Combining and Evaluating Probabilistic Forecasts

Roopesh Ranjan

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2009

Program Authorized to Offer Degree: Statistics

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Roopesh Ranjan

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Tilmann Gneiting

Reading Committee:

Tilmann Gneiting

Donald Percival

Adrian Raftery

Date:

University of Washington

Abstract

Combining and Evaluating Probabilistic Forecasts

Roopesh Ranjan

Chair of the Supervisory Committee:
Professor Tilmann Gneiting
Statistics

Over the past one to two decades, there has been a shift of paradigms from deterministic (or point) forecasts to probabilistic (or distributional) forecasts. Probabilistic forecasts take uncertainty in the prediction into account and forecast a probability distribution function (pdf) for the unknown quantity of interest. In the case of binary events, the probabilistic forecast is the probability that the event will occur. In the case of continuous variables, the probabilistic forecast is the predictive density or distribution for the variable of interest. Calibration and sharpness are two important components of a probabilistic forecast. Calibration refers to statistical consistency between the forecasts and the realizations. Sharpness refers to the spread of the forecast pdf. The narrower the pdf, the sharper the forecast. Proper scoring rules combine calibration and sharpness together. They are a function of the probability forecast and observation that materializes. Using proper scoring rules a forecaster maximizes his expected gain by giving his true belief.

We propose a method for comparing density forecasts which is based on weighted versions of the continuous ranked probability score (CRPS). The weighting emphasizes regions of interest, such as the tails or the center of a variable's range, while encouraging the forecaster to give his true belief (propriety), as opposed to a recently developed weighted likelihood ratio test which encourages forecasters to deviate from their true beliefs (hedging). Threshold and quantile based decompositions of the CRPS can be illustrated graphically and prompt insights into the strengths and deficiencies of a forecasting method. We illustrate the use

of the weighted CRPS and graphical tools in case studies on the Bank of England's density forecasts of quarterly inflation rates in the United Kingdom, and probabilistic predictions of wind resources in the Pacific Northwest.

We also consider the problem of combining probabilistic forecasts. Linear pooling is by far the most popular method for combining probabilistic forecasts. However, any nontrivial weighted average of two or more distinct calibrated probability or density forecasts is necessarily uncalibrated and lacks sharpness. In view of this, linear pooling requires recalibration, even in the ideal case in which the individual forecasts are calibrated. Toward this end, we propose a beta transformed linear opinion pool (BLP) for the aggregation of probability forecasts or densities from distinct, calibrated or uncalibrated sources. The BLP method fits an optimal nonlinearly recalibrated forecast combination, by compositing a beta transform and the traditional linear opinion pool. The technique is illustrated in simulation examples and case studies on probability of precipitation forecasts in the Pacific Northwest and density forecasts of temperature at the Sea-Tac Airport.

TABLE OF CONTENTS

Page

List of Figures . . . iii

List of Tables . . . vi

Chapter 1: Introduction . . . 1
- 1.1 Evaluation of Probabilistic Forecasting . . . 2
- 1.2 Weighted Proper Scoring Rules . . . 6
- 1.3 Combination of Probabilistic Forecasts . . . 7

Chapter 2: Comparing Density Forecasts Using Threshold and Quantile Weighted Scoring Rules . . . 11
- 2.1 Introduction . . . 11
- 2.2 Hedging Strategies For the Weighted Likelihood Ratio Test . . . 17
- 2.3 Weighting and Testing With the Continuous Ranked Probability Score . . . 19
- 2.4 Case Studies . . . 25
- 2.5 Discussion . . . 33

Chapter 3: Combining Probability Forecasts . . . 36
- 3.1 Introduction . . . 36
- 3.2 Some Shortcomings of Linearly Combined Probability Forecasts . . . 38
- 3.3 Recalibration . . . 43
- 3.4 Case Study: Probability of Precipitation Forecasts . . . 53
- 3.5 Discussion . . . 60

Chapter 4: Density Forecast Combination, Calibration, and Recalibration . . . 66
- 4.1 Introduction . . . 66
- 4.2 Theory and methods . . . 67
- 4.3 Density forecasts for daily maximum temperature at Seattle-Tacoma Airport 81
- 4.4 Density forecasts for S&P 500 returns . . . 86
- 4.5 Discussion . . . 88

Chapter 5: Concluding Remarks 92

LIST OF FIGURES

Figure Number Page

1.1 Reliability Diagram of a calibrated forecast, left, an under-confident forecast, middle, and an over-confident forecast, right. 4

1.2 Sharpness Diagram of a more sharp forecast left, and a less sharp forecast, right. 4

1.3 PIT histograms of calibrated (left), under-dispersed (middle) and over-dispersed (right) forecasts. 10

2.1 Threshold and quantile decomposition of the mean continuous ranked probability score for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (2.10). 26

2.2 Threshold and quantile decomposition of the mean continuous ranked probability score for Bank of England (BoE) and autoregressive (AR) density forecasts of inflation rates, at a prediction horizon of one quarter. 28

2.3 Threshold and quantile decomposition of the mean continuous ranked probability score for Bank of England (BoE) and autoregressive (AR) density forecasts of inflation rates, at a prediction horizon of seven quarters. 29

2.4 Bank of England (BoE) and autoregressive (AR) forecasts of inflation rates, at a prediction horizon of seven quarters ahead, for the third quarter of 1994 through the third quarter of 2005. The plot shows the 50th and 90th percentiles of the density forecasts for the two methods along with the observed rates. 30

2.5 Threshold and quantile decomposition of the mean continuous ranked probability score for regime-switching space-time (RST) and autoregressive (AR) probabilistic forecasts of hourly average wind speed at the Stateline wind energy center, at a prediction horizon of two hours. 31

3.1 Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the individual and linearly combined forecasts in the simulation example of Section 3.2.2. The histograms show the empirical distribution of the forecast values over the unit interval. 44

3.2 Calibration curve and 95% bootstrap intervals under the null hypothesis of calibration for the BLP forecast in the simulation example of Sections 3.2.2 and 3.3.2, for the same independent test sample as that used before. The histogram shows the empirical distribution of the forecast values over the unit interval. 49

3.3 Calibration curve and 95% bootstrap intervals under the null hypothesis of calibration for the uncalibrated source p_2^* in the simulation example of Section 3.3.3. The histogram shows the empirical distribution of the forecast values over the unit interval. 52

3.4 Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the BLP forecasts in the simulation example of Section 3.3.3. The histograms show the empirical distribution of the forecast values over the unit interval. 54

3.5 Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the four individual probability of precipitation forecasts in the test period. The histograms show the empirical distribution of the forecast values over the unit interval. 55

3.6 Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the OLP and BLP probability of precipitation forecasts in the test period, using the statistical forecasts only. The histograms show the empirical distribution of the forecast values over the unit interval. 57

3.7 Same as Figure 3.6 but now using all four individual forecasts, including the NWS forecast. 59

4.1 The variance (4.2) of the PIT, $Z_\sigma = f_\sigma(Y)$, for the density forecast f_σ in Example 4.2.3 as a function of the predictive standard deviation, σ. The horizontal line is at $\frac{1}{12}$ and indicates a neutrally dispersed forecast. 69

4.2 PIT histograms for the density forecast f_σ in Example 4.2.3 where $\sigma = \frac{3}{4}$ (underdispersed), $\sigma = 1$ (neutrally dispersed and calibrated) and $\sigma = \frac{5}{4}$ (overdispersed). Details are given in the text. 69

4.3 PIT histograms for the individual and combined density forecasts in the simulation example of Section 4.2.4, for the test set. 78

4.4 PIT histograms for the individual and combined density forecasts in the simulation example of Section 4.2.5, for the test set. 79

4.5 Two-day ahead density forecasts for the maximum temperature at Seattle-Tacoma Airport on June 28, 2008. The vertical line is at the verifying maximum, at 32.8 degrees Celsius or 91 degrees Fahrenheit. 82

4.6 PIT histograms for the individual and combined density forecasts in the temperature example, for the test period. 84

4.7 PIT histograms for the combined density forecasts in the S&P 500 example, for the test period. 88

LIST OF TABLES

Table Number Page

1.1 Strictly proper scoring rules for binary forecasts. 5

1.2 Strictly proper scoring rules for density forecasts, where f is the density and F is the corresponding cumulative distribution function. 6

2.1 Weighted likelihood ratio tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1} = \mathcal{N}(0, \frac{1}{2}\hat{\sigma}^2_{t+1})$ uses a deliberately misspecified predictive variance. The width of the sliding training window is $m = 100$, and we consider $n = 900$ one-step-ahead density forecasts. Counterintuitive test statistic is shown in bold. See text for details. 13

2.2 Weighted CRPS tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1} = \mathcal{N}(0, \frac{1}{2}\hat{\sigma}^2_{t+1})$ uses a deliberately misspecified predictive variance. The width of the sliding training window is $m = 100$, and we consider $n = 900$ one-step-ahead density forecasts. In contrast to the weighted likelihood ratio test, all tests prefer f over g. 16

2.3 Weighted likelihood ratio tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (2.10). Counterintuitive test statistic is shown in bold. See text for details. 18

2.4 Proposed weight functions for threshold and quantile weighted versions of the continuous ranked probability score. The threshold weight functions are specified in terms of the probability density function $\phi_{a,b}$ and the cumulative distribution function $\Phi_{a,b}$ of the normal distribution with mean a and standard deviation b. 20

2.5 Threshold and quantile weighted CRPS tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (2.10). 26

2.6 Threshold and quantile weighted CRPS tests for density forecasts of inflation rates, at a prediction horizon of one quarter, in percent. The Bank of England forecast takes the role of f and the autoregressive benchmark the role of g. . 28

2.7 Threshold and quantile weighted CRPS tests for density forecasts of inflation rates, at a prediction horizon of seven quarters. The Bank of England forecast takes the role of f and the autoregressive benchmark the role of g. 29

2.8 Threshold and quantile weighted CRPS tests in the wind example. The regime-switching space-time (RST) forecast takes the role of f and the autoregressive benchmark the role of g. 31

3.1 Maximum likelihood estimates of OLP and BLP parameters in the simulation example of Sections 3.2.2 and 3.3.2, with standard errors in brackets. 48

3.2 Out-of-sample mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for the probability forecasts in the simulation example of Sections 3.2.2 and 3.3.2. 51

3.3 Maximum likelihood estimates of OLP and BLP parameters in the simulation example of Section 3.3.3, with standard errors in brackets. 53

3.4 Out-of-sample mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for the probability forecasts in the simulation example of Section 3.3.3. 53

3.5 Combined probability forecasts in the precipitation example, using the statistical forecasts only. Maximum likelihood estimates for the OLP and BLP parameters from the training period with standard errors in brackets. 58

3.6 Mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for individual and combined probability of precipitation forecasts in the test period, using the statistical forecasts only. 59

3.7 Same as Table 3.5 but now using all four individual forecasts, including the NWS forecast. 60

3.8 Same as Table 3.6 but now using all four individual forecasts, including the NWS forecast. 65

4.1 Maximum likelihood estimates with approximate standard errors (in brackets) for the parameters of the combined density forecasts in the simulation example of Section 4.2.4. 77

4.2 Mean logarithmic score for the individual and combined density forecasts in the simulation example of Section 4.2.4, for the training set and an independent test set. 77

4.3 Maximum likelihood estimates with approximate standard errors (in brackets) for the parameters of the combined density forecasts in the simulation example of Section 4.2.5. 77

4.4 Mean logarithmic score for the individual and combined density forecasts in the simulation example of Section 4.2.5, for the training set and an independent test set. 77

4.5 Composition of the eight-member University of Washington Mesoscale Ensemble (UWME; Eckel and Mass 2005), with member acronyms and organizational sources for initial and lateral boundary conditions. The United States National Centers for Environmental Prediction supply two distinct sets of initial and lateral boundary conditions, namely, from its Global Forecast System (GFS) and Limited-Area Mesoscale Model (ETA). 83

4.6 Maximum likelihood estimates for the predictive standard deviation, σ_i, for the individual, member specific density forecasts in the temperature example. 83

4.7 Maximum likelihood estimates for the parameters of the combined density forecasts in the temperature example. 83

4.8 Mean logarithmic score for the individual and combined density forecasts in the temperature example, for the training period and the test period. 83

4.9 Maximum likelihood estimates of the parameters for the combined density forecasts in the S&P 500 example. 87

4.10 Mean logarithmic score for the individual and combined density forecasts in the S&P 500 example, for the training period and the test period. 87

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation of the Students, Faculty and Staff of the University of Washington, Department of Statistics.

Chapter 1

INTRODUCTION

Probabilistic forecasts take forecast uncertainty into account by giving forecast distribution or predictive probability density function (pdf) of the future quantity of interest. The simplest case is that of a future binary event, such as a recession versus no recession, or rain versus no rain. In the binary case, a predictive pdf is simply the probability for the event to occur. While the roots of probability forecasting can be traced back to 18th century, the transition to probability of precipitation forecasts by the U.S. National Weather Service in 1965 was perhaps the most influential and important event in its development (Murphy 1998; Winkler and Jose 2008). In the continuous case, the Bank of England's Monetary Policy Committee (MPC) has issued probabilistic forecasts of inflation rates every quarter since February 1996 using fan charts to visualize the deciles of the predictive distributions (Wallis 2003, 2004; Clements 2004; Mitchell and Hall 2005). The archived inflation forecasts can be downloaded at `http://www.bankofengland.co.uk/publications/inflationreport/irprobab.htm`. As another example, the University of Washington Mesoscale Ensemble system routinely produces probabilistic forecasts of temperature and precipitation which are postprocessed using Bayesian model averaging (Raftery et al. 2005; Sloughter et al. 2005). This information is communicated to the user via the website `http://www.probcast.com/`. It is widely acknowledged that these distributional forecasts are more useful than merely giving point forecasts, which do not take into account the uncertainty in forecasting. Of course, there are many other important applications of probabilistic forecasts including medical diagnosis (Pepe 2005), educational testing, and political and socio-economic foresight (Tetlock 2005). One can discern a transition from point forecasts to distributional forecasts in a multidisciplinary strand of literature (Gneiting 2008).

1.1 Evaluation of Probabilistic Forecasting

With the continued use of probabilistic forecasts evaluation methods have been developed. These methods differ from the classical evaluation techniques like mean square error or mean absolute error used for point forecasts. We present below some techniques used for evaluating probability forecasts of binary events and density forecasts of continuous quantities.

1.1.1 Evaluation of binary forecasts

A special set of tools have been developed for use in the case of binary forecasts: Calibration diagram, sharpness diagram and proper scoring rules.

Calibration or Reliability diagram

Suppose a forecaster gives the probability forecast of rain on the coming day, a day in advance. Assume that we have collected information on past forecasts and realized observation (rain or no rain) for the last couple of years. We can now look at all those days when the probability forecast is close to 80%. For a good forecaster we would expect that the proportion of rainy days in those days when the forecast was 80% is close to 0.8. If this happens, we say that the forecaster is *calibrated (reliable) at 0.8.* If a forecaster is calibrated at all probabilities which she forecasts, then we say that she is *calibrated or reliable.* The drawing which plots against each probability the empirical frequency of the event is called the *reliability diagram or calibration plot.* For a well calibrated forecaster his calibration diagram is close to a diagonal line. Similarly, a forecaster is said to be under (or over) confident if his forecasts are less (or more) extreme than empirical frequencies. For an under-confident forecaster the reliability diagram is S-shaped. Figure 1.1 gives an example of a well calibrated, an over-confident and an under-confident forecast.

The above definition of calibration is an empirical one and will be made mathematically rigorous in Chapter 3.

Sharpness

From the preceding paragraph we see that reliability is an important criteria for evaluating forecast probabilities. However, we shall quickly see that it is not enough to have a reliable forecast. Let's assume a forecaster who forecasts every day the chance of rain to be the empirical frequency of rain in the last 5 years (say, 30%). Now assuming that precipitation pattern is relatively stationary over the last six years in the region of interest, he is going to be calibrated. But, this forecaster is not very useful in making decisions. So, we need more than calibration to evaluate a probabilistic forecast. We need the forecasts to be as extreme as possible. A forecast which is close to 0 or 1 is more informative than a forecast which falls in the middle of the interval. So, a sharp forecast is one whose forecast probabilities are close to 0 or 1. Therefore, we can say that the goal of probabilistic forecasting is to maximize sharpness subject to calibration (Murphy and Winkler 1987; Gneiting, Balabdaoui and Raftery 2007; Pal 2009). To assess sharpness we plot the histogram of forecast probabilities. For a sharp forecast the histogram is U-shaped. Figure 1.2 gives examples of forecasts with high and low sharpness.

Proper scoring rules for binary forecasts

Scoring rules are a way to measure the quality of a probabilistic forecast. They are a function of the forecast probability p and the binary outcome Y and are interpreted as the reward obtained by the forecaster (Jolliffe and Stephenson 2003; Gneiting and Raftery 2007). Let $S(p, Y)$ denote the score obtained when the forecast is p and the observation realized is Y. We can define its expected score when the true probability is q by,

$$S(p, q) = E_q s(p, Y) = qS(p, 1) + (1 - q)S(p, 0).$$

A scoring rule is called proper, if it is maximized by forecasting the true probability i.e.

$$S(p, q) \leq S(q, q) \; \forall \; p, q.$$

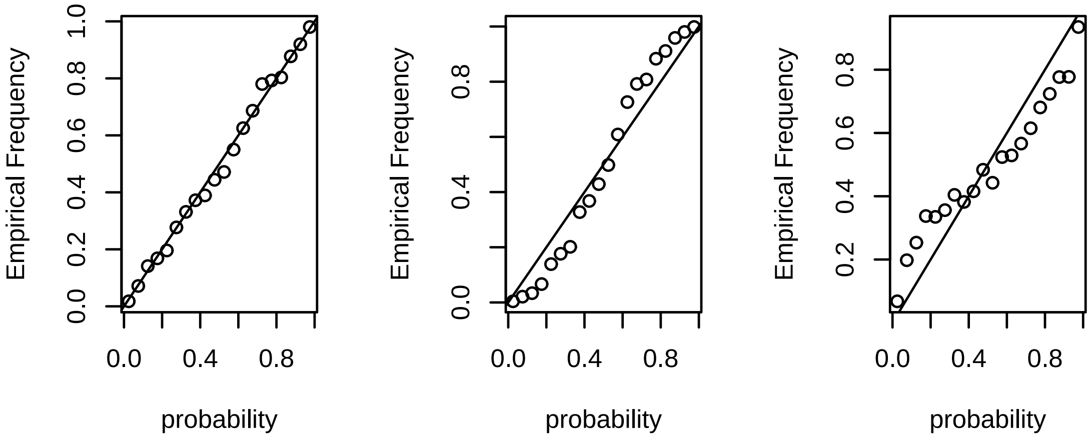

Figure 1.1: Reliability Diagram of a calibrated forecast, left, an under-confident forecast, middle, and an over-confident forecast, right.

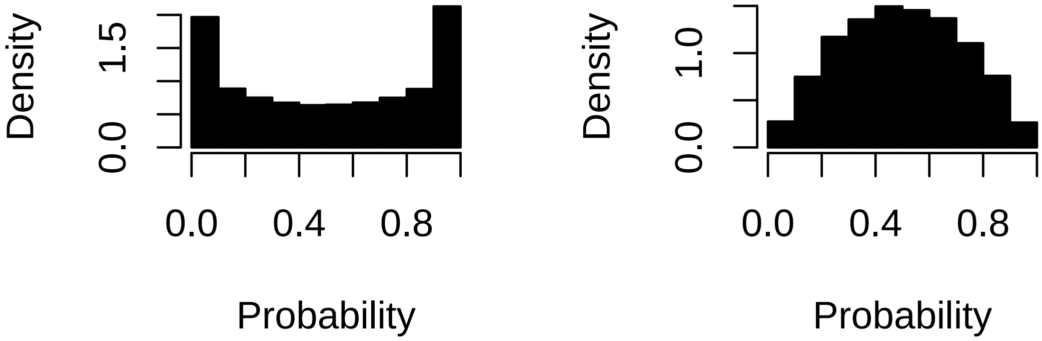

Figure 1.2: Sharpness Diagram of a more sharp forecast left, and a less sharp forecast, right.

Table 1.1: Strictly proper scoring rules for binary forecasts.

Name	S(p,0)	S(p,1)
Brier score (Brier 1950)	$-p^2$	$-(1-p)^2$
Log score (Good 1952)	$\log(1-p)$	$\log(p)$
Spherical score	$\frac{1-p}{\sqrt{p^2+(1-p)^2}}$	$\frac{p}{\sqrt{p^2+(1-p)^2}}$

It is called strictly proper when strict inequality holds for all $p \neq q$. Strictly proper scoring rules reward honesty and truthfulness in reporting in that, a forecaster maximizes his expected score by reporting his true belief (Winkler and Murphy 1968). Proper scoring rules combine calibration and sharpness together (Gneiting et al. 2008). Table 1.1 gives examples of strictly proper scoring rules for binary forecasts.

1.1.2 Evaluation of density forecasts

Probability Integral Transform (PIT) histogram

Let $F(\cdot)$ be a continuous predictive distribution and Y the realized value. Then, we define the *probability integral transform (PIT)* of Y by,

$$Z = F(Y).$$

If Y has the distribution F, then $Z \sim U(0,1)$. Now assume that we have a sequence of observations in time $Y_1, Y_2, \ldots, Y_T$. Under the assumption that F_t is the true conditional distribution of Y_t given the past $Y_1, Y_2, \ldots, Y_{t-1}$, Diebold et al. (1998) show that $Z_t,\ t = 1, 2, \ldots, T$ are independently distributed as U(0,1). Thus, checking for calibration requires checking for the uniformity of the PIT histogram of $Z_t,\ t = 1, 2, \ldots, T$. Although Diebold et al. (1998) derived their results in the context of time series, operationally the use of PIT doesn't require that the probabilistic forecasts be evaluated in a time series framework, and is quite general. For a calibrated sequence of density forecasts the PIT histogram resembles uniform histogram. If the density forecasts are under-dispersed or over-dispersed

Table 1.2: Strictly proper scoring rules for density forecasts, where f is the density and F is the corresponding cumulative distribution function.

Name	$S(f, Y)$
Log score (Good 1952)	$\log f(Y)$
CRPS score (Matheson and Winkler 1976)	$-\int_{-\infty}^{\infty}(F(z) - \mathbb{I}\{Y \leq z\})^2 \, \mathrm{d}z$
Quadratic score	$2f(Y) - E_f f(Y)$
Spherical score	$\frac{f(Y)}{\\|f\\|}, \\|f\\| = (\int f(z)^2 \, \mathrm{d}z)^{\frac{1}{2}}$

the PIT histogram will be U-shaped and hump shaped respectively. Figure 1.3 gives an example of each type of forecast.

Scoring rules for density forecasts

Scoring rule for a density forecast is the reward $S(f, Y)$ the forecaster obtains when he forecasts a density f and the observation Y is realized. If the true density of Y is g then the expected score of the forecaster who forecasts f is given by,

$$S(f, g) = E_g S(f, Y).$$

A scoring rule is called proper if

$$s(f, g) \leq S(g, g) \; \forall \; f, g.$$

It is called strictly proper if the inequality is strict for all $f \neq g$. Strictly proper scoring rules encourage honesty in reporting the density forecast, in that, the expected score of the forecaster is maximized for his true belief g (Winkler and Murphy 1968). Table 1.2 gives examples of strictly proper scoring rules for density forecasts.

1.2 Weighted Proper Scoring Rules

As discussed above scoring rules are used to evaluate density forecasts. However, in certain situations different regions of the density are of higher importance. In such situations, it is necessary to use scoring rules which emphasize different regions of the density like tails, center, right-tail and left-tail while retaining propriety. Recently, Amisano and Giacomini (2007) have proposed the use of weighted version of log score for comparing the performance of competing density forecasts. In Chapter 2, we will show that the score proposed by them is not proper and can be hedged. Hence, we suggest the use of weighted versions of continuous ranked probability score (CRPS), which are proper. Threshold and quantile based decompositions of the CRPS can be illustrated graphically and give insights into the strengths and deficiencies of a forecasting method in different regions of interest. We illustrate the use of weighted scoring rules and graphical tools in case studies on the Bank of England's density forecasts of quarterly inflation rates in the United Kingdom, and probabilistic predictions of wind resources in the Pacific Northwest.

1.3 Combination of Probabilistic Forecasts

In many situations, we can have multiple models or experts generating probabilistic forecasts. In such situations, it would be important to combine various forecasts for the same quantity to generate a single probabilistic forecast. In the density case, Mitchell & Hall (2005) and Hora (2004) have among others considered the problem of combining. In the case of binary forecast, various ways of combining probability forecasts into a single aggregated forecast have been proposed. For comprehensive reviews we refer the reader to Genest and Zidek (1986), Wallsten et al. (1997), Clemen and Winkler (2007) and Primo et al. (2009).

1.3.1 Combination of binary forecasts

In practice, the most commonly used approach of aggregating forecast probabilities is to take a simple average or a weighted average of the individual probability forecasts, which is often referred to as the linear opinion pool (Stone 1961; Genest and McConway 1990; DeGroot and Mortera 1991). Empirical evidence shows the benefits of linear opinion pools over individual forecasts, with successful applications in meteorology (Sanders 1963; Vislocky and Fritsch 1995; Baars and Mass 2005), economics (Graham 1996), psychology (Ariely et al. 2000), and medical diagnosis (Winkler and Poses 1993), among other fields.

Gneiting, Balabdaoui and Raftery (2007) contend that the goal in probabilistic forecasting is to maximize sharpness of the forecast subject to calibration. In view of this it is important to aggregate probability forecasts in a way such that the combined forecast is both calibrated and sharp. In Chapter 3, we will point out some of the major deficiencies of the linear opinion pool. In particular, we will show that the linear opinion pool lacks calibration even when the individual forecasts are calibrated. We will also show that linear opinion pool lacks sharpness in the sense it tends to move away from the extreme probabilities 0 and 1. Hence, linear pooling requires recalibration, even in the ideal case in which the individual forecasts are calibrated. To accomplish this we need a non-linear generalization of the linear opinion pool to bring the linear average towards 0 or 1. We propose a beta transformed linear opinion pool (BLP) for the combination of probability forecasts from distinct, calibrated or uncalibrated sources. The BLP method applies a non-linear beta cumulative distribution function (cdf) transform to the linear average to aggregate probabilities. The weights and the parameters of the beta cdf transform are obtained simultaneously by maximizing the log-score of the combined forecast.

The method is illustrated in a simulation example and in a case study on statistical and National Weather Service probability of precipitation forecasts at 29 major cities in the continental US. In practice, we fit the combination parameters in a training set and use it to generate combined forecasts on the test set. The BLP combined forecast is both calibrated and sharp and outperforms the individual forecasts and linear opinion pools.

1.3.2 Combination of density forecasts

Here again, the most commonly used method appears to be a weighted linear combination of the individual density forecasts. The weights are determined using some optimality criteria. However, Hora (2004) proves a very interesting result. He shows that for combining two experts the linear combination of two distinct calibrated density forecasts is necessarily uncalibrated.

In Chapter 4 we will generalize Hora's result to the case of more than two experts and provide a more general proof. Unlike Hora's proof which doesn't show the direction of departure from calibration, our proof also indicates that the deviation from calibration is towards overdispersion. So, the linearly combined forecasts are overdispersed and give prediction intervals that are too wide on average. This result undermines the use of linear combination for combining density forecasts. The overdispersion of the linear opinion pool can be addressed empirically by spread adjustments to the density components, as implemented in the deflated linear pool (DLP), or via nonlinear recalibration transforms, such as the beta transformed linear pool (BLP). In the DLP method, we require the component densities to be parameterized by a scale parameter. The scale parameter is adjusted for spread by deflating it by a constant and then the spread adjusted densities are combined linearly. The adjustment parameter and the weights are estimated simultaneously by maximizing the log score. In the BLP method, a beta cdf is applied to the weighted average of the forecast distributions to get the combined cumulative distribution function. Again the weights and the recalibration parameters are estimated simultaneously.

Both methods can be used effectively to combine calibrated as well as uncalibrated sources. The effects and techniques are demonstrated theoretically, in simulation examples and in case studies on density forecasts for S&P 500 returns and daily maximum temperature at Seattle-Tacoma Airport. They also have relevance in the fusion of expert opinions that are expressed in terms of probability densities.

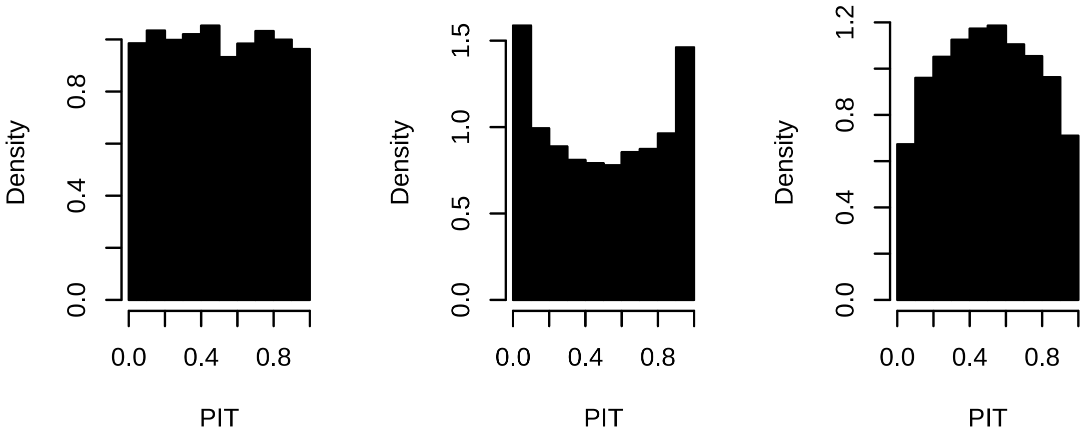

Figure 1.3: PIT histograms of calibrated (left), under-dispersed (middle) and over-dispersed (right) forecasts.

Chapter 2

COMPARING DENSITY FORECASTS USING THRESHOLD AND QUANTILE WEIGHTED SCORING RULES

2.1 Introduction

One of the major tasks of statistical analysis is to make forecasts for the future. To realize their full potential, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities or events (Dawid 1984). Here we are concerned with density forecasts of a continuous variable, such as inflation rate, gross domestic product, temperature or wind speed, to name but a few examples. With the continued proliferation of probabilistic forecasts in economic, environmental and meteorological applications, among others, there is a critical need for principled techniques for the comparison and ranking of density forecasts (Timmermann 2000; Elliott and Timmermann 2008; Gneiting 2008a).

Following Amisano and Giacomini (2007), we consider density forecasts in a time series context, in which a rolling window consisting of the past m observations is used to fit a density forecast for a future observation that lies k time steps ahead. The reason for using a rolling window of size m instead of using all past observations is that the rolling window approach will be able to exploit the non-stationary in data more effectively. Specifically, suppose that $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ is a stochastic process which can be partitioned as $\boldsymbol{Z}_t = (Y_t, \boldsymbol{X}_t)$ where Y_t is the variable of interest and $\boldsymbol{X}_t$ is a vector of predictors. Suppose that $T = m+n$. At times $t = m, \ldots, m+n-k$, density forecasts $\hat{f}_{t+k}$ and $\hat{g}_{t+k}$ for Y_{t+k} are generated, each of which depends only on $\boldsymbol{Z}_{t-m+1}, \ldots, \boldsymbol{Z}_t$. In this framework, the only requirement imposed on how the forecasts are produced is that they are measurable functions of the data in the rolling estimation window. We are interested in comparing and ranking the competing density forecasting methods.

The comparison typically uses a proper scoring rule. A scoring rule is a loss function $\mathrm{S}(f, y)$ whose arguments are the density forecast, f, and the realization, y, of the future observation, Y. The density forecast is ideal if the conditional sampling density of Y is indeed f. Diebold et al. (1998) and Granger and Pesaran (2000) argue powerfully that the ideal forecast is preferred by any rational user, irrespectively of the cost-loss structure at hand. Hence, it is critically important that a scoring rule be proper, in the sense that

$$\begin{aligned} \mathbb{E}_f \, \mathrm{S}(f, Y) &= \int f(y) \, \mathrm{S}(f, y) \, \mathrm{d}y \\ &\leq \int f(y) \, \mathrm{S}(g, y) \, \mathrm{d}y = \mathbb{E}_f \, \mathrm{S}(g, Y) \end{aligned} \tag{2.1}$$

for all density functions f and g. A scoring rule is strictly proper if (2.1) holds, with equality if and only if $f = g$ almost surely. Clearly, a strictly proper scoring rule prefers the ideal forecaster over any other. Prominent examples of strictly proper scoring rules include the logarithmic, quadratic, spherical, and continuous ranked probability scores (Matheson and Winkler 1976; Good 1952; Diebold and Rudebusch 1989; Winkler 1996; Gneiting and Raftery 2007). We take scoring rules to be negatively oriented penalties, so the lower the score, the better.

Density forecast methods are then ranked by comparing their average scores. Specifically, if

$$\overline{\mathrm{S}}_n^f = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{S}(\hat{f}_{t+k}, y_{t+k}) \quad \text{and} \quad \overline{\mathrm{S}}_n^g = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{S}(\hat{g}_{t+k}, y_{t+k}),$$

then we prefer f if $\overline{\mathrm{S}}_n^f < \overline{\mathrm{S}}_n^g$, and prefer g otherwise. Amisano and Giacomini (2007) consider tests of equal forecast performance based on the test statistic

$$t_n = \sqrt{n} \, \frac{\overline{\mathrm{S}}_n^f - \overline{\mathrm{S}}_n^g}{\hat{\sigma}_n}, \tag{2.2}$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n-k+1} \sum_{j=-(k-1)}^{k-1} \sum_{t=m}^{m+n-k-|j|} \Delta_{t,k} \Delta_{t+|j|,k} \quad \text{and} \quad \Delta_{t,k} = \mathrm{S}(\hat{f}_{t+k}, y_{t+k}) - \mathrm{S}(\hat{g}_{t+k}, y_{t+k}), \tag{2.3}$$

as proposed by Diebold and Mariano (1995). Assuming suitable regularity conditions, the

Table 2.1: Weighted likelihood ratio tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1} = \mathcal{N}(0, \frac{1}{2}\hat{\sigma}^2_{t+1})$ uses a deliberately misspecified predictive variance. The width of the sliding training window is $m = 100$, and we consider $n = 900$ one-step-ahead density forecasts. Counterintuitive test statistic is shown in bold. See text for details.

Weight Function	Emphasis	$\overline{\mathrm{S}}_n^f$	$\overline{\mathrm{S}}_n^g$	$\hat{\sigma}_n$	t_n	P
$w_0(x) = 1$	uniform	1.312	1.490	0.862	−6.20	< 0.001
$w_1(x) = \phi(x)$	center	0.294	0.267	0.100	**7.98**	< 0.001
$w_2(x) = 1 - \phi(x)/\phi(0)$	tails	0.575	0.820	0.759	−9.69	< 0.001
$w_3(x) = \Phi(x)$	right tail	0.667	0.767	0.633	−4.73	< 0.001
$w_4(x) = 1 - \Phi(x)$	left tail	0.645	0.723	0.542	−4.34	< 0.001

statistic t_n is asymptotically standard normal under the null hypothesis of vanishing expected score differentials (Amisano and Giacomini, 2007). In the case of rejection, f is chosen if t_n is negative and g is chosen if t_n is positive.[1]

What scoring rule should be used? Amisano and Giacomini (2007) consider a weighted logarithmic scoring rule,

$$\mathrm{S}(f, y) = w\left(\frac{y - \mu}{\sigma}\right) \mathrm{S}_0(f, y), \tag{2.4}$$

where w is a fixed, nonnegative weight function, μ and σ are estimates of the unconditional mean and standard deviation of the predictand, based on the past m observations, and S_0 is the logarithmic scoring rule, $\mathrm{S}_0(f, y) = -\log f(y)$. The weight function emphasizes regions of interest, such as the tails or the center of a variable's range. With ϕ and Φ denoting the standard normal probability density and cumulative distribution function, the weight functions $w_1(x) = \phi(x)$, $w_2(x) = 1 - \phi(x)/\phi(0)$, $w_3(x) = \Phi(x)$ and $w_4(x) = 1 - \Phi(x)$ emphasize the center, the tails, the right tail and the left tail, respectively. The approach of Mitchell and Hall (2005) and Bao, Lee and Saltoğlu (2007) employs the unweighted, original logarithmic score.

The weighting approach seems appealing; however, it corresponds to the use of an im-

[1] Amisano and Giacomini (2007) use the logarithmic score in positive orientation, so they choose f if t_n is positive and g if t_n is negative.

proper scoring rule and incurs misguided inferences. For instance, consider the Gaussian GARCH(1,1) process $Y_1, Y_2, \ldots$, where

$$Y_{t+1} = \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim \mathcal{N}(0, \sigma_{t+1}^2), \qquad \sigma_{t+1}^2 = \alpha\epsilon_t^2 + \beta\sigma_t^2 + \gamma. \tag{2.5}$$

Following Christoffersen and Diebold (1996), we set the GARCH parameters at $\alpha = 0.2$ and $\beta = 0.75$, which are typical of estimates reported in the literature, and we let $\gamma = 0.05$, which normalizes the unconditional process variance to 1.[2] The rolling estimation window is of size $m = 100$, and we consider $n = 900$ density forecasts at the prediction horizon $k = 1$. The density forecast $\hat{f}_{t+1}$ is Gaussian with mean zero and variance $\hat{\sigma}^2$, which is derived from a GARCH fit for (2.5). Except for uncertainty in parameter estimation, this is the ideal density forecast. In contrast, the density forecast $\hat{g}_{t+1}$ is Gaussian with mean zero and variance one half time times $\hat{\sigma}^2$, deliberately misspecifying the conditional variance. Results for the weighted likelihood ratio test are shown in Table 2.1. Using the weight functions w_0, w_2, w_3 and w_4 the test prefers f, as expected. With the weight function w_1, the test prefers the misspecified density forecast g, which is a counterintuitive result.

The goal of this chapter is to propose a test that adopts the weighting approach of Amisano and Giacomini (2007), avoids misguided inferences, and comes with associated graphical tools that can be used to diagnose strengths and weaknesses of a forecasting method. We retain the test statistic (2.2), but base our test on appropriately weighted, proper versions of the continuous ranked probability score (CRPS; Matheson and Winkler 1976; Gneiting and Raftery 2007; Laio and Tamea 2007). Any density forecast f induces a probability forecast for the binary event $\{Y \le z\}$ via the value of the corresponding cumulative distribution function (cdf) $F(z)$ at the threshold $z \in \mathbb{R}$. Similarly, it induces the quantile forecast $F^{-1}(\alpha)$ at the level $\alpha \in (0,1)$. The continuous ranked probability score is then defined as

$$\mathrm{CRPS}(f, y) = \int_{-\infty}^{\infty} \mathrm{PS}(F(z), \mathbb{I}\{y \le z\})\,\mathrm{d}z = \int_0^1 \mathrm{QS}_\alpha(F^{-1}(\alpha), y)\,\mathrm{d}\alpha, \tag{2.6}$$

[2] See Engle (1982) and Bollerslev (1986) for details on ARCH and GARCH processes. We set the initial conditional variance equal to $\sqrt{609}/7$, that is, the unconditional variance plus one standard deviation of the conditional variance, and discard the first 1,000 values.

where

$$\mathrm{PS}(F(z), \mathbb{I}\{y \le z\}) = (F(z) - \mathbb{I}\{y \le z\})^2$$

is the Brier probability score (Selten 1998; Gneiting and Raftery 2007) for the probability forecast $F(z)$ of the binary event $\{Y \le z\}$ at the threshold $z \in \mathbb{R}$, and

$$\mathrm{QS}_\alpha(F^{-1}(\alpha), y) = 2\,(\mathbb{I}\{y \le F^{-1}(\alpha)\} - \alpha)\,(F^{-1}(\alpha) - y)$$

is the quantile score (Gneiting and Raftery 2007) for the quantile forecast $F^{-1}(\alpha)$ at the level $\alpha \in (0,1)$. Here and in the following, the symbol $\mathbb{I}$ stands for an indicator function. The second equality in (2.6) is due to Laio and Tamea (2007) and will be reviewed below.

Following Matheson and Winkler (1976) and Gneiting and Raftery (2007), it is straightforward to construct weighted versions of the continuous ranked probability score (2.6) that emphasize regions of interest and retain propriety. A threshold weighted version of the continuous ranked probability score is obtained as

$$\mathrm{S}(f, y) = \int_{-\infty}^{\infty} \mathrm{PS}(F(z), \mathbb{I}\{y \le z\})\, u(z)\, \mathrm{d}z, \tag{2.7}$$

where u is a nonnegative weight function on the real line. Similarly, a quantile weighted version is obtained as

$$\mathrm{S}(f, y) = \int_0^1 \mathrm{QS}_\alpha(F^{-1}(\alpha), y)\, v(\alpha)\, \mathrm{d}\alpha, \tag{2.8}$$

where v is a nonnegative weight function on the unit interval. For a constant weight function, both (2.7) and (2.8) reduce to the unweighted score (2.6).

Table 2.2 returns to the simulation study for the GARCH model (2.5) and reports results based on the test statistic (2.2) and threshold or quantile weighted versions of the continuous ranked probability score, which are proper (Matheson and Winkler 1976), as opposed to the weighted logarithmic score. In contrast to the results for the weighted likelihood ratio test, all t_n values in Table 2.2 are negative, favoring the nearly ideal density forecast f over its deliberately misspecified competitor g.

The remainder of the chapter is organized as follows. In Section 2.2 we show that the

Table 2.2: Weighted CRPS tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1} = \mathcal{N}(0, \frac{1}{2}\hat{\sigma}^2_{t+1})$ uses a deliberately misspecified predictive variance. The width of the sliding training window is $m = 100$, and we consider $n = 900$ one-step-ahead density forecasts. In contrast to the weighted likelihood ratio test, all tests prefer f over g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^f$	$\overline{\mathrm{S}}_n^g$	$\hat{\sigma}_n$	t_n	P
$u_0(z) = 1$	uniform	0.511	0.521	0.070	-3.94	< 0.001
$u_1(z) = \phi(z)$	center	0.153	0.155	0.018	-4.24	< 0.001
$u_2(z) = 1 - \phi(z)/\phi(0)$	tails	0.129	0.132	0.030	-2.88	0.004
$u_3(z) = \Phi(z)$	right tail	0.258	0.262	0.046	-2.83	0.005
$u_4(z) = 1 - \Phi(z)$	left tail	0.254	0.259	0.046	-3.24	0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^f$	$\overline{\mathrm{S}}_n^g$	$\hat{\sigma}_n$	t_n	P
$v_0(\alpha) = 1$	uniform	0.511	0.521	0.070	-3.95	< 0.001
$v_1(\alpha) = \alpha(1 - \alpha)$	center	0.100	0.101	0.009	-2.79	0.005
$v_2(\alpha) = (2\alpha - 1)^2$	tails	0.113	0.118	0.036	-4.85	< 0.001
$v_3(\alpha) = \alpha^2$	right tail	0.157	0.161	0.041	-2.53	0.012
$v_4(\alpha) = (1 - \alpha)^2$	left tail	0.155	0.159	0.041	-3.00	0.003

weighted likelihood ratio test incurs the use of an improper scoring rule, and we explore ways in which the test can be hedged. In Section 2.3 we study threshold and quantile weighted versions of the continuous ranked probability score in further detail, and discuss conditions under which the test statistic t_n is asymptotically standard normal. We also note graphical representations of the threshold and quantile decompositions of the continuous ranked probability score, which can be used diagnostically to assess strengths and deficiencies of forecasting techniques. Section 2.4 applies these methods to compare density forecasts for quarterly inflation rates in the United Kingdom and wind resources in the North American Pacific Northwest. The chapter closes with a discussion in Section 2.5.

2.2 Hedging Strategies For the Weighted Likelihood Ratio Test

Recall that a scoring rule $\mathrm{S}(f, y)$ for a density forecast is proper if

$$\begin{aligned}\mathbb{E}_f\,\mathrm{S}(f, Y) &= \int f(y)\,\mathrm{S}(f, y)\,\mathrm{d}y \\ &\leq \int f(y)\,\mathrm{S}(g, y)\,\mathrm{d}y \;=\; \mathbb{E}_f\,\mathrm{S}(g, Y)\end{aligned}$$

for all density functions f and g. It is strictly proper if the above holds, with equality if and only if $f = g$ almost surely. Examples of proper scoring rules for density forecasts include the logarithmic score, $\mathrm{S}(f, y) = -\log f(y)$, the quadratic score, $\mathrm{S}(f, y) = -2f(y) + \|f\|^2$, and the spherical score $\mathrm{S}(f, y) = -f(y)/\|f\|$, where

$$\|f\| = \left(\int_{-\infty}^{\infty} f(y)^2\,\mathrm{d}y\right)^{1/2}.$$

The continuous ranked probability score and its weighted versions are also proper (Matheson and Winkler 1976; Gneiting and Raftery 2007).

The following result shows that if $\mathrm{S}_0(f, y)$ is a strictly proper scoring rule, then its product with a non-negative weight function $w(y)$ is improper, unless the weight function is constant.

Theorem 2.2.1. *Suppose that f is the sampling density of the random variable Y. Let S_0 be any proper scoring rule and let w be a weight function such that $0 < \int w(y) f(y)\,\mathrm{d}y < \infty$. Then the expected value of the weighted score*

$$\mathrm{S}(g, Y) = w(Y)\,\mathrm{S}_0(g, Y) \tag{2.9}$$

is minimized if we issue the density forecast

$$g(y) = \frac{w(y) f(y)}{\int w(y) f(y)\,\mathrm{d}y}.$$

Table 2.3: Weighted likelihood ratio tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (2.10). Counterintuitive test statistic is shown in bold. See text for details.

Weight Function	Emphasis	$\overline{\mathrm{S}}_n^f$	$\overline{\mathrm{S}}_n^g$	$\hat{\sigma}_n$	t_n	P
$w_0(x) = 1$	uniform	1.312	1.611	0.727	−12.31	< 0.001
$w_1(x) = \phi(x)$	center	0.294	0.436	0.215	−19.84	< 0.001
$w_2(x) = 1 - \phi(x)/\phi(0)$	tails	0.575	0.518	0.331	**5.23**	< 0.001
$w_3(x) = \Phi(x)$	right tail	0.667	0.744	0.515	−4.48	< 0.001
$w_4(x) = 1 - \Phi(x)$	left tail	0.645	0.867	0.310	−21.51	< 0.001

Proof. Let h be any density forecast. Then

$$\begin{aligned}\mathbb{E}_f \,\mathrm{S}(g, Y) &= \int w(y) f(y) \,\mathrm{S}_0(g, y) \,\mathrm{d}y = \int w(y) f(y) \,\mathrm{d}y \int g(y) \,\mathrm{S}_0(g, y) \,\mathrm{d}y \\ &\leq \int w(y) f(y) \,\mathrm{d}y \int g(y) \,\mathrm{S}_0(h, y) \,\mathrm{d}y = \int w(y) f(y) \,\mathrm{S}_0(h, y) \,\mathrm{d}y = \mathbb{E}_f \,\mathrm{S}(h, Y),\end{aligned}$$

where the inequality reflects the propriety of S_0. □

In particular, we are now in a position to explain the failure of the weighted likelihood ratio test in the simulation example in table 2.1 of the introduction. The weighted logarithmic score (2.4) is similar to the composite scoring rule (2.9) where S_0 is the logarithmic score, and the composite score is improper, unless the weight function is constant. Moreover, Theorem 2.2.1 suggests a hedging strategy if forecasters are compared by the weighted likelihood ratio test, namely to issue the density function g that is proportional to the product of the forecaster's true belief, f, and the weight function, w. For example, if both $f = \phi$ and $w = \phi$ are standard normal, the suggested hedge uses a normal density function g with mean zero and variance one half. Essentially, this is the situation in the simulation study in the introduction. The misspecified density forecast $\hat{g}_{t+1}$ halves the estimated Gaussian variance; hence, to a good degree of approximation, it is proportional to the product of the true belief, $\hat{f}_{t+1}$, and the weight function, $w_1 = \phi$. Not surprisingly, the weighted likelihood ratio test with weight function w_1 fails.

Before closing this section, we present another simulation study in which the weighted likelihood ratio test yields counterintuitive results. Once again, we study density forecasts for the conditionally heteroscedastic process (2.5) with parameter values $\alpha = 0.2$, $\beta = 0.75$ and $\gamma = 0.05$. The rolling estimation window is of size $m = 100$, and we issue $n = 900$ density forecasts at the prediction horizon $k = 1$. As previously, the density forecast $\hat{f}_{t+1}$ is Gaussian with mean zero and variance $\hat{\sigma}^2_{t+1}$, derived from a GARCH fit under the correct model specification. Except for estimation uncertainty, this is the ideal density forecast. Its competitor is the density forecast $\hat{g}_{t+1}$, which is deliberately misspecified as

$$\hat{g}_{t+1}(y) = \hat{f}_{t+1}(y)\left(\mathbb{I}\{y < -\hat{\sigma}_{t+1}\} + \frac{1}{2}\,\mathbb{I}\{|y| \leq \hat{\sigma}_{t+1}\} + \frac{1}{2(1-\Phi(1))}\,\mathbb{I}\{y > \hat{\sigma}_{t+1}\}\right). \quad (2.10)$$

Note that $\hat{g}_{t+1}$ is identical to $\hat{f}_{t+1}$ in the left tail, underspecifies the center of the distribution, and makes this up in the right tail. Table 2.3 shows results for the weighted likelihood ratio test, which are misguided and inconsistent. Specifically, the test suggests that both in the left tail and in the right tail f is preferable. Looking at both tails simultaneously, the test stipulates that g is better.

2.3 Weighting and Testing With the Continuous Ranked Probability Score

2.3.1 Threshold and quantile weighting for the continuous ranked probability score

Suppose that the density forecast is f and y realizes. Let F denote the CDF corresponding to the density f, and write $F^{-1}(\alpha)$ for the quantile at level $\alpha \in (0,1)$. The continuous ranked probability score then can be defined in three equivalent ways, as

$$\begin{aligned} \mathrm{CRPS}(f,y) &= \int_{-\infty}^{\infty} (F(z) - \mathbb{I}\{y \leq z\})^2 \,\mathrm{d}z & (2.11)\\ &= 2\int_0^1 (\mathbb{I}\{y \leq F^{-1}(\alpha)\} - \alpha)\,(F^{-1}(\alpha) - y)\,\mathrm{d}\alpha & (2.12)\\ &= \mathbb{E}_f|Y - y| - \frac{1}{2}\mathbb{E}_f|Y - Y'|, & (2.13) \end{aligned}$$

where Y and Y' are independent random variables with common sampling density f. Laio and Tamea (2007) showed the equivalence of the traditional form (2.11), to which we refer to

Table 2.4: Proposed weight functions for threshold and quantile weighted versions of the continuous ranked probability score. The threshold weight functions are specified in terms of the probability density function $\phi_{a,b}$ and the cumulative distribution function $\Phi_{a,b}$ of the normal distribution with mean a and standard deviation b.

Emphasis	Threshold Weight Function	Quantile Weight Function
center	$u_1(z) = \phi_{a,b}(z)$	$v_1(\alpha) = \alpha(1-\alpha)$
tails	$u_2(z) = 1 - \phi_{a,b}(z)/\phi_{a,b}(a)$	$v_2(\alpha) = (2\alpha - 1)^2$
right tail	$u_3(z) = \Phi_{a,b}(z)$	$v_3(\alpha) = \alpha^2$
left tail	$u_4(z) = 1 - \Phi_{a,b}(z)$	$v_4(\alpha) = (1-\alpha)^2$

as the threshold decomposition of the continuous ranked probability score, and the quantile score representation (2.12). The equivalence to the kernel score representation (2.13) was noted and proved by Gneiting and Raftery (2007). Both (2.12) and (2.13) show that the continuous ranked probability score is reported in the same unit as the observations. The score is strictly proper within the class of the forecast densities that have finite first moment, and attains an infinite value otherwise. It applies to predictive distributions with discrete components, and reduces to the absolute error in the case of a point forecast.

The integrand in the traditional representation (2.11) equals the quadratic or Brier probability score (Selten 1998; Gneiting and Raftery 2007)

$$\mathrm{PS}(F(z), \mathbb{I}\{y \le z\}) = (F(z) - \mathbb{I}\{y \le z\})^2$$

for the probability forecast $F(z)$ of the binary event $\{Y \le z\}$ at the threshold $z \in \mathbb{R}$. The integrand in (2.12) equals the quantile score

$$\mathrm{QS}_\alpha(F^{-1}(\alpha), y) = 2\,(\mathbb{I}\{y \le F^{-1}(\alpha)\} - \alpha)\,(F^{-1}(\alpha) - y)$$

for the quantile forecast $F^{-1}(\alpha)$ (Cervera and Muñoz 1996; Gneiting and Raftery 2007). It has also been referred to as the tick loss function (Giacomini and Komunjer 2005) or, more traditionally, as the asymmetric linear or lin-lin loss function (Koenker and Basset 1978; Christoffersen and Diebold 1996).

Using the Brier probability score, we define threshold weighted versions of the continuous ranked probability score as

$$\mathrm{S}(f, y) = \int_{-\infty}^{\infty} \mathrm{PS}(F(z), \mathbb{I}\{y \leq z\}) \, u(z) \, \mathrm{d}z, \tag{2.14}$$

where u is a nonnegative weight function on the real line; if $u \equiv 1$, this reduces to the unweighted score (2.11). Table 2.4 lists some potential weight functions that emphasize the center or tails of a variable's range. The threshold weight functions resemble the suggestions of Amisano and Giacomini (2007); however, in our implementation, the parameters are fixed and user specified, depending on the application at hand. For instance, in the case of inflation rates we fix a at the policy target. If the weight function is integrable, such as in the case of the center weight function $\phi_{a,b}$, the threshold-weighted continuous ranked probability score (2.14) is finite and bounded by the integral of the weight function. Other options for integrable weight functions with center emphasis include t and Laplace densities.

Similarly, we define quantile weighted versions of the continuous ranked probability score as

$$\mathrm{S}(f, y) = \int_{0}^{1} \mathrm{QS}_\alpha(F^{-1}(\alpha), y) \, v(\alpha) \, \mathrm{d}\alpha, \tag{2.15}$$

where v is a nonnegative weight function on the unit interval. Table 2.4 suggests weight functions with center or tail emphasis.

From a decision-theoretic perspective, the quantile score $\mathrm{QS}_\alpha(F^{-1}(\alpha), y)$ can be interpreted as follows. Suppose that a point forecast, x, for the future quantity y is sought, and that the ex post loss is $\mathrm{L}(x, y) = 2(1 - \alpha)|y - x|$ in the case of an overprediction ($y \leq x$), and $\mathrm{L}(x, y) = 2\alpha|y - x|$ in the case of an underprediction ($y > x$). In this setting, the optimal point forecast or Bayes rule is the α-quantile, $F^{-1}(\alpha)$, of the predictive distribution, F (Granger 1969; Matheson and Winkler 1976), and the quantile score, $\mathrm{QS}_\alpha(F^{-1}(\alpha), y)$, equals the corresponding loss. The unweighted continuous ranked probability score (2.12) assigns uniform weight, $v(\alpha) \equiv 1$, to the potential values of the asymmetry parameter $\alpha \in (0, 1)$. In many applications, such as those described by Pinson, Chevallier and Kariniotakis (2007) and Laio and Tamea (2008), this may not represent realistic assumptions on actual cost-loss ratios. However, the application at hand may suggest alternative, non-uniform quantile weight functions, of which we give an example in Section 2.4.2.

The threshold and quantile weighting approaches can be traced back at least to Matheson and Winkler (1976), who showed that the scoring rules in (2.14) and (2.15) are proper. The threshold weighting idea is also employed by Corradi and Swanson (2006a, pp. 194–195), though their emphases and terminology differ from ours.

Closed form expressions for the evaluation of (2.14) or (2.15) may or may not be available, but the computation of a suitably discretized approximate version is always feasible, to any degree of accuracy. In the case of threshold weighting, we approximate (2.14) by

$$\mathrm{S}(f,y) = \frac{y_u - y_l}{I-1} \sum_{i=1}^{I} w(y_i)\, \mathrm{PS}(F(y_i), \mathbb{I}\{y \le y_i\}) \quad \text{where} \quad y_i = y_l + i\, \frac{y_u - y_l}{I} \tag{2.16}$$

and (y_l, y_u) is the range of interest. In the case of the quantile weighted score, we approximate the integral in (2.15) by a discrete version,

$$\mathrm{S}(f,y) = \frac{1}{J-1} \sum_{j=1}^{J-1} v(\alpha_j)\, \mathrm{QS}_{\alpha_j}(F^{-1}(\alpha_j), y) \quad \text{where} \quad \alpha_j = \frac{j}{J}. \tag{2.17}$$

Note that the discrete versions themselves are proper scoring rules, that arise as special cases in (2.14) and (2.15) if the integral is taken with respect to a discrete Stieltjes measure rather than a weight function.

2.3.2 Asymptotic normality of the test statistic

Following Amisano and Giacomini (2007), we consider tests of equal forecast performance based on the test statistic

$$t_n = \sqrt{n}\, \frac{\overline{\mathrm{S}}_n^f - \overline{\mathrm{S}}_n^g}{\hat{\sigma}_n},$$

where

$$\overline{\mathrm{S}}_n^f = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{S}(\hat{f}_{t+k}, y_{t+k}) \quad \text{and} \quad \overline{\mathrm{S}}_n^g = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{S}(\hat{g}_{t+k}, y_{t+k}) \tag{2.18}$$

and $\hat{\sigma}_n^2$ is defined in (2.3). Under general conditions, t_n is asymptotically standard normal under the null hypothesis of vanishing expected score differentials, and the test will reject with probability tending to 1 under a fixed alternative. When S is a weighted logarithmic rule, Amisano and Giacomini (2007) prove these claims under regularity assumptions,[3] which include a mixing condition on the process $\{\boldsymbol{Z}_t\}$ defined in the introduction, boundedness of the weight function, consistency of $\hat{\sigma}_n^2$ as an estimate of

$$\sigma_n^2 = \mathrm{var}(\sqrt{n}\,(\overline{\mathrm{S}}_n^f - \overline{\mathrm{S}}_n^g)) > 0,$$

and moment conditions. In our case, in which S is a weighted version of the continuous ranked probability score, the same result holds, except for the moment condition, which now requires that

$$\mathbb{E}_{\hat{f}_{t+k}}|X|, \quad \mathbb{E}_{\hat{g}_{t+k}}|X| \quad \text{and} \quad \mathbb{E}|Y_{t+k}|^{2r} \quad \text{are finite for all} \quad t, \tag{2.19}$$

where the power $r \geq 2$ depends on the mixing condition. In the case of threshold weighting with an integrable (rather than just bounded) weight function, the moment condition can be dropped. In analogy to the arguments of Amisano and Giacomini (2007), these results can be proved by verifying the assumptions of Theorem 4 of Giacomini and White (2006). The only novel argument is in the derivation of the moment condition (2.19), which is presented in an appendix.

In real-world applications, the full set of assumptions cannot be verified; yet, the assumptions are plausible as approximations. Recall that the continuous ranked probability score attains an infinite value if the forecast density has infinite first moment. In this light, the first two conditions in (2.19) assure that each individual score is finite. The third condition stipulates that the true data generating density has a finite moment of order $2r$, where typically one can take $r = 2$. Hence, as a rule of thumb, the normal approximation for t_n is

[3] Amisano and Giacomini consider the case $k = 1$ only. The extension to a general prediction horizon $k \geq 1$ is straightforward. We wish to emphasize that our aforementioned concerns are not with the asymptotic arguments in Amisano and Giacomini (2007) nor with the weighting idea, which is appealing indeed. However, we disagree with the particular choice of a weighted logarithmic scoring rule for the test, which can lead to rejection in favor of an inferior forecast.

appropriate, unless the forecast densities have infinite moments of low order. In the case of threshold weighting with an integrable weight function, the moment condition can just be ignored.

Table 2.5 summarizes results for weighted CRPS tests in the simulation example of Section 2.2. The density forecasts $\hat{f}_{t+1}$ and $\hat{g}_{t+1}$ and the true data generating density have Gaussian tails, so the normal approximation for t_n is justified. In contrast to the respective results for weighted likelihood ratio tests, all t_n values are strongly negative, favoring f over its deliberately misspecified competitor, g.

2.3.3 Forecast diagnostics via threshold and quantile decomposition

The threshold and quantile decompositions of the CRPS carry over to mean scores, and in this latter form they can be used diagnostically, to assess strengths and deficiencies of density forecasting techniques.

Consider a mean score of the form (2.18). The threshold decomposition (2.11) applies to the mean score, in that

$$\overline{\mathrm{CRPS}}_n^f = \int_{-\infty}^{\infty} \overline{\mathrm{PS}}_n^f(z)\,\mathrm{d}z \tag{2.20}$$

where

$$\overline{\mathrm{PS}}_n^f(z) = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{PS}(\hat{F}_{t+k}(z), y_{t+k}) \tag{2.21}$$

denotes the mean Brier probability score for the probability forecast of the binary event $\{Y_{t+k} \leq z\}$ at the threshold $z \in \mathbb{R}$. Schumacher, Graf and Gerds (2003) and Gneiting, Balabdaoui and Raftery (2007) proposed a plot of the mean Brier score (2.21) versus z as a diagnostic tool, and coined the terms prediction error curve and Brier score plot, respectively. The representation (2.20) shows that the plot illustrates the threshold decomposition of the continuous ranked probability score.

Similarly, the quantile decomposition (2.12) suggests the representation

$$\overline{\mathrm{CRPS}}_n^f = \int_0^1 \overline{\mathrm{QS}}_n^f(\alpha)\,\mathrm{d}z, \tag{2.22}$$

where

$$\overline{\mathrm{QS}}_n^f(\alpha) = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{QS}_\alpha(\hat{F}_{t+k}^{-1}(\alpha), y_{t+k}). \tag{2.23}$$

Laio and Tamea (2007) proposed a plot of the mean quantile score (2.23) versus α as a diagnostic tool in the assessment of density forecasts. We adopt their suggestion, which illustrates the quantile decomposition (2.22) of the mean continuous ranked probability score.

Figure 2.1 applies the threshold decomposition (2.20) and the quantile decomposition (2.22) to the density forecasting techniques f and g in the simulation study described in Sections 2.2 and 2.3.2. It is apparent that f and g are on comparable footing in the lower tail, while f is superior in the center, which is in accordance with (2.10). As shown in Table 2.5, the mean continuous ranked probability score is 0.511 for f and 0.625 for g; this equals the integral under the respective curves. The weighted scores in the table correspond to weighted integrals.

2.4 Case Studies

2.4.1 Bank of England projections of quarterly inflation rates

The Bank of England's Monetary Policy Committee (MPC) has issued probabilistic forecasts of inflation rates and gross domestic product every quarter since February 1996 and November 1997, respectively, using fan charts to visualize the deciles of the predictive distributions (Wallis 2003, 2004; Clements 2004; Elder, Kapetanios, Taylor and Yates 2005; Mitchell and Hall 2005).[4]

We compare the Bank of England's density forecasts of inflation rates (RPIX) to those derived from a simplistic autoregressive time series model. The Bank of England employs potentially asymmetric two-piece normal distributions with parameters $\mu \in \mathbb{R}$ and $\sigma_1, \sigma_2 > 0$

[4] The quarterly Bank of England inflation report is available online at `http://www.bankofengland.co.uk/publications/inflationreport/`. Archived forecasts can be downloaded at `http://www.bankofengland.co.uk/publications/inflationreport/irprobab.htm`. Observed RPIX inflation rates are available at `http://www.statistics.gov.uk/StatBase/tsdataset.asp?vlnk=7173&More=Y` under Office of National Statistics code CDKQ. The rates are percentage changes over 12 months. The first quarter ranges from March to May, the second from June to August, and so on. Prior to the inception of the MPC, the Bank of England issued inflation forecasts from February 1993 to May 1997, which were retrospectively converted into density forecasts and added to the forecast archive.

Table 2.5: Threshold and quantile weighted CRPS tests for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (2.10).

Threshold Weight	Emphasis	$\overline{\mathrm{S}}^f_n$	$\overline{\mathrm{S}}^g_n$	$\hat{\sigma}_n$	t_n	P
$u_0(z) = 1$	uniform	0.511	0.625	0.317	-10.72	< 0.001
$u_1(z) = \phi(z)$	center	0.153	0.184	0.095	-10.01	< 0.001
$u_2(z) = 1 - \phi(z)/\phi(0)$	tails	0.129	0.163	0.097	-10.37	< 0.001
$u_3(z) = \Phi(z)$	right tail	0.258	0.343	0.227	-11.32	< 0.001
$u_4(z) = 1 - \Phi(z)$	left tail	0.254	0.281	0.098	-8.39	< 0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}^f_n$	$\overline{\mathrm{S}}^g_n$	$\hat{\sigma}_n$	t_n	P
$v_0(\alpha) = 1$	uniform	0.511	0.625	0.317	-10.72	< 0.001
$v_1(\alpha) = \alpha(1-\alpha)$	center	0.100	0.125	0.069	-10.99	< 0.001
$v_2(\alpha) = (2\alpha - 1)^2$	tails	0.113	0.125	0.045	-7.98	< 0.001
$v_3(\alpha) = \alpha^2$	right tail	0.157	0.198	0.116	-10.44	< 0.001
$v_4(\alpha) = (1-\alpha)^2$	left tail	0.155	0.177	0.069	-9.60	< 0.001

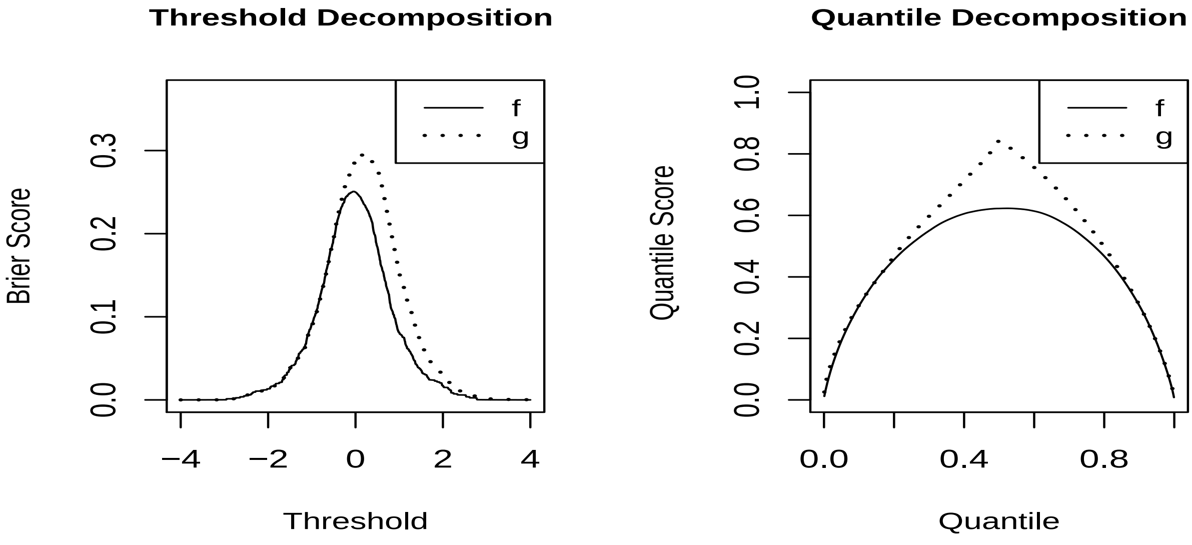

Figure 2.1: Threshold and quantile decomposition of the mean continuous ranked probability score for density forecasts for the conditionally heteroscedastic process (2.5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (2.10).

and forecast density

$$f(y) = \begin{cases} \left(\frac{\pi}{2}\right)^{-1/2} (\sigma_1 + \sigma_2)^{-1} \exp\left(-\frac{(y-\mu)^2}{2\sigma_1^2}\right) & \text{if} \quad y \leq \mu, \\ \left(\frac{\pi}{2}\right)^{-1/2} (\sigma_1 + \sigma_2)^{-1} \exp\left(-\frac{(y-\mu)^2}{2\sigma_2^2}\right) & \text{if} \quad y \geq \mu. \end{cases}$$

The simplistic competitor is a Gaussian autoregression of order one that uses a rolling estimation window of length $m = 6$ quarters. This method results in Gaussian density forecasts.

Figure 2.2 and Table 2.6 compare the two methods at a prediction horizon of $k = 1$ quarters ahead, for a test period ranging from the first quarter of 1993 to the first quarter of 2004, for a total of $n = 45$ density forecast cases. Figure 2.2 shows the threshold and quantile decompositions (2.20) and (2.22) of the continuous ranked probability score for the two techniques. The Bank of England forecast has a clear edge at almost all thresholds and quantiles, with a mean continuous ranked probability score of 0.112%, as opposed to 0.246% for the autoregressive forecast. The integrals under the corresponding curves in Figure 2.2 equal these values. The superiority of the Bank of England forecast is corroborated by Table 2.6, which reports the results of weighted CRPS tests, using the weight functions of Table 2.4, where $a = 2.5\%$ equals the MPC's 1997–2003 policy target and $b = 1.0\%$ reflects the relative constancy of the inflation rate during the evaluation period.

Figure 2.3 and Table 2.7 show results at a prediction horizon of $k = 7$ quarters ahead, for the third quarter of 1994 to the third quarter of 2005. Perhaps surprisingly, the dominance of the Bank of England forecast is much less pronounced. In Figure 2.3, the simplistic autoregressive forecast seems competitive at moderately large thresholds and quantiles. The mean continuous ranked probability score is 0.304% for the Bank of England forecast, as opposed to 0.382% for the autoregressive forecast. None of the tests in Table 2.7 rejects the null hypothesis of vanishing expected score differentials.

To explain this we consider Figure 2.4, which shows quantiles of the two density forecasts at a prediction horizon of seven quarters along with the realized inflation rates. The 90th percentile of the Bank of England forecast was much too conservative, resulting in

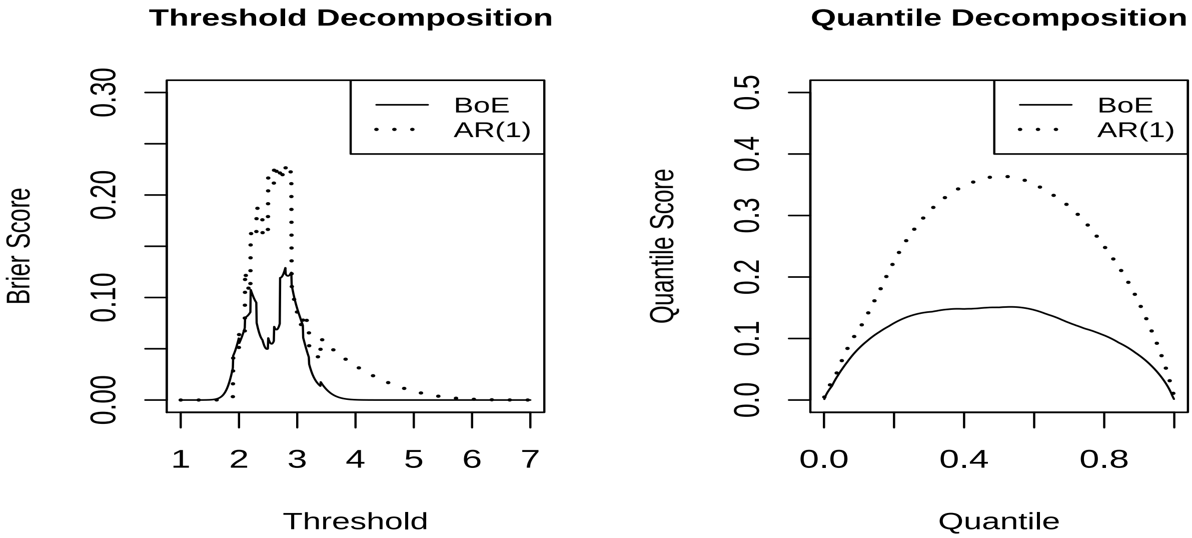

Figure 2.2: Threshold and quantile decomposition of the mean continuous ranked probability score for Bank of England (BoE) and autoregressive (AR) density forecasts of inflation rates, at a prediction horizon of one quarter.

Table 2.6: Threshold and quantile weighted CRPS tests for density forecasts of inflation rates, at a prediction horizon of one quarter, in percent. The Bank of England forecast takes the role of f and the autoregressive benchmark the role of g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$u_0(z) = 1$	uniform	0.112	0.246	0.248	-3.62	< 0.001
$u_1(z) = \phi_{2.5,1}(z)$	center	0.041	0.081	0.064	-4.16	< 0.001
$u_2(z) = 1 - \phi_{2.5,1}(z)/\phi_{2.5,1}(2.5)$	tails	0.010	0.044	0.137	-1.69	0.090
$u_3(z) = \Phi_{2.5,1}(z)$	right tail	0.061	0.152	0.200	-3.07	0.002
$u_4(z) = 1 - \Phi_{2.5,1}(z)$	left tail	0.051	0.094	0.076	-3.75	< 0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$v_0(\alpha) = 1$	uniform	0.112	0.246	0.248	-3.62	< 0.001
$v_1(\alpha) = \alpha(1-\alpha)$	center	0.022	0.049	0.050	-3.67	< 0.001
$v_2(\alpha) = (2\alpha - 1)^2$	tails	0.026	0.050	0.049	-3.35	< 0.001
$v_3(\alpha) = \alpha^2$	right tail	0.033	0.077	0.078	-3.78	< 0.001
$v_4(\alpha) = (1-\alpha)^2$	left tail	0.036	0.071	0.076	-3.12	0.002

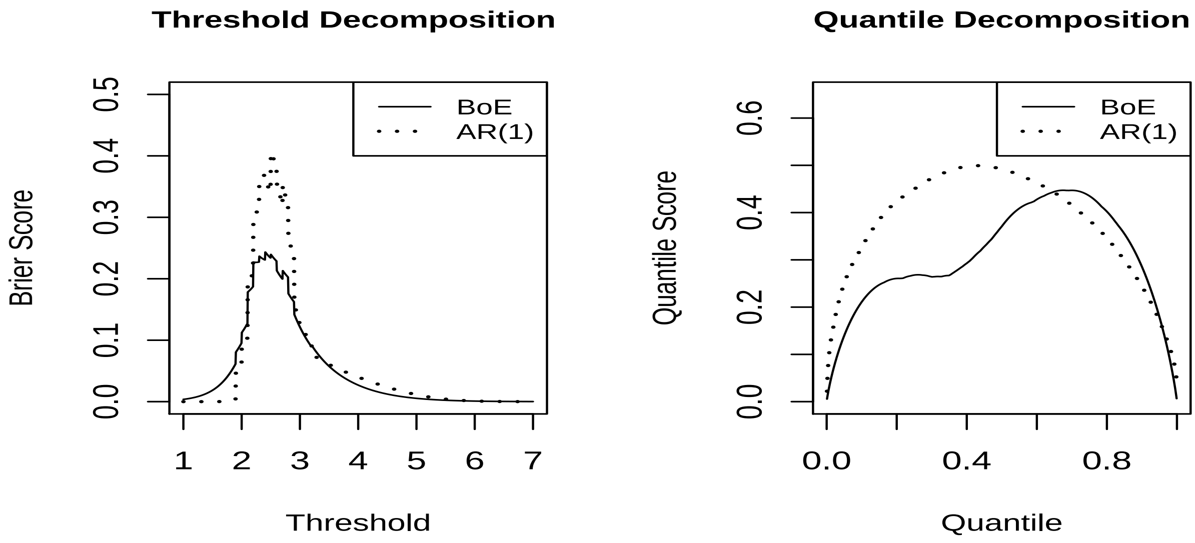

Figure 2.3: Threshold and quantile decomposition of the mean continuous ranked probability score for Bank of England (BoE) and autoregressive (AR) density forecasts of inflation rates, at a prediction horizon of seven quarters.

Table 2.7: Threshold and quantile weighted CRPS tests for density forecasts of inflation rates, at a prediction horizon of seven quarters. The Bank of England forecast takes the role of f and the autoregressive benchmark the role of g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$u_0(z) = 1$	uniform	0.304	0.381	0.437	-1.19	0.235
$u_1(z) = \phi_{2.5,1}(z)$	center	0.102	0.129	0.131	-1.43	0.152
$u_2(z) = 1 - \phi_{2.5,1}(z)/\phi_{2.5,1}(2.5)$	tails	0.049	0.057	0.166	-0.30	0.761
$u_3(z) = \Phi_{2.5,1}(z)$	right tail	0.170	0.226	0.324	-1.15	0.251
$u_4(z) = 1 - \Phi_{2.5,1}(z)$	left tail	0.134	0.155	0.148	-0.99	0.321

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$v_0(\alpha) = 1$	uniform	0.304	0.381	0.437	-1.19	0.235
$v_1(\alpha) = \alpha(1-\alpha)$	center	0.057	0.072	0.081	-1.23	0.217
$v_2(\alpha) = (2\alpha - 1)^2$	tails	0.077	0.095	0.124	-0.96	0.338
$v_3(\alpha) = \alpha^2$	right tail	0.108	0.111	0.080	-0.24	0.813
$v_4(\alpha) = (1-\alpha)^2$	left tail	0.083	0.127	0.263	-1.14	0.255

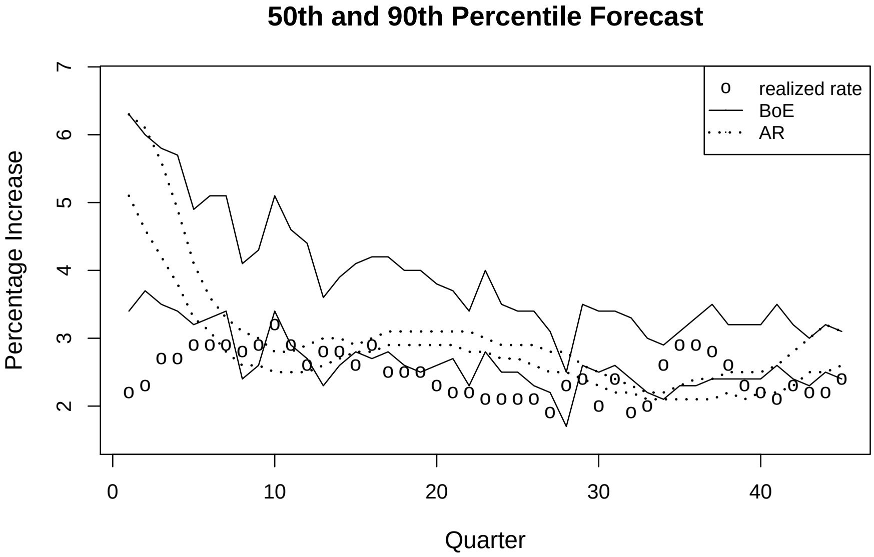

Figure 2.4: Bank of England (BoE) and autoregressive (AR) forecasts of inflation rates, at a prediction horizon of seven quarters ahead, for the third quarter of 1994 through the third quarter of 2005. The plot shows the 50th and 90th percentiles of the density forecasts for the two methods along with the observed rates.

unnecessarily wide prediction intervals that are penalized by the scores.

2.4.2 Probabilistic forecasts of wind resources at the Stateline wind energy center

With the proliferation of wind power, probabilistic short-term forecasts of wind resources at wind energy sites are becoming a critical requirement. Gneiting, Larson, Westrick, Genton and Aldrich (2006) introduced the regime-switching space-time (RST) technique that merges meteorological and statistical expertise to obtain accurate and calibrated, fully probabilistic forecasts of wind speed and wind power. Briefly, the RST method identifies forecast regimes at the wind energy site and fits a conditionally heteroscedastic predictive model for each regime. Geographically dispersed meteorological observations in the vicinity of the wind farm are used as predictor variables. The forecast densities are truncated normal.

Gneiting et al. (2006) applied the RST technique to obtain probabilistic forecasts of hourly average wind speed near the Stateline wind energy center in the US states of Oregon

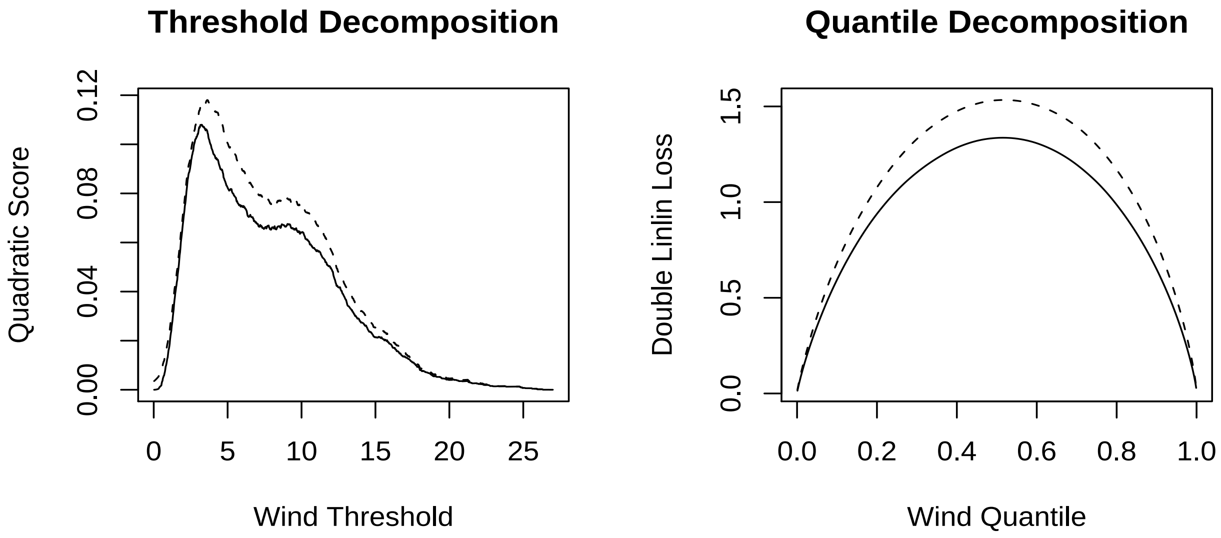

Figure 2.5: Threshold and quantile decomposition of the mean continuous ranked probability score for regime-switching space-time (RST) and autoregressive (AR) probabilistic forecasts of hourly average wind speed at the Stateline wind energy center, at a prediction horizon of two hours.

Table 2.8: Threshold and quantile weighted CRPS tests in the wind example. The regime-switching space-time (RST) forecast takes the role of f and the autoregressive benchmark the role of g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{RST}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$u_0(z) = 1$	uniform	0.961	1.115	0.838	-13.16	< 0.001
$u_1(z) = \phi_{10,5}(z)$	center	0.051	0.060	0.050	-12.29	< 0.001
$u_2(z) = 1 - \phi_{10,5}(z)/\phi_{10,5}(10)$	tails	0.318	0.364	0.293	-11.26	< 0.001
$u_3(z) = \Phi_{10,5}(z)$	right tail	0.342	0.398	0.386	-10.37	< 0.001
$u_4(z) = 1 - \Phi_{10,5}(z)$	left tail	0.619	0.718	0.552	-12.73	< 0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{RST}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$v_0(\alpha) = 1$	uniform	0.961	1.115	0.838	-13.17	< 0.001
$v_1(\alpha) = \alpha(1-\alpha)$	center	0.187	0.216	0.162	-12.93	< 0.001
$v_2(\alpha) = (2\alpha - 1)^2$	tails	0.213	0.250	0.201	-13.07	< 0.001
$v_3(\alpha) = \alpha^2$	right tail	0.299	0.351	0.302	-12.35	< 0.001
$v_4(\alpha) = (1-\alpha)^2$	left tail	0.288	0.331	0.252	-12.31	< 0.001
$v_5(\alpha) = \Delta_{0.73}(\alpha)$	peak at 0.73	0.564	0.654	0.504	-12.85	< 0.001

and Washington, at a prediction horizon of $k = 2$ hours. In what follows, we compare the RST density forecasts to probabilistic forecasts derived from autoregressive time series models, as proposed by Brown, Katz and Murphy (1984) and widely implemented since. Both methods employ a rolling estimation window of 45 days or $1,080$ hours. The evaluation period ranges from 1 May through 30 November 2003, for a total of $n = 5,136$ density forecast cases. See Gneiting et al. (2006) for details.[5]

Figure 2.5 shows the threshold and quantile decompositions of the continuous ranked probability score for the two probabilistic forecasting methods. The RST technique is superior at all thresholds and quantiles, with a mean continuous ranked probability score of 0.961 meters per second, as opposed to 1.115 meters per second for the autoregressive benchmark. Table 2.8 shows the results of weighted CRPS tests with the weight functions in Table 2.4, where $a = 10$ meters per second and $b = 5$ meters per second, a choice that is motivated by the marginal climatological distribution of wind speed at Stateline (Gneiting et al. 2006). All tests are overwhelmingly in favor of the RST technique.

In deregulated electricity markets power producers propose quantity-price bids in advance, and are charged for any imbalances. In this context, the optimal point forecast of a future wind speed is often the α-quantile of the predictive distribution (Pinson et al. 2007; Gneiting 2008b). The relevant value of α depends on the current market conditions, with the above references arguing that a typical value is $\alpha = 0.73$.[6] This suggests the use of a triangular quantile weight function, $v_5(\alpha) = \Delta_{0.73}(\alpha)$, which has a peak of height one at $\alpha = 0.73$ and decays to zero at $\alpha = 0$ and 1. The corresponding results are also shown in

[5] Gneiting et al. (2006) refer to the methods considered here as the RST-D-CH and AR-D-CH techniques. The autoregressive method assumes Gaussian forecast densities that assign small but positive probability mass to the negative halfaxis, which we reassign to wind speed zero. The continuous ranked probability score handles this point mass naturally.

[6] More specifically, the quantity of economic interest is the power output, which is commonly modeled as a nondecreasing nonlinear function, g, of the wind speed. If x denotes the point forecast and y the realizing wind speed, the predicted (and contracted) power output equals $g(x)$, while $g(y)$ units are generated. In the case of an underforecast, the producer's loss is proportional to the uncontracted surplus, $g(y) - g(x)$. In the case of an overprediction, it is proportional to the unrealized contracted power, $g(x) - g(y)$. The optimal point forecast then is the α-quantile of the predictive distribution (Gneiting 2008b). The proportionality constants are called regulation unit costs and typically are distinct. In a detailed analysis, Pinson et al. (2007) argue that regulation unit costs relative to negative imbalances ($y > x$) tend to be higher than those associated with positive imbalances, with the ratio, $r = \alpha/(1 - \alpha)$, of their yearly averages being equal to 2.7, which corresponds to $\alpha = 0.73$.

Table 2.8.

2.5 Discussion

We have proposed a method for comparing density forecasts that is based on threshold and quantile weighted versions of the continuous ranked probability score. R code is available from the author upon request.

Our approach is similar in spirit to the weighted likelihood ratio test of Amisano and Giacomini (2007); however, it is based on proper scoring rules, and therefore avoids misguided inferences. In the case of threshold weighting, it is formally equivalent to the approach of Corradi and Swanson (2006a), who provide a wealth of relevant theoretical results under rolling and recursive estimation schemes. The threshold and quantile decompositions of the continuous ranked probability score can be illustrated graphically, to provide diagnostic tools that prompt insights into the strengths and deficiencies of forecasting methods, as we have illustrated in the case studies.

Diks, Panchenko and van Dijk (2008) discuss the use of scoring rules for evaluating density forecasts in tails. Their paper also notes the impropriety of the weighted logarithmic score, and instead proposes test statistics of the form (2.2) that are based on conditional likelihood (CL; their eq. (10)) or censored likelihood (CSL; their eq. (11)) scoring rules. These scoring rules equal the standard logarithmic score on collapsed sample spaces and thus they are proper, but not strictly proper. The CL scoring rule is tailored to situations in which one wishes to assess predictive performance under conditions that depend on the outcome of the predictand. For example, it can be used to assess inflation forecasts conditional on the verifying rate exceeding the policy target. The CSL scoring rule allows for evaluations that emphasize regions of interest, similarly to the setting in our case studies.

Gneiting et al. (2007) contend that the goal of probabilistic forecasting is to maximize the sharpness of the forecast densities subject to calibration. Calibration refers to the statistical consistency between the forecast densities and the observations, and is a joint property of the forecasts and the values that materialize. Sharpness refers to the concentration of the forecast densities: The sharper the densities, the less the uncertainty, and the sharper, the better, subject to calibration.

The probability integral transform (PIT) histogram is the primary diagnostic tool for calibration checks (Diebold et al. 1998; Corradi and Swanson 2006b; Gneiting et al. 2007; Laio and Tamea 2007). The PIT is simply the value that the predictive CDF attains at the observation (Dawid 1984). If the observation is drawn from the forecast density, the PIT has a uniform distribution. Hence, to assess the calibration of a density forecasting method, one finds the PIT, repeats over a sizable number of forecast cases, and checks the PIT histogram for uniformity. However, this does not take the sharpness of the density forecasts into account, as opposed to proper scoring rules, which provide a combined assessment of calibration and sharpness (Gneiting et al. 2007).

A possible limitation of our method is that the unweighted continuous ranked probability score is infinite if the forecast density has infinite first moment, such as in the case of a Cauchy density. Even then, the mean scores (2.21) and (2.23) can be plotted versus the threshold z and the quantile α, and the resulting plots can be interpreted diagnostically. Furthermore, the threshold-weighted continuous ranked probability score (2.14) is finite if the weight function is integrable, and in this latter form the weighted CRPS test continues to apply.

Appendix: Moment Conditions

We supply the remaining nontrivial arguments in Section 2.3.2. To verify the assumptions of Theorem 4 of Giacomini and White (2006), we need to show that the moment condition (2.19) implies

$$\mathbb{E}\,|\mathrm{S}(\hat{f}_{t+k}, Y_{t+k}) - \mathrm{S}(\hat{g}_{t+k}, Y_{t+k})|^{2r} \tag{2.24}$$

to be finite, where S is the threshold-weighted continuous ranked probability score (2.14) or the quantile-weighted score (2.15), and the weight function is bounded. For ease of notation, we substitute f, g and Y for $\hat{f}_{t+k}$, $\hat{g}_{t+k}$ and Y_{t+k}, respectively. If the weight function is bounded above by the constant $M > 0$, then

$$\mathbb{E}\,|\mathrm{S}(f, Y) - \mathrm{S}(g, Y)|^{2r} \leq (2M)^{2r}\,(\mathbb{E}\,\mathrm{CRPS}(f, Y)^{2r} + \mathbb{E}\,\mathrm{CRPS}(g, Y)^{2r}).$$

We proceed to show that under (2.19) both $\mathbb{E}\,\mathrm{CRPS}(f,Y)^{2r}$ and $\mathbb{E}\,\mathrm{CRPS}(g,Y)^{2r}$ are finite. If X and X' are independent random variables with density f that are independent of Y, then

$$\mathrm{CRPS}(f,Y) = \mathbb{E}|X-Y| - \frac{1}{2}\mathbb{E}_f|X-X'| \leq 2\mathbb{E}_f|X| + |Y|$$

by (2.13) and the triangle inequality, and therefore

$$\mathbb{E}\,\mathrm{CRPS}(f,Y)^{2r} \leq\, 2^{2r}\left((2\mathbb{E}_f|X|)^{2r} + \mathbb{E}|Y|^{2r}\right).$$

A similar result holds for $\mathbb{E}\,\mathrm{CRPS}(g,Y)^{2r}$; hence, (2.19) is a sufficient condition for the expectation (2.24) to be finite.

Finally, if the threshold weight function u in (2.14) is integrable, the score differential in (2.24) is bounded and its moments of order $r \geq 2$ are finite.

Chapter 3

COMBINING PROBABILITY FORECASTS

3.1 Introduction

Probabilistic forecasts take account of the uncertainty in a prediction, by taking the form of a predictive probability distribution for a future quantity or event. The simplest case is that of a future binary or dichotomous event, such as a recession versus no recession, or rain versus no rain. In the binary case, a predictive probability distribution is simply an ex ante probability for the event to happen. While the roots of probability forecasting can be traced back to the 18th century, the transition to probability of precipitation forecasts by the US National Weather Service in 1965 was perhaps the most influential and important event in their development (Murphy 1998; Winkler and Jose 2008). In economics, the Survey of Professional Forecasters has included probability variables since 1968 (Croushore 1993). Of course, there are many other important applications of probability forecasts, including but not limited to medical diagnosis (Spiegelhalter 1986), educational testing, and political and socio-economic foresight (Tetlock 2005). Arguably, a far-reaching transdisciplinary transition to distributional forecasting is well under way (Gneiting 2008).

In many instances, multiple probability forecasts for the same event are available. In surveys, economic experts might provide diverse probability assessments of a future recession. Distinct numerical and/or statistical models might provide a collection of probability of precipitation forecasts, and a group of physicians might assign individual survival probabilities. In this type of situation, there is strong empirical evidence that combined probability forecasts that draw an all the experts' or models' strengths result in improved predictive performance. This is very much in the spirit of model averaging, which has primarily been developed for the purpose of statistical inference (Hoeting et al. 1999).

Various ways of combining probability forecasts into a single aggregated forecast have been proposed. Genest and Zidek (1986), Wallsten et al. (1997), Clemen and Winkler (2007) and Primo et al. (2009) provide excellent reviews. In practice, most aggregation techniques rely on a weighted linear combination of the individual probability forecasts, which is often referred to as a linear opinion pool (Stone 1961; Genest and McConway 1990; DeGroot and Mortera 1991). Substantial empirical evidence attests to the benefits of linear opinion pools, with successful applications ranging from meteorology (Sanders 1963; Vislocky and Fritsch 1995; Baars and Mass 2005) to economics (Graham 1996), psychology (Ariely et al. 2000), and medical diagnosis (Winkler and Poses 1993), among other fields.

The goal in probability forecasting is to maximize the sharpness of the forecast distributions subject to calibration (Murphy and Winkler 1987; Gneiting, Balabdaoui and Raftery 2007; Pal 2009). Calibration or reliability measures how close conditional event frequencies are to the forecast probabilities. Sharpness describes how far away the forecasts are from the naive, climatological baseline forecast, that is, the marginal event frequency (Gneiting et al. 2008; Winkler and Jose 2008). The more extreme the forecast probabilities are, that is, the closer to the most confident values of zero or one, the sharper the forecast. Strictly proper scoring rules such as the Brier or quadratic score (Brier 1950; Selten 1998) and the logarithmic score (Good 1952) provide summary measures of predictive performance that address calibration and sharpness simultaneously (Gneiting and Raftery 2007).

It is therefore critical that probability assessments are aggregated in ways that promote calibrated and sharp combined forecasts. In Section 3.2 we demonstrate a striking result, in that any weighted linear combination of distinct, individually calibrated probability forecasts is necessarily uncalibrated and lacks sharpness. In this light, linear opinion pools are suboptimal, so in Section 3.3 we propose a nonlinear generalization, the beta-transformed linear opinion pool (BLP). The BLP method fits an optimally recalibrated forecast combination, by compositing a beta transform and the traditional linear opinion pool. Section 3.4 illustrates the BLP method in a case study on statistical and National Weather Service probability of precipitation forecasts at 29 major cities in the continental US. The BLP combined forecast is calibrated and sharp and outperforms the individual and linearly combined forecasts. The chapter closes with a discussion in Section 3.5.

3.2 Some Shortcomings of Linearly Combined Probability Forecasts

The overarching message in this section is that linear opinion pools are generally uncalibrated, even in the ideal case in which each individual source is calibrated. We give a rigorous probabilistic version of this result in Theorem 3.2.1, which is then illustrated in a simulation study.

3.2.1 Theoretical results

We work within a probabilistic framework which considers the joint distribution of the random vector

$$(Y, p_1, \ldots, p_k),$$

where $Y \in \{0, 1\}$ is a binary or dichotomous event, and $0 \leq p_1, \ldots, p_k \leq 1$ are probability forecasts that take values in the closed unit interval. This is akin to the setting in DeGroot and Fienberg (1982, 1983) and Murphy and Winkler (1987), but considers an arbitrary number, k, of individual probability forecasts, each of which is a random variable, with full generality in the joint dependence structure. In this framework a probability forecast is any random variable, p, that is measurable with respect to the σ-algebra generated by $p_1, \ldots, p_k$, with the linear opinion pool,

$$p = w_1 p_1 + \cdots + w_k p_k \quad \text{where} \quad w_1, \ldots, w_k \geq 0 \quad \text{and} \quad w_1 + \cdots + w_k = 1, \tag{3.1}$$

being one such example. The probability forecast p is calibrated for Y if

$$\mathbb{P}\,(Y = 1|p) = \mathbb{E}\,(Y|p) = p \qquad \text{almost surely.}$$

This definition is in accordance with the economic, meteorological, psychological and statistical forecasting literature and can be traced to Murphy and Winkler (1987) and Schervish (1989). It differs from the game-theoretic approach to calibration that has been developed in a far-reaching, related strand of literature (Dawid 1982; Foster and Vohra 1998; Lehrer 2001; Sandroni, Smorodinsky and Vohra 2003; Vovk and Shafer 2005). From the basic properties of conditional expectations, it is immediate that if p is a calibrated probability forecast then

$$\mathbb{E}p = \mathbb{E}\mathbb{E}\,(Y|p) = \mathbb{E}Y.$$

This latter property can be thought of as a weak form of calibration, and we refer to it as marginal consistency. It resembles the notion of marginal calibration for probabilistic forecasts of continuous variables (Gneiting, Balabdaoui and Raftery 2007).

A scoring rule assigns a numerical score, $\mathrm{S}(x, y)$, to the probability forecast $x \in [0, 1]$ and the binary event y, where $y = 1$ if the event occurs and $y = 0$ otherwise. We consider scoring rules to be negatively oriented penalties, that is, the smaller the better. A scoring rule is strictly proper if it encourages honest assessments, that is, if

$$x\,\mathrm{S}(x, 1) + (1 - x)\mathrm{S}(x, 0) < x\,\mathrm{S}(x', 1) + (1 - x)\mathrm{S}(x', 0) \qquad \text{for all} \qquad 0 \le x \ne x' \le 1.$$

See Dawid (1986), Winkler (1996) and Gneiting and Raftery (2007) for reviews and discussion.

We are now in a position to state our key result. The proof is deferred to the Appendix.

Theorem 3.2.1. *Suppose that $p_1, \ldots, p_k$ are calibrated for the binary event Y and such that $p_i \ne p_j$ with strictly positive probability for at least one pair $i \ne j$. Consider the linear opinion pool,*

$$p = w_1 p_1 + \cdots + w_k p_k,$$

where $w_1, \ldots, w_k > 0$ and $w_1 + \cdots + w_k = 1$. Let

$$q = \mathbb{P}(Y = 1|p) = \mathbb{E}(Y|p)$$

denote the recalibrated version of p, that is, the conditional probability of Y given p. Then the following holds.

(a) *The linear opinion pool p lacks calibration, in that $q \ne p$ with strictly positive probability.*

(b) *The linear opinion pool p lacks sharpness, in that*

$$\mathbb{E}(p-p_0)^2 < \mathbb{E}(q-p_0)^2 \qquad \textit{where} \qquad p_0 = \mathbb{E}p = \mathbb{E}q = \mathbb{E}Y.$$

In words, both p and q are marginally consistent, but on average p is closer to its expectation, the naive climatological forecast p_0, than its recalibrated version, q.

(c) *The recalibrated forecast q is calibrated, that is, $\mathbb{P}(Y = 1|q) = q$ almost surely, and it outperforms p, in that*

$$\mathbb{E}\mathrm{S}(q,Y) \; < \; \mathbb{E}\mathrm{S}(p,Y)$$

for every strictly proper scoring rule S*.*

The statement about the lack of calibration of the linear opinion pool in part (a) is our key result. It is counter to a natural belief that linear pools of calibrated probability forecasts are calibrated, as recently expressed by Iversen, Parmigiