forecast may not be the best choice. In any event it can be argued that to focus on uncertainty the point forecast should be suppressed, and only the interval reported.

The requirement that an interval (a, b) be constructed so that it has a given probability π of containing the outcome *y*, that is,

$$\Pr\left(a \le y \le b\right) = F(b) - F(a) = \pi$$

where $F(\cdot)$ is the cumulative distribution function, does not by itself pin down the location of the interval. Additional specification is required, and the question is what is the best choice. The forecasting literature assumes unimodal densities and considers two possible specifications, namely the shortest interval, with $b - a$ as small as possible, and the central interval, which contains the stated probability in the centre of the distribution, defined such that there is equal probability in each of the tails:

$$\Pr\left(y < a\right) = \Pr\left(y > b\right) = \left(1 - \pi\right)/2 .$$

(This usage of “central” is in accordance with the literature on confidence intervals, see Stuart, Ord and Arnold, 1999, p.121.) If the distribution of outcomes is symmetric then the two intervals are the same; if the distribution is asymmetric then the shortest and central intervals do not coincide. Each can be justified as the optimal interval forecast with respect to a particular loss function or cost function, as we now show.

It is assumed that there is a cost proportional to the length of the interval, $c_0(b - a)$, which is incurred irrespective of the outcome. The distinction between the two cases arises from the assumption about the additional cost associated with the interval not containing the outcome.

All-or-nothing loss function If the costs associated with the possible outcomes have an all-or-nothing form, being zero if the interval contains the outcome and a constant $c_1 > 0$ otherwise, then the loss function is

$$L(y) = \begin{cases} c_0(b-a)+c_1 & y < a \\ c_0(b-a) & a \le y \le b \\ c_0(b-a)+c_1 & y > b \end{cases}$$

The expected loss is

$$\begin{aligned} E\left(L(y)\right) &= c_0(b-a) + \int_{-\infty}^{a} c_1 f(y)dy + \int_{b}^{\infty} c_1 f(y)dy \\ &= c_0(b-a) + c_1 F(a) + c_1\left(1 - F(b)\right). \end{aligned}$$

To minimise expected loss subject to the correct interval probability consider the Lagrangean

$$\mathcal{L} = c_0(b-a) + c_1 F(a) + c_1\left(1-F(b)\right) + \lambda\left(F(b) - F(a) - \pi\right).$$

The first-order conditions with respect to *a* and *b* give

$$f(a) = f(b) = c_0 \big/ \left(c_1 - \lambda\right),$$

thus for given coverage the limits of the optimal interval correspond to ordinates of the probability density function (pdf) of equal height on either side of the mode. As the coverage is reduced, the interval closes in on the mode of the distribution.

The equal height property is also a property of the interval with shortest length $b-a$ for given coverage π. To see this consider the Lagrangean

$$\mathcal{L} = b - a + \lambda\left(F(b) - F(a) - \pi\right).$$

This is a special case of the expression considered above, and the first-order conditions for a minimum again give $f(a) = f(b)$ as required. The shortest interval has unequal tail probabilities in the asymmetric case, and these should be reported, in case a user might erroneously think that they are equal.

Linear loss function Here it is assumed that the additional cost is proportional to the amount by which the outcome lies outside the interval. Thus the loss function is

$$L(y) = \begin{cases} c_0(b-a) + c_2(a-y) & y < a \\ c_0(b-a) & a \le y \le b \\ c_0(b-a) + c_2(y-b) & y > b \end{cases}$$

The expected loss is

$$E\left(L(y)\right) = c_0(b-a) + \int_{-\infty}^{a} c_2(a-y) f(y)dy + \int_{b}^{\infty} c_2(y-b) f(y)dy$$

and the first-order conditions for minimum expected loss give

$$c_0 = c_2 \int_{-\infty}^{a} f(y)dy = c_2 \int_{b}^{\infty} f(y)dy\,.$$

Hence the best forecast interval under a linear loss function is the central interval with equal tail probabilities $\Pr(y < a) = \Pr(y > b)$, the limits being the corresponding quantiles

$$a = F^{-1}\left(\frac{1-\pi}{2}\right), \quad b = F^{-1}\left(\frac{1+\pi}{2}\right).$$

As the coverage, π, is reduced, the central interval converges on the median.

In some applications a pre-specified interval may be a focus of attention. In a monetary policy regime of inflation targeting, for example, the objective of policy is sometimes expressed as a target range for inflation, whereupon it is of interest to report the forecast probability that the future outcome will fall in the target range. This is equivalent to an event probability forecasting problem, the forecast being stated as the probability of the future event “inflation on target” occurring.

Density forecasts

The preceding discussion includes cases where the density forecast has a known functional form and cases where it is estimated by non-parametric methods. In the former case features of the forecast may not be immediately apparent from an algebraic expression for the density, and in both cases numerical presentations are used, either as histograms, with intervals of equal length, or based on quantiles of the distribution. In the present context the conventional discretisation of a distribution based on quantiles amounts to representing the density forecast as a set of central forecast intervals with different coverage probabilities. Graphical presentations are widespread, but before discussing them we present a further density function that is used to represent forecast uncertainty, particularly when the balance of risks to the forecast is asymmetric.

The density forecasts of inflation published by the Bank of England and the Sveriges Riksbank assume the functional form of the two-piece normal distribution (Blix and Sellin, 1998; Britton, Fisher and Whitley, 1998). A random variable *X* has a two-piece normal distribution with parameters μ, σ_1 and σ_2 if it has probability density function (pdf)

$$f(x)=\begin{cases} A\exp\left(-(x-\mu)^2 / 2\sigma_1^2\right) & x \le \mu \\ A\exp\left(-(x-\mu)^2 / 2\sigma_2^2\right) & x \ge \mu \end{cases}$$

where $A=\left(\sqrt{2\pi}\,(\sigma_1+\sigma_2)/2\right)^{-1}$ (John, 1982; Johnson, Kotz and Balakrishnan, 1994; Wallis, 1999). The distribution is formed by taking the left half of a normal distribution with parameters $(\mu,\ \sigma_1)$ and the right half of a normal distribution with parameters $(\mu,\ \sigma_2)$, and scaling them to give the common value $f(\mu)=A$ at the mode, as above. An illustration is presented in Figure 1. The scaling factor applied to the left half of the $N(\mu,\sigma_1)$ pdf is $2\sigma_1/(\sigma_1+\sigma_2)$ while that applied to the right half of the $N(\mu,\sigma_2)$ pdf is $2\sigma_2/(\sigma_1+\sigma_2)$. If $\sigma_2>\sigma_1$ this reduces the probability mass to the left of the mode to below one-half and correspondingly increases the probability mass above the mode, hence in this case the two-piece normal distribution is positively skewed with mean>median>mode. Likewise, when $\sigma_1>\sigma_2$ the distribution is negatively skewed. The mean and variance of the distribution are

$$E(X)=\mu+\sqrt{\frac{2}{\pi}}(\sigma_2-\sigma_1)$$

$$\operatorname{var}(X)=\left(1-\frac{2}{\pi}\right)(\sigma_2-\sigma_1)^2+\sigma_1\sigma_2\ .$$

The two-piece normal distribution is a convenient representation of departures from the symmetry of the normal distribution, since probabilities can be readily calculated by referring to standard normal tables and scaling by the above factors; however, the asymmetric distribution has no convenient multivariate generalisation.

In the case of the Bank of England, the density forecast describes the subjective assessment of inflationary pressures by its Monetary Policy Committee, and the three parameters are calibrated to represent this judgement, expressed in terms of the location, scale and skewness of the distribution. A point forecast – mean and/or mode – fixes the location of the distribution. The level of uncertainty or scale of the distribution is initially assessed with reference to forecast errors over the preceding ten years, and is then adjusted with respect to known or anticipated future developments. The degree of skewness, expressed in terms of the difference between the mean and the mode, is determined by the Committee's collective assessment of the balance of risks on the upside and downside of the forecast.

Graphical presentations

In real-time forecasting, a sequence of forecasts for a number of future periods from a fixed initial condition (the "present") is often presented as a time-series plot. The point forecast may be shown as a continuation of the plot of actual data recently observed, and limits may be attached, either as standard error bands or quantiles, becoming wider as the forecast horizon increases. Thompson and Miller (1986) note that "typically forecasts and limits are graphed as dark lines on a white background, which tends to make the point forecast the focal point of the display." They argue for and illustrate the use of selective shading of quantiles, as "a deliberate attempt to draw attention away from point forecasts and toward the *uncertainty* in forecasting" (1986, p. 431, emphasis in original).

In presenting its density forecasts of inflation the Bank of England takes this argument a stage further, by suppressing the point forecast. The density forecast is presented graphically as a set of forecast intervals covering 10, 20, 30,…, 90% of the probability distribution, of lighter shades for the outer bands. This is done for quarterly forecasts up to two years ahead, and since the dispersion increases and the intervals "fan out" as the forecast horizon increases, the result has become known as the "fan chart". Rather more informally, and noting its red colour, it also became known as the "rivers of blood". (In their recent textbook, Stock and Watson (2003) refer to the fan chart using only the "river of blood" title; since their reproduction is coloured green, readers are invited to use their imagination.)

An example of the Bank of England's presentation of the density forecasts is shown in Figure 2. This uses the shortest intervals for the assigned probabilities, which converge on the mode. (The calibrated parameter values for the final quarter's forecast are also used in the illustration of the two-piece normal distribution in Figure 1.) As the distribution is asymmetric the probabilities in the upper and lower same-shade segments are not equal. The Bank does not report the consequences of this, which are potentially misleading. For the final quarter Wallis (1999, Table 1) calculates the probability of inflation lying below the darkest 10% interval as 32½%, and correspondingly a probability of 57½% that inflation will lie above the middle 10% interval. Visual inspection of the fan chart does not by itself reveal the extent of this asymmetry. Similarly the lower and upper tail probabilities in the final quarter are 3.6% and 6.4% respectively.

An alternative presentation of the same density forecasts by Wallis (1999) is shown in Figure 3: this uses central intervals defined by percentiles, with equal tail probabilities, as discussed above. There is no ambiguity about the probability content of the upper and lower bands of a given shade: they are all 5%, as are the tail probabilities. It is argued that a preference for this alternative fan chart is implicit in the practice of the overwhelming majority of statisticians of summarising densities by presenting selected percentiles.

Additional examples

We conclude this section by describing three further examples of the reporting of forecast uncertainty by the use of density forecasts. First is the National Institute of Economic and Social Research in London, England, which began to publish density forecasts of inflation and GDP growth in its quarterly *National Institute Economic Review* in February 1996, the same month in which the Bank of England's fan chart first appeared. The forecast density is assumed to be a normal distribution centred on the point forecast, since the hypothesis of unbiased forecasts with normally distributed errors could not be rejected in testing the track record of earlier forecasts. The standard deviation of the normal distribution is set equal to the standard deviation of realised forecast errors at the same horizon over a previous period. The distribution is presented as a histogram, in the form of a table reporting the probabilities of outcomes falling in various intervals. For inflation, those used in 2004, for example, were: less than 1.5%, 1.5 to 2.0%, 2.0 to 2.5%, and so on.

A second example is the budget projections prepared by the Congressional Budget Office (CBO) of the US Congress. Since January 2003 the uncertainty of the CBO's projections of the budget deficit or surplus under current policies has been represented as a fan chart. The method of construction of the density forecast is described in CBO (2003); in outline it follows the preceding paragraph, with a normal distribution calibrated to the historical record. On the CBO website (www.cbo.gov) the fan chart appears in various shades of blue.

Our final example is the work of Garratt, Lee, Pesaran and Shin (2003). They have previously constructed an eight-equation conditional vector error-correction model of the UK economy. In the present article they develop density and event probability forecasts for inflation and growth, singly and jointly, based on this model. These are computed by stochastic simulation allowing for parameter uncertainty. The density forecasts are presented by plotting the estimated cumulative distribution function at three forecast horizons.

2.4 *Forecast scenarios*

Variant forecasts that highlight the sensitivity of the central forecast to key assumptions are commonly published by forecasting agencies. The US Congressional Budget Office (2004), for example, presents in addition to its baseline budget projections variants that assume lower real growth, higher interest rates or higher inflation. The Bank of England has on occasion shown the sensitivity of its central projection for inflation to various alternative assumptions preferred by individual members of the Monetary Policy Committee: with respect to the behaviour of the exchange rate, the scale of the slowdown in the global economy, and the degree of spare capacity in the domestic economy, for example. The most highly developed and documented use of forecast scenarios is that of the CPB Netherlands Bureau for Economic Policy Analysis, which is a good example for fuller discussion.

Don (2001), who was CPB Director 1994-2006, describes the CPB's practice of publishing a small number of scenarios rather than a single forecast, arguing that this communicates forecast uncertainty more properly than statistical criteria for forecast quality, since "*ex post* forecast errors can at best provide a rough guide to *ex ante* forecast errors". Periodically the CPB publishes a medium-term macroeconomic outlook for the Dutch economy over the next Cabinet period, looking four or five years ahead. The outlook is the basis for the CPB's analysis of the platforms of the competing parties at each national election, and for the programme of the new Cabinet. It comprises two scenarios, which in the early outlooks were termed "favourable" and "unfavourable" in relation to the exogenous assumptions supplied to the model of the domestic economy. "The idea is that these scenarios show between which margins economic growth in the Netherlands for the projection period is likely to lie, barring extreme conditions. There is no numerical probability statement; rather the flavour is informal and subjective, but coming from independent experts" (Don, 2001, p.172). The first sentence of this quotation almost describes an interval forecast, but the word “likely” is not translated into a probability statement, as noted in the second sentence.

The practical difficulty facing the user of these scenarios is not knowing where they lie in the complete distribution of possible outcomes. What meaning should be attached to the words “favourable” and “unfavourable”? And how likely is “likely”? Indeed, in 2001, following a review, the terminology was changed to “optimistic” and “cautious”. The change was intended to indicate that the range of the scenarios had been reduced, so that “optimistic” is less optimistic than “favourable” and “cautious” is less pessimistic than “unfavourable”. It was acknowledged that this made the probability of the actual outcome falling outside the bands much larger, but no quantification was given. (A probability range for potential GDP growth, a key element of the scenarios, can be found in Huizinga (2001), but no comparable estimate of actual outcomes.) All the above terminology lacks precision and is open to subjective interpretation, and ambiguity persists in the absence of a probability statement. Its absence also implies that *ex post* evaluation of the forecasts can only be undertaken descriptively, and that no systematic statistical evaluation is possible.

The objections in the preceding two sentences apply to all examples of the use of scenarios in an attempt to convey uncertainty about future outcomes. How to assess the reliability of statements about forecast uncertainty, assuming that these are quantitative, not qualitative, is the subject of Section 3 below.

2.5 *Uncertainty and disagreement in survey forecasts*

In the absence of direct measures of future uncertainty, early researchers turned to the surveys of forecasters that collected their point forecasts, and suggested that the disagreement among forecasters invariably observed in such surveys might serve as a useful proxy measure of uncertainty. In 1968 the survey now known as the Survey of Professional Forecasters (SPF) was inaugurated, and since this collects density forecasts as well as point forecasts in due course it allowed study of the relationship between direct measures of uncertainty and such proxies, in a line of research initiated by Zarnowitz and Lambros (1987) that remains active to the present time.

The SPF represents the longest-running series of density forecasts in macroeconomics, thanks to the agreement of the Business and Economic Statistics Section of the American Statistical Association and the National Bureau of Economic Research jointly to establish a quarterly survey of macroeconomic forecasters in the United States, originally known as the ASA-NBER survey. Zarnowitz (1969) describes its objectives, and discusses the first results. In 1990 the Federal Reserve Bank of Philadelphia assumed responsibility for the survey, and changed its name to the Survey of Professional Forecasters. Survey respondents are asked not only to report their point forecasts of several variables, but also to attach a probability to each of a number of preassigned intervals, or bins, into which future GNP growth and inflation might fall. In this way, respondents provide their density forecasts of these two variables, in the form of histograms. The probabilities are then averaged over respondents to obtain the mean or aggregate density forecasts, again in the form of histograms, and these are published on the Bank's website. A recent example is shown in Table 1.

Zarnowitz and Lambros (1987) define "consensus" as the degree of agreement among point forecasts of the same variable by different forecasters, and "uncertainty" as the dispersion of the corresponding probability distributions. Their emphasis on the distinction between them was motivated by several previous studies in which high dispersion of point forecasts had been interpreted as indicating high uncertainty, as noted above. Access to a direct measure of uncertainty now provided the opportunity for Zarnowitz and Lambros to check this presumption, among other things. Their definitions are made operational by calculating time series of: (a) the mean of the standard deviations calculated from the individual density forecasts, and (b) the standard deviations of the corresponding sets of point forecasts, for two variables and four forecast horizons. As the strict sense of "consensus" is unanimous agreement, we prefer to call the second series a measure of disagreement. They find that the uncertainty (a) series are typically larger and more stable than the disagreement (b) series, thus measures of uncertainty based on the forecast distributions "should be more dependable". The two series are positively correlated, however, hence in the absence of direct measures of uncertainty a measure of disagreement among point forecasts may be a useful proxy.

A formal relationship among measures of uncertainty and disagreement can be obtained as follows. Denote *n* individual density forecasts of a variable *y* at some future time as $f_i(y),\ i=1,...,n$. In the

SPF these are expressed numerically, as histograms, but the statistical framework also accommodates density forecasts that are expressed analytically, for example, via the normal or two-piece normal distributions. For economy of notation time subscripts and references to the information sets on which the forecasts are conditioned are suppressed. The published mean or aggregate density forecast is then

$$f_A(y) = \frac{1}{n}\sum_{i=1}^{n} f_i(y),$$

which is an example of a finite mixture distribution. The finite mixture distribution is well known in the statistical literature, though not hitherto in the forecasting literature; it provides an appropriate statistical model for a combined density forecast. (Note that in this section *n* denotes the size of a cross-section sample, whereas elsewhere it denotes the size of a time-series sample. We do not explicitly consider panel data at any point, so the potential ambiguity should not be a problem.)

The moments about the origin of $f_A(y)$ are given as the same equally-weighted sum of the moments about the origin of the individual densities. We assume that the individual point forecasts are the means of the individual forecast densities and so denote these means as $\hat{y}_i$; the individual variances are σ_i^2. Then the mean of the aggregate density is

$$\mu_1' = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \hat{y}_A,$$

namely the average point forecast, and the second moment about the origin is

$$\mu_2' = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i^2 + \sigma_i^2\right).$$

Hence the variance of f_A is

$$\sigma_A^2 = \mu_2' - \mu_1'^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \hat{y}_A\right)^2 .$$

This expression decomposes the variance of the aggregate density, σ_A^2, a possible measure of collective uncertainty, into the average individual uncertainty or variance, plus a measure of the dispersion of, or disagreement between, the individual point forecasts. The two components are analogous to the measures of uncertainty and disagreement calculated by Zarnowitz and Lambros, although their use of standard deviations rather than variances breaks the above equation; in any event Zarnowitz and Lambros seem unaware of their role in decomposing the variance of the aggregate distribution. The decomposition lies behind more recent analyses of the SPF data, by Giordani and Soderlind (2003), for example, although their statistical framework seems less appropriate. The choice of measure of collective uncertainty – the variance of the aggregate density forecast or the average individual variance – is still under discussion in the recent literature. (*Note*. This use of the finite mixture distribution was first presented in the May 2004 lectures, then extended in an article in a special issue of the *Oxford Bulletin of Economics and Statistics*; see Wallis, 2005.)

3. Evaluating interval and density forecasts

Decision theory considerations suggest that forecasts of all kinds should be evaluated in a specific decision context, in terms of the gains and losses that resulted from using the forecasts to solve a sequence of decision problems. As noted above, however, macroeconomic forecasts are typically published for general use, with little knowledge of users' specific decision contexts, and their evaluation is in practice based on their statistical performance. How this is done is the subject of this section, which considers interval and density forecasts in turn, and includes two applications.

3.1 *Likelihood ratio tests of interval forecasts*

Given a time series of interval forecasts with announced probability π that the outcome will fall within the stated interval, *ex ante*, and the corresponding series of observed outcomes, the first question is whether this coverage probability is correct *ex post*. Or, on the other hand, is the relative frequency with which outcomes were observed to fall inside the interval significantly different from π? If in *n* observations there are n_1 outcomes falling in their respective forecast intervals and n_0 outcomes falling outside, then the *ex post* coverage is $p = n_1/n$. From the binomial distribution the likelihood under the null hypothesis is

$$L(\pi) \propto (1-\pi)^{n_0}\,\pi^{n_1},$$

and the likelihood under the alternative hypothesis, evaluated at the maximum likelihood estimate *p*, is

$$L(p) \propto \left(1-p\right)^{n_0} p^{n_1} .$$

The likelihood ratio test statistic $-2\,\log\left(L(\pi)/L(p)\right)$ is denoted $\mathrm{LR}_{\mathrm{uc}}$ by Christoffersen (1998), and is then

$$\mathrm{LR}_{\mathrm{uc}} = 2\left(n_0\ \log(1-p)/(1-\pi) + n_1\ \log\left(p/\pi\right)\right) .$$

It is asymptotically distributed as chi-squared with one degree of freedom, denoted χ_1^2, under the null hypothesis.

The $\mathrm{LR}_{\mathrm{uc}}$ notation follows Christoffersen's argument that this is a test of *unconditional* coverage, and that this is inadequate in a time-series context. He defines an efficient sequence of interval forecasts as one which has correct *conditional* coverage and develops a likelihood ratio test of this hypothesis, which combines the test of unconditional coverage with a test of independence. This supplementary hypothesis is directly analogous to the requirement of lack of autocorrelation of orders greater than or equal to the forecast lead time in the errors of a sequence of efficient point forecasts. It is implemented in a two-state (the outcome lies in the interval or not) Markov chain, as a likelihood ratio test of the null hypothesis that successive observations are statistically independent, against the alternative hypothesis that the observations are from a first-order Markov chain.

A test of independence against a first-order Markov chain alternative is based on the matrix of transition counts $[n_{ij}]$, where n_{ij} is the number of observations in state i at time t–1 and j at time t. The maximum likelihood estimates of the transition probabilities are the cell frequencies divided by the corresponding row totals. For an interval forecast there are two states – the outcome lies inside or outside the interval – and these are denoted 1 and 0 respectively. The estimated transition probability matrix is

$$P = \begin{bmatrix} 1-p_{01} & p_{01} \\ 1-p_{11} & p_{11} \end{bmatrix} = \begin{bmatrix} n_{00}/n_{0\cdot} & n_{01}/n_{0\cdot} \\ n_{10}/n_{1\cdot} & n_{11}/n_{1\cdot} \end{bmatrix},$$

where replacing a subscript with a dot denotes that summation has been taken over that index. The likelihood evaluated at *P* is

$$L(P) \propto \left(1-p_{01}\right)^{n_{00}} p_{01}^{n_{01}} \left(1-p_{11}\right)^{n_{10}} p_{11}^{n_{11}} .$$

The null hypothesis of independence is that the state at *t* is independent of the state at *t*–1, that is, $\pi_{01} = \pi_{11}$, and the maximum likelihood estimate of the common probability is $p = n_{\cdot 1}/n$. The likelihood under the null, evaluated at *p*, is

$$L(p) \propto \left(1-p\right)^{n_{\cdot 0}} p^{n_{\cdot 1}} .$$

This is identical to *L*(*p*) defined above if the first observation is ignored. The likelihood ratio test statistic is then

$$\mathrm{LR}_{\mathrm{ind}} = -2 \log\left(L(p)/L(P)\right)$$

which is asymptotically distributed as χ_1^2 under the independence hypothesis.

Christoffersen proposes a likelihood ratio test of conditional coverage as a joint test of unconditional coverage and independence. It is a test of the original null hypothesis against the alternative hypothesis of the immediately preceding paragraph, and the test statistic is

$$\mathrm{LR}_{\mathrm{cc}} = -2 \log\left(L(\pi)/L(P)\right) .$$

Again ignoring the first observation the test statistics obey the relation

$$\mathrm{LR}_{\mathrm{cc}} = \mathrm{LR}_{\mathrm{uc}} + \mathrm{LR}_{\mathrm{ind}} .$$

Asymptotically LR$_{\mathrm{cc}}$ has a χ_2^2 distribution under the null hypothesis. The alternative hypothesis for LR$_{\mathrm{ind}}$ and LR$_{\mathrm{cc}}$ is the same, and these tests form an ordered nested sequence.

3.2 *Chi-squared tests of interval forecasts*

It is well known that the likelihood ratio tests for such problems are asymptotically equivalent to Pearson's chi-squared goodness-of-fit tests. For general discussion and proofs, and references to earlier literature, see Stuart, Ord and Arnold (1999, ch 25). In discussing this equivalence for the Markov chain tests they develop, Anderson and Goodman (1957) note that the chi-squared tests, which are of the form used in contingency tables, have the advantage that "for many users of these methods, their motivation and their application seem to be simpler". This point of view leads Wallis (2003) to explore the equivalent chi-squared tests for interval forecasts, and their extension to density forecasts.

To test the unconditional coverage of interval forecasts, the chi-squared statistic that is asymptotically equivalent to LR$_{\mathrm{uc}}$ is the square of the standard normal test statistic of a sample proportion, namely

$$X^2 = n\left(p-\pi\right)^2 \big/ \pi\left(1-\pi\right).$$

The asymptotic result rests on the asymptotic normality of the binomial distribution of the observed frequencies, and in finite samples an exact test can be based on the binomial distribution.

For testing independence, the chi-squared test of independence in a 2×2 contingency table is asymptotically equivalent to $\mathrm{LR}_{\mathrm{ind}}$. Denoting the matrix $[n_{ij}]$ of observed frequencies alternatively as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

the statistic has the familiar expression

$$X^2 = \frac{n\left(ad - bc\right)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Equivalently, it is the square of the standard normal test statistic for the equality of two binomial proportions. In finite samples computer packages such as StatXact are available to compute exact *P*-values, by enumerating all possible tables that give rise to a value of the test statistic greater than or equal to that observed, and cumulating their null probabilities.

Finally for the conditional coverage joint test, the asymptotically equivalent chi-squared test compares the observed contingency table with the expected frequencies under the joint null hypothesis of row independence and correct coverage probability π. In the simple formula for Pearson's statistic memorised by multitudes of students, $\Sigma(O–E)^2/E$, the observed (*O*) and expected (*E*) frequencies are, respectively,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} (1-\pi)(a+b) & \pi(a+b) \\ (1-\pi)(c+d) & \pi(c+d) \end{bmatrix}.$$

The test has two degrees of freedom since the column proportions are specified by the hypothesis under test and not estimated. The statistic is equal to the sum of the squares of two standard normal test statistics of sample proportions, one for each row of the table. Although the chi-squared statistics for the separate and joint hypotheses are asymptotically equivalent to the corresponding likelihood ratio statistics, in finite samples they obey the additive relation satisfied by the LR statistics only approximately, and not exactly.

To illustrate the two approaches to testing we consider the data on the SPF mean density forecasts of inflation, 1969-1996, analysed by Diebold, Tay and Wallis (1999) and used by Wallis (2003) to illustrate the chi-squared tests. The series of forecasts and outcomes are shown in Figure 4. The density forecasts are represented by box-and-whisker plots, the box giving the interquartile range and the whiskers the 10th and 90th percentiles; these are obtained by linear interpolation of the published histograms. For the present purpose we treat the interquartile range as the relevant interval forecast. Taking the first observation as the initial condition for the transition counts leaves 27 further observations, of which 19 lie inside the box and 8 outside. Christoffersen's $\mathrm{LR}_{\mathrm{uc}}$ statistic is equal to 4.61, and Pearson's chi-squared statistic is equal to 4.48. The asymptotic critical value at the 5% level is 3.84, hence the null hypothesis of correct coverage, unconditionally, with $\pi = 0.5$, is rejected.

The matrix of transition counts is

$$\begin{bmatrix} 5 & 4 \\ 3 & 15 \end{bmatrix}$$

which yields values of the $\mathrm{LR}_{\mathrm{ind}}$ and X^2 statistics of 4.23 and 4.35 respectively. Thus the null hypothesis of independence is rejected. Finally, summing the two likelihood ratio statistics gives the value 8.84 for $\mathrm{LR}_{\mathrm{cc}}$, whereas the direct chi-squared statistic of the preceding paragraph is 8.11, which illustrates the lack of additivity among the chi-squared statistics. Its exact *P*-value in the two binomial proportions model is 0.018, indicating rejection of the conditional coverage joint hypothesis. Overall the two asymptotically equivalent approaches give different values of the test statistics in finite samples, but in this example they are not sufficiently different to result in different conclusions.

3.3 *Extension to density forecasts*

For interval forecasts the calibration of each tail may be of interest, to check the estimation of the balance of risks to the forecast. If the forecast is presented as a central interval, with equal tail probabilities, then the expected frequencies under the null hypothesis of correct coverage are $n(1-\pi)/2$, $n\pi$, $n(1-\pi)/2$ respectively, and the chi-squared statistic comparing these with the observed frequencies has two degrees of freedom.

This is a step towards goodness-of-fit tests for complete density forecasts, where the choice of the number of classes, *k*, into which to divide the observed outcomes is typically related to the size of the sample. The conventional answer to the question of how class boundaries should be determined is to use equiprobable classes, so that the expected class frequencies under the null hypothesis are equal, at *n*/*k*. With observed class frequencies n_i, i=1,...,k, $\Sigma n_i = n$, the chi-squared statistic for testing goodness-of-fit is

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{(n/k)} \quad .$$

It has a limiting χ^2_{k-1} distribution under the null hypothesis.

The asymptotic distribution of the test statistic rests on the asymptotic *k*-variate normality of the multinomial distribution of the observed frequencies. Placing these in the $k \times 1$ vector *x*, under the null hypothesis this has mean vector $\mu = (n/k, ..., n/k)$ and covariance matrix

$$V = (n/k)\left[I - ee'/k\right],$$

where *e* is a $k \times 1$ vector of ones. The covariance matrix is singular, with rank $k-1$. Defining its generalised inverse V^-, the limiting distribution of the quadratic form $(x-\mu)'V^-(x-\mu)$ is then χ^2_{k-1} (Pringle and Rayner, 1971, p.78). Since the above matrix in square brackets is symmetric and idempotent it coincides with its generalised inverse, and the chi-squared statistic given in the preceding paragraph is equivalently written as

$$X^2 = (x-\mu)'\left[I - ee'/k\right](x-\mu)\Big/(n/k)$$

(note that $e'(x-\mu) = 0$). There exists a $(k-1) \times k$ transformation matrix *A* such that (Rao and Rao, 1998, p.252)

$$AA' = I, \quad A'A = \left[I - ee'/k\right].$$

Hence defining $y = A(x-\mu)$ the statistic can be written as an alternative sum of squares

$$X^2 = y'y\Big/(n/k)$$

where the $k-1$ components $y_i^2\Big/(n/k)$ are independently distributed as χ^2_1 under the null hypothesis.

Anderson (1994) introduces this decomposition in order to focus on particular characteristics of the distribution of interest. For example, with *k*=4 and

$$A = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

the three components focus in turn on departures from the null distribution with respect to location, scale and skewness. Such decompositions are potentially more informative about the nature of departures from the null distribution than the single “portmanteau” goodness-of-fit statistic. Anderson (1994) claims that the decomposition also applies in the case of non-equiprobable classes, but Boero, Smith and Wallis (2004) show that this is not correct. They also show how to construct the matrix *A* from Hadamard matrices.

The test of independence of interval forecasts in the Markov chain framework generalises immediately to density forecasts grouped into *k* classes. However the matrix of transition counts is now $k \times k$, and with sample sizes that are typical in macroeconomic forecasting this matrix is likely to be sparse once *k* gets much beyond 2 or 3, the values relevant to interval forecasts. The investigation of possible higher-order dependence becomes even less practical in the Markov chain approach, since the dimension of the transition matrix increases with the square of the order of the chain. In these circumstances other approaches based on transformation rather than grouping of the data are more useful, as discussed next.

3.4 *The probability integral transformation*

The chi-squared goodness-of-fit tests suffer from the loss of information caused by grouping of the data. The leading alternative tests of fit all make use, directly or indirectly, of the probability integral transform. In the present context, if a forecast density $f(y)$ with corresponding distribution function $F(y)$ is correct, then the transformed variable

$$u = \int_{-\infty}^{y} f(x)dx = F(y)$$

is uniformly distributed on (0,1). For a sequence of one-step-ahead forecasts $f_{t-1}(y)$ and corresponding outcomes y_t, a test of fit can then be based on testing the departure of the sequence $u_t = F_{t-1}(y_t)$ from uniformity. Intuitively, the *u*-values tell us in which percentiles of the forecast densities the outcomes fell, and we should expect to see all the percentiles occupied equally in a long run of correct probability forecasts. The advantage of the transformation is that, in order to test goodness-of-fit, the "true" density does not have to be specified.

Diebold, Gunther and Tay (1998), extending the perspective of Christoffersen (1998) from interval forecasts to density forecasts, show that if a sequence of density forecasts is correctly conditionally calibrated, then the corresponding *u*-sequence is iid *U*(0,1). They present histograms of *u* for visual assessment of unconditional uniformity, and various autocorrelation tests.

A test of goodness-of-fit that does not suffer the disadvantage of grouping can be based on the sample cumulative distribution function of the *u*-values. The distribution function of the *U*(0,1) distribution is a 45-

degree line, and the Kolmogorov-Smirnov test is based on the maximum absolute difference between this null distribution function and the sample distribution function. Miller (1956) provides tables of critical values for this test. It is used by Diebold, Tay and Wallis (1999) in their evaluation of the SPF mean density forecasts of inflation. As in most classical statistics, the test is based on an assumption of random sampling, and although this corresponds to the joint null hypothesis of independence and uniformity in the density forecast context, little is known about the properties of the test in respect of departures from independence. Hence to obtain direct information about possible directions of departure from the joint null hypothesis, separate tests have been employed, as noted above. However standard tests for autocorrelation face difficulties when the variable is bounded, and a further transformation has been proposed to overcome these.

3.5 *The inverse normal transformation*

Given probability integral transforms u_t, we consider the inverse normal transformation

$$z_t = \Phi^{-1}(u_t)$$

where $\Phi(\cdot)$ is the standard normal distribution function. Then if u_t is iid $U(0,1)$, it follows that z_t is iid $N(0,1)$. The advantages of this second transformation are that there are more tests available for normality, it is easier to test autocorrelation under normality than uniformity, and the normal likelihood can be used to construct likelihood ratio tests.

We note that in cases where the density forecast is explicitly based on the normal distribution, centred on a point forecast $\hat{y}_t$ with standard deviation σ_t, as in some examples discussed above, then the double transformation returns the standardised value of the outcome $\left(y_t - \hat{y}_t\right)/\sigma_t$, which could be calculated directly.

Berkowitz (2001) proposes likelihood ratio tests for testing hypotheses about the transformed series z_t. In the AR(1) model

$$z_t - \mu = \phi\left(z_{t-1} - \mu\right) + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

the hypotheses of interest are $\mu = 0,\ \sigma_\varepsilon^2 = 1$ and $\phi = 0$. The exact likelihood function of the normal AR(1) model is well known; denote it $L(\mu, \sigma_\varepsilon^2, \phi)$. Then a test of independence can be based on the statistic

$$\mathrm{LR}_{\mathrm{ind}} = -2\left(\log L(\hat{\mu}, \hat{\sigma}_\varepsilon^2, 0) - \log L(\hat{\mu}, \hat{\sigma}_\varepsilon^2, \hat{\phi})\right)$$

and a joint test of the above three hypotheses on

$$\mathrm{LR} = -2\left(\log L(0,1,0) - \log L(\hat{\mu}, \hat{\sigma}_\varepsilon^2, \hat{\phi})\right)$$

where the hats denote estimated values. However this approach does not provide tests for more general departures from iid *N*(0,1), in particular non-normality.

Moment-based tests of normality are an obvious extension, with a long history. Defining the central moments

$$\mu_j = E(z - \mu)^j$$

the conventional moment-based measures of skewness and kurtosis are

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

respectively. Sometimes $\sqrt{\beta_1}$ and $(\beta_2 - 3)$ are more convenient measures; both are equal to zero if *z* is normally distributed. Given the equivalent sample statistics,

$$\hat{\mu}_j = \frac{1}{n}\sum_{t=1}^{n}\left(z_t - \overline{z}\right)^j, \quad \sqrt{b_1} = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}}, \quad b_2 = \frac{\hat{\mu}_4}{\hat{\mu}_2^2},$$

Bowman and Shenton (1975) showed that the test statistic

$$B = n\left(\frac{\left(\sqrt{b_1}\right)^2}{6} + \frac{\left(b_2 - 3\right)^2}{24}\right)$$

is asymptotically distributed as χ_2^2 under the null hypothesis of normality. This test is often attributed to Jarque and Bera (1980) rather than Bowman and Shenton. Jarque and Bera’s contributions were to show that *B* is a score or Lagrange multiplier test statistic and hence asymptotically efficient, and to derive a correction for the case of hypotheses about regression disturbances, when the statistic is based on regression residuals. However the correction drops out if the residual sample mean is zero, as is the case in many popular regression models, such as least squares regression with a constant term.

A second possible extension, due to Bao, Lee and Saltoglu (2007), is to specify a flexible alternative distribution for ε_t that nests the normal distribution, for example a semi-nonparametric density function, and include the additional restrictions that reduce it to normality among the hypotheses under test.

Bao, Lee and Saltoglu also show that the likelihood ratio tests are equivalent to tests based on the Kullback-Leibler information criterion (KLIC) or distance measure between the forecast and “true” densities.

For a density forecast $f_1(y)$ and a "true" density $f_0(y)$ the KLIC distance is defined as

$$I\left(f_0, f_1\right) = E_0\left(\log f_0(y) - \log f_1(y)\right).$$

With E replaced by a sample average, and using transformed data z, a KLIC-based test is equivalent to a test based on

$$\log g_1(z) - \log \phi(z),$$

the likelihood ratio, where g_1 is the forecast density of z and ϕ is the standard normal density. Equivalently, the likelihood ratio statistic measures the distance of the forecast density from the "true" density. Again the transformation from $\{y\}$ to $\{z\}$ obviates the need to specify the "true" density of y, but some assumption about the density of z is still needed for this kind of test, such as their example in the previous paragraph. Berkowitz specifies g_1 as autoregressive $N\left(\mu, \sigma^2\right)$, as discussed above.

3.6 *The Bank of England's inflation forecasts*

To illustrate some of these procedures we present an evaluation of the Bank of England's density forecasts of inflation, drawn from Wallis (2004). The density forecast first published in the Bank of England's quarterly *Inflation Report* in February 1996 became the responsibility of the Monetary Policy Committee (MPC) on its establishment in 1997, when the Bank was given operational independence. Our evaluation follows the practice of the analyses of the MPC's forecasting record published in the August issue of the *Inflation Report* each year since 1999, by starting from the MPC's first inflation projection published in August 1997, and by focusing on the one-year-ahead forecasts. Strictly speaking, the forecasts are conditional projections, based on the assumption that interest rates remain at the level just agreed by the MPC. They begin with a current-quarter forecast, and extend up to eight quarters ahead. Nevertheless it is argued that the one-year-ahead projections can be evaluated as unconditional forecasts, using standard forecast evaluation procedures, since inflation does not react quickly to changes in the interest rate. On the other hand the inflation outcome two years ahead is likely to be influenced by intervening policy shifts, whose impact is difficult to estimate when comparing the outcome to a forecast with a strong judgemental component, as here. The two-year projection has played an important part in establishing policy credibility, with the central projection seldom deviating far from the inflation target.

The forecast parameters, inflation outcomes and associated *u*-values for 22 one-year-ahead forecasts are shown in Table 2. Forecasts are dated by the publication date of the *Inflation Report* in which they appear, and the inflation outcome refers to the corresponding quarter one year later. The inflation measure is the annual percentage change in the quarterly Retail Prices Index excluding mortgage interest payments (RPIX, Office for National Statistics code CHMK). Over the sample period 1997q3-2003q4 its mean is 2.40 and its standard deviation is 0.34.

With respect to the asymmetry of the forecast densities, it is seen that 13 of them exhibit positive skewness, with the mean exceeding the mode, whereas five are symmetric and four are negatively skewed. The balance of risks was thought to be on the upside of the forecast more often than not, although the average of the Bank's preferred skew measure (mean minus mode), at 0.075, is small.

Evaluations of point forecasts typically focus on the conditional expectation, the mean of the forecast density, and the *Inflation Report* forecast analyses follow suit, despite the focus on the mode, the most likely outcome, in the MPC's forecast commentary and press releases. The mean forecasts in Table 2 have an average error of zero (0.01, to be precise), thus these forecasts are unbiased. The tendency to overestimate inflation in the early part of the sample is offset by the more recent underestimation. Important contributions to this experience were the unanticipated persistence of the strength of sterling in the early years, followed more recently by surprisingly high house price inflation, which contributes to the housing depreciation component of RPIX inflation.

The standard deviation of the forecast errors is 0.42, indicating that the standard deviation of the fan chart distributions is an overestimate. A 90% confidence interval is (0.34, 0.56), and the recent entries in column (3) of Table 2 cluster around its upper limit. The dispersion of the fan charts has tended to decrease over the period, perhaps in recognition of a decline in the volatility of inflation, although the realised uncertainty is less than that assumed by the MPC at any time. This finding can be expected to dominate assessments of the goodness-of-fit of the complete distributions. A simple approach is to assess the coverage of the interquartile range, as in the SPF illustration in Section 3.2. We find that, rather than containing the nominal 50% of the outcomes, they actually contain some two-thirds of the outcomes, with 15 of the 22 *u*-values falling between 0.25 and 0.75. The forecast interquartile ranges were too wide. More generally the class frequencies in the four classes defined by the quartiles, which are equiprobable under the hypothesis of correct distributions, are 4, 6, 9, 3. The chi-squared goodness-of-fit statistic is 3.82, compared to the asymptotic critical value at the 5% level of 7.81. The data show little evidence of asymmetry, although it is only the first three outcomes in 2003 that have delivered this finding by falling in the uppermost quarter of the fan charts.

A more complete picture of the correspondence or otherwise of the fan chart forecasts to the correct distribution is given in Figure 5. This compares the sample distribution function of the observed *u*-values with the uniform distribution function, the 45° line representing the hypothesis that the densities are correct. It is again seen that there are fewer observations than there "should" be in the outer ranges of the forecasts, with the sample distribution function being correspondingly steeper than the 45° line in the central region. The fan charts fanned out too much. Whether exaggerated views of uncertainty led to undue caution in the setting of interest rates is an open research question.

3.7 *Comparing density forecasts*

A recent development in density forecasting is the comparative evaluation of forecasts, given the existence in some circumstances of competing density forecasts of the same outcome. This is a reflection, to date a small one, of the extensive literature on the comparison of point forecasts. In both cases such forecasts are sometimes genuinely competitive, having been constructed by different groups using different models or methods, and sometimes artificially competitive, a competing “benchmark” or “naïve” forecast having been constructed by forecasters wishing to undertake a comparative evaluation of their own forecasts. Either way, two activities are usually distinguished, namely hypothesis testing – is there a significant difference in forecast performance? – and model selection – which forecast is best? And in each activity, how sensitive are the results to the choice of measure of performance? We consider three groups of possible measures of performance, namely scoring rules, test statistics and distance measures, and an equivalence between them.

Scoring rules have been principally developed in probability forecasting, which has a long history in meteorology and medicine. The two leading measures are the quadratic probability or “Brier” score and the logarithmic score, which can be readily adapted to density forecasts. Given a series of density forecasts presented as *k*-bin histograms with bin probabilities p_{jt}, $j=1,...,k$, and defining an indicator variable $I_{jt}=1$ if the outcome y_t, $t=1,...,n$, falls in bin *j*, otherwise $I_{jt}=0$, the quadratic probability score is

$$QPS = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{k}\left(p_{jt}-I_{jt}\right)^2 , \quad 0 \le QPS \le 2 .$$

The logarithmic score, abbreviated to *Slog*, is

$$Slog = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{k} I_{jt}\log\left(p_{jt}\right) \quad \text{or} \quad \frac{1}{n}\sum_{t=1}^{n}\log\left(f_{t-1}\left(y_t\right)\right)$$

if $f_{t-1}(y)$ is a continuous (one-step-ahead) forecast density. It is entirely possible that different rankings of competing forecasts are given by the different scoring rules.

For two density forecasts $f_1(y)$ and $f_2(y)$, Bao, Lee and Saltoglu (2007) consider the KLIC difference

$$I\left(f_0, f_1\right) - I\left(f_0, f_2\right).$$

Again replacing *E* by a sample average, but without transforming the data, a likelihood ratio test of equal forecast performance can be based on the sample average of

$$\log f_2\left(y_t\right) - \log f_1\left(y_t\right).$$

Amisano and Giacomini (2007) develop the same test by starting from the logarithmic score as a comparative measure of forecast performance. Using outcomes $\{y\}$ rather than transformed data $\{z\}$ (or $\{u\}$) is preferred, because the need to specify and/or estimate the density of *z* (or *u*) is avoided. This is not an issue in comparing goodness-of-fit statistics of competing forecasts based on transforms, such as statistics assessing departures from uniformity of $\{u\}$. While these can be readily applied to model selection problems, few hypothesis testing procedures are as yet available.

Comparisons of the goodness-of-fit of competing forecasts may also be undertaken as a preliminary to determining the weights that might be employed in the construction of a combined forecast. Mitchell and Hall (2005) consider combinations of two competing density forecasts of UK inflation, using weights based on their relative Berkowitz LR test statistics, which they interpret as “KLIC weights”. They find that the combined forecast performs worse than the better of the two individual forecasts. That combining with an inferior forecast could improve matters seems counter intuitive, but for point forecasts this is what the original result of Bates and Granger (1969) shows, if the “optimal” weights are used. Their result is that a linear combination of two competing point forecasts using the optimal (variance minimising) weight in general has a smaller forecast mse than either of the two competing forecasts. The only case in which no improvement is possible is that in which one forecast is already the minimum mse forecast; its optimal weight is then 1, and there is no gain in combining with an inferior forecast. Bates and Granger work analytically in the widely accepted least squares framework, but there is as yet no comparable setting in which to consider density forecast combination (Wallis, 2005).

4. Conclusion

It is now widely recognised that a point forecast is seldom sufficient for well-informed decision-making in the face of an uncertain future, and that it needs to be supplemented with some indication of the degree of uncertainty. The first main section of these lecture notes surveys the different ways in which economic forecasters measure and report uncertainty, and discusses some of the technical issues that arise. It is seen that much progress has been made in recent years in measuring and reporting forecast uncertainty. However there is still reluctance in some quarters to adopt the international language of uncertainty, namely probability.

The second main section surveys recent research on statistical methods for the evaluation of interval and density forecasts. The discussion aims to convey the principles and key ideas that underlie these methods, and some technical issues of interest to specialists are left on one side. These include the distinction between in-sample and out-of-sample analysis, the effects of parameter estimation error, finite-sample issues and the use of the bootstrap to estimate exact *p*-values, and the possibility of direct estimation of the "true" $f_0(y)$. Their importance and possible resolution is often related to the size of the available sample of data, and there is a major contrast in this respect between macroeconomic forecasting, which is our main concern, and financial analysis based on high-frequency data, discussed elsewhere in this program. In both fields many outstanding research problems remain, and this is an active and fruitful area in which to work.

References

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, 25, 177-190.

Anderson, G.J. (1994). Simple tests of distributional form. *Journal of Econometrics*, 62, 265-276.

Anderson, T.W. and Goodman, L.A. (1957). Statistical inference about Markov chains. *Annals of Mathematical Statistics*, 28, 89-110.

Bao, Y., Lee, T-H. and Saltoglu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, 26, 203-255.

Bates, J.M. and 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.P. (1996). Forecast error bounds by stochastic simulation. *National Institute Economic Review*, No.156, 72-79.

Blix, M. and Sellin, P. (1998). Uncertainty bands for inflation forecasts. Working Paper No.65, Sveriges Riksbank, Stockholm.

Boero, G., Smith, J. and Wallis, K.F. (2004). Decompositions of Pearson's chi-squared test. *Journal of Econometrics*, 123, 189-193.

Bowman, K.O. and Shenton, L.R. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, 62, 243-250.

Bray, M. and Goodhart, C.A.E. (2002). "You might as well be hung for a sheep as a lamb": the loss function of an agent. Discussion Paper 418, Financial Markets Group, London School of Economics.

Britton, E., Fisher, P.G. and Whitley, J.D. (1998). The *Inflation Report* projections: understanding the fan chart. *Bank of England Quarterly Bulletin,* 38, 30-37.

Calzolari, G. and Panattoni, L. (1990). Mode predictors in nonlinear systems with identities. *International Journal of Forecasting,* 6, 317-326.

Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review,* 39, 841-862.

Clements, M.P. and Hendry, D.F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Congressional Budget Office (2003). The uncertainty of budget projections: a discussion of data and methods. Congressional Budget Office Report, US Congress, Washington DC.

Congressional Budget Office (2004). *The Budget and Economic Outlook: Fiscal Years 2005 to 2014; Appendix A: The Uncertainty of Budget Projections; Appendix B: How Changes in Economic Assumptions Can Affect Budget Projections*. Congressional Budget Office, US Congress, Washington DC.

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

Diebold, F.X., Tay, A.S. and Wallis, K.F. (1999). Evaluating density forecasts of inflation: the Survey of Professional Forecasters. In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger* (R.F. Engle and H. White, eds), pp.76-90. Oxford: Oxford University Press.

Don, F.J.H. (2001). Forecasting in macroeconomics: a practitioner's view. *De Economist,* 149, 155-175.

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

Giordani, P. and Soderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, 47, 1037-1059.

Hall, S.G. (1986). The importance of non-linearities in large forecasting models with stochastic error processes. *Journal of Forecasting*, 5, 205-215.

Huizinga, F. (2001). Economic outlook 2003-2006. *CPB Report*, 2001/4, 16-22.

Jarque, C.M. and Bera, A.K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6, 255-259.

John, S. (1982). The three-parameter two-piece normal family of distributions and its fitting. *Communications in Statistics – Theory and Methods*, 11, 879-885.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, 2nd ed., vol 1. New York: Wiley.

Miller, L.H. (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association*, 51, 111-121.

Mitchell, J. and 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.

Pringle, R.M. and Rayner, A.A. (1971). *Generalized Inverse Matrices with Applications to Statistics*. London: Charles Griffin.

Rao, C.R. and Rao, M.B. (1998). *Matrix Algebra and its Applications to Statistics and Econometrics*. Singapore: World Scientific Publishing Co.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Stock, J.H. and Watson, M.W. (2003). *Introduction to Econometrics*. Boston, MA: Pearson Education.

Stuart, A., Ord, J.K. and Arnold, S. (1999). *Kendall's Advanced Theory of Statistics*, 6th ed., vol. 2A. London: Edward Arnold.

Tay, A.S. and Wallis, K.F. (2000). Density forecasting: a survey. *Journal of Forecasting*, 19, 235-254. Reprinted in *A Companion to Economic Forecasting* (M.P. Clements and D.F. Hendry, eds), pp.45-68. Oxford: Blackwell, 2002.

Thompson, P.A. and Miller, R.B. (1986). Sampling the future: a Bayesian approach to forecasting from univariate time series models. *Journal of Business and Economic Statistics*, 4, 427-436.

Wallis, K.F. (1995). Large-scale macroeconometric modeling. In *Handbook of Applied Econometrics* (M.H. Pesaran and M.R. Wickens, eds), pp.312-355. Oxford: Blackwell.

Wallis, K.F. (1999). Asymmetric density forecasts of inflation and the Bank of England's fan chart. *National Institute Economic Review*, No.167, 106-112.

Wallis, K.F. (2003). Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting*, 19, 165-175.

Wallis, K.F. (2004). An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review*, No.189, 64-71.

Wallis, K.F. (2005). Combining density and interval forecasts: a modest proposal. *Oxford Bulletin of Economics and Statistics*, 67, 983-994.

Zarnowitz, V. (1969). The new ASA-NBER survey of forecasts by economic statisticians. *American Statistician*, 23(1), 12-16.

Zarnowitz, V. and Lambros, L.A. (1987). Consensus and uncertainty in economic prediction. *Journal of Political Economy*, 95, 591-621.

Table 1.

SPF mean probability of possible percent changes in GDP and prices, quarter 4, 2006

	2005-2006	**2006-2007**
Real GDP		
≥ 6.0	0.11	0.39
5.0 to 5.9	0.30	0.72
4.0 to 4.9	2.41	3.30
3.0 to 3.9	73.02	19.43
2.0 to 2.9	19.49	48.30
1.0 to 1.9	3.30	19.59
0.0 to 0.9	0.92	5.43
−1.0 to −0.1	0.21	1.88
−2.0 to −1.1	0.16	0.62
< −2.0	0.07	0.33
GDP price index		
≥ 8.0	0.17	0.20
7.0 to 7.9	0.28	0.28
6.0 to 6.9	0.37	0.93
5.0 to 5.9	1.17	1.59
4.0 to 4.9	5.65	5.04
3.0 to 3.9	40.48	23.96
2.0 to 2.9	48.20	49.93
1.0 to 1.9	3.13	15.63
0.0 to 0.9	0.52	2.24
< 0	0.02	0.20

Notes. Number of forecasters reporting is 46. Released 13 November 2006.

Source: http://www.phil.frb.org/files/spf/spfq406.pdf (Table 4).

Table 2.

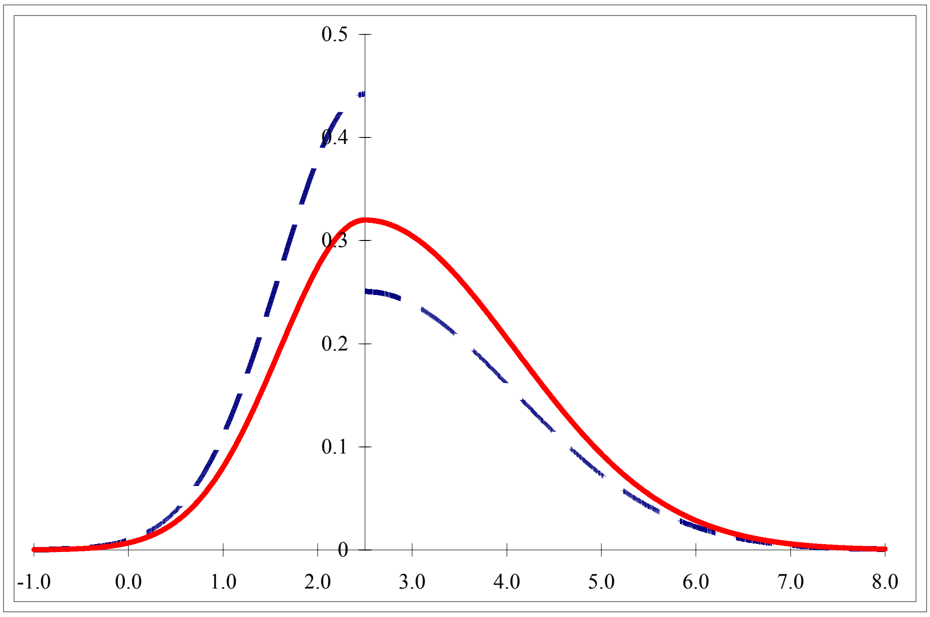
Bank of England Monetary Policy Committee inflation forecasts: one-year-ahead forecasts and outcomes (*n*=22)

Inflation	(1)	(2)	(3)	(4)	(5)
Report	Mode	Mean	Std. Dev.	Outcome	*u*
Aug 97	1.99	2.20	0.79	2.55	0.68
Nov 97	2.19	2.72	0.75	2.53	0.45
Feb 98	2.44	2.53	0.50	2.53	0.51
May 98	2.37	2.15	0.66	2.30	0.56
Aug 98	2.86	3.00	0.62	2.17	0.08
Nov 98	2.59	2.72	0.64	2.16	0.19
Feb 99	2.52	2.58	0.62	2.09	0.22
May 99	2.23	2.34	0.60	2.07	0.34
Aug 99	1.88	2.03	0.59	2.13	0.58
Nov 99	1.84	1.79	0.55	2.11	0.72
Feb 00	2.32	2.42	0.57	1.87	0.17
May 00	2.47	2.52	0.55	2.26	0.32
Aug 00	2.48	2.48	0.54	2.38	0.43
Nov 00	2.19	2.24	0.56	1.95	0.31
Feb 01	2.09	2.04	0.55	2.37	0.72
May 01	1.94	1.89	0.55	1.86	0.47
Aug 01	1.96	1.96	0.55	2.00	0.52
Nov 01	2.06	2.26	0.60	2.61	0.73
Feb 02	2.13	2.33	0.59	2.89	0.83
May 02	2.05	2.05	0.52	2.90	0.95
Aug 02	2.31	2.31	0.51	2.87	0.87
Nov 02	2.41	2.41	0.48	2.58	0.64

Notes on sources: (1),(2): Bank of England spreadsheets, see http://www.bankofengland.co.uk/inflationreport/irprobab.htm;
(3),(5): calculated using code written in the Gauss Programming Language by Michael Clements. The standard deviation is the square root of the variance given on p. 23; *u* is the probability integral transform of the inflation outcome in the forecast distribution;
(4): annual percentage growth in quarterly RPIX, ONS code CHMK.

Figure 1.

The probability density function of the two-piece normal distribution



dashed line :	two halves of normal distributions with $\mu = 2.5$, $\sigma_1 = 0.902$ (left) and $\sigma_2 = 1.592$ (right)
solid line :	the two-piece normal distribution

Figure 2.

The August 1997 *Inflation Report* fan chart

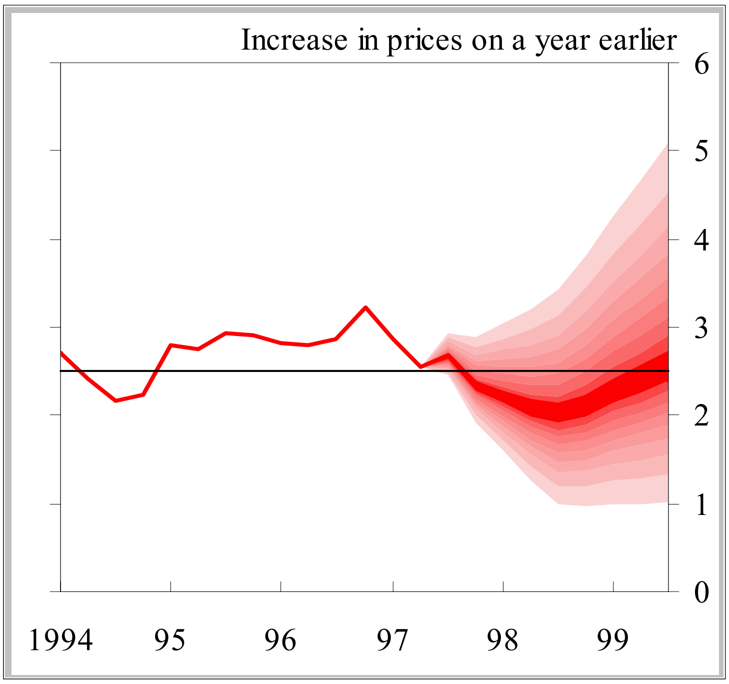


Figure 3.

Alternative fan chart based on central prediction intervals

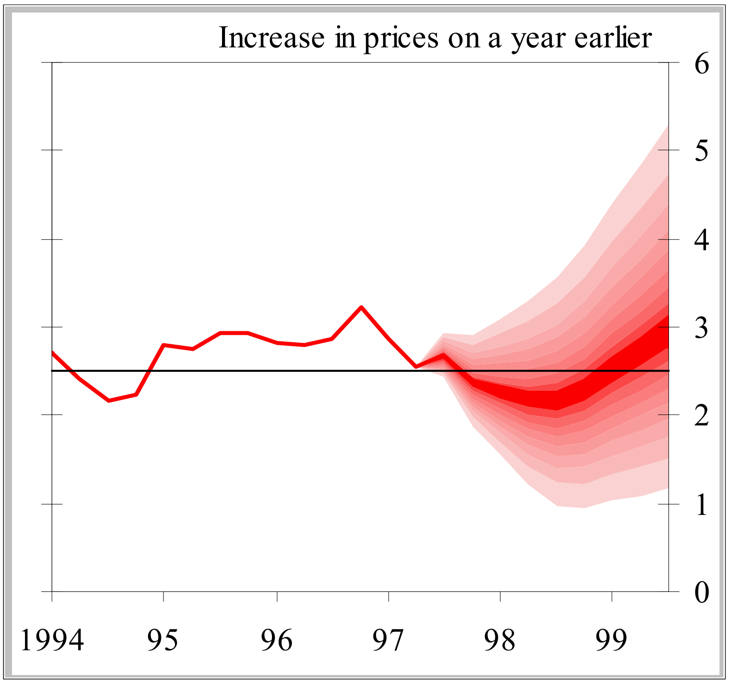
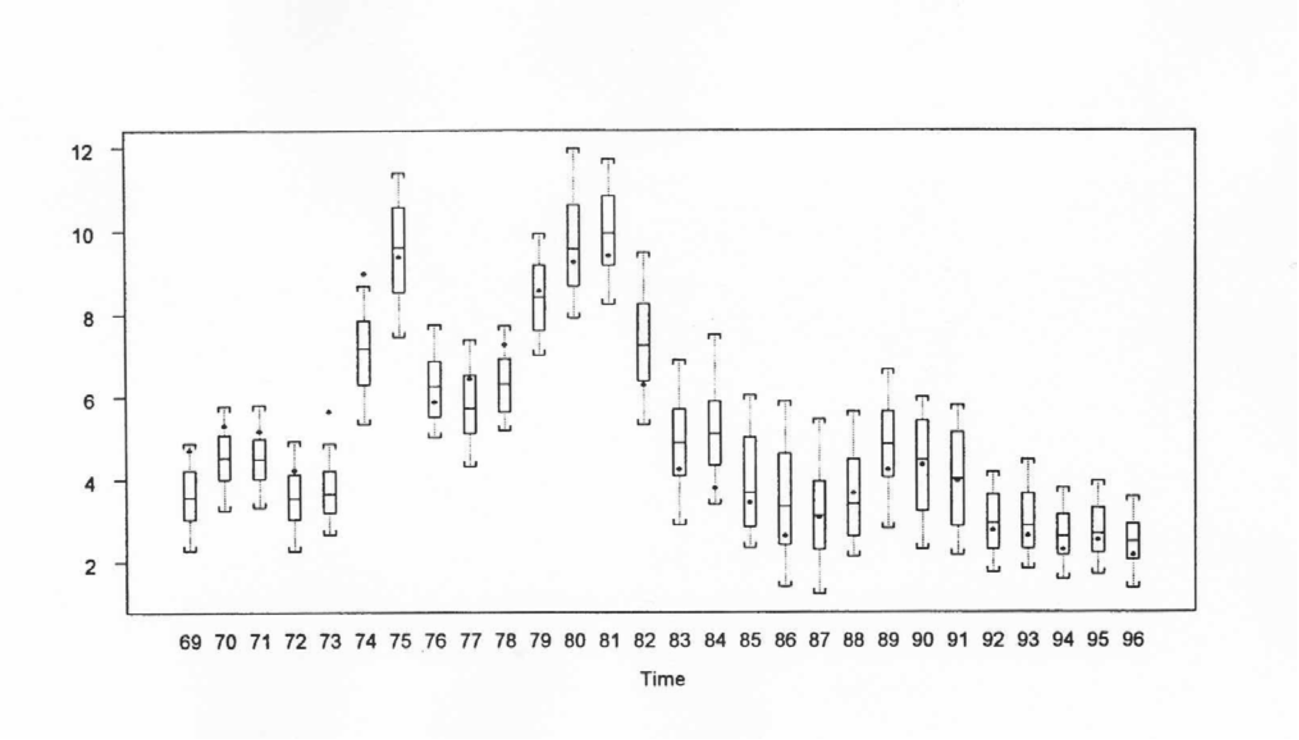


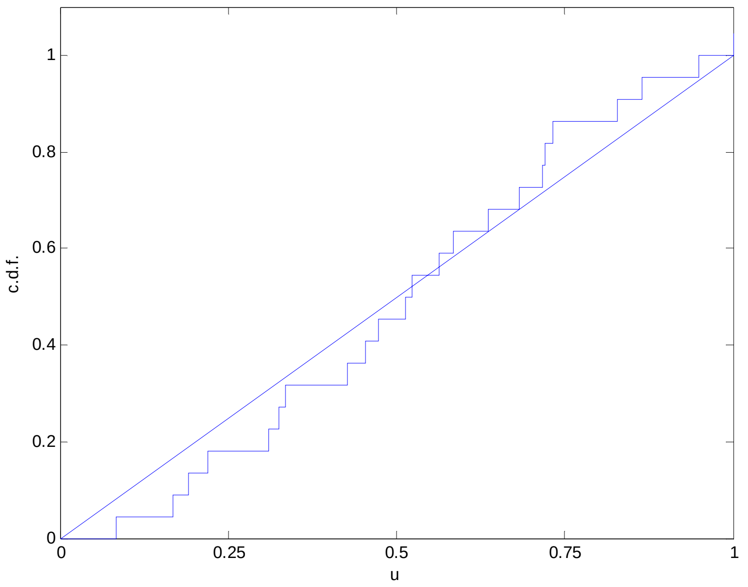
Figure 4. US inflation: SPF mean density forecasts and outcomes, 1969-1996



Note: outcomes are denoted by diamonds; forecast inter-quartile ranges by boxes. Source: Diebold, Tay and Wallis (1999).

Figure 5.

Bank of England Monetary Policy Committee inflation forecasts: cumulative distribution functions of sample *u*-values (*n*=22) and uniform distribution


1
0.8
0.6
0.4
0.2
0
c.d.f.
0
0.25
0.5
0.75
1
u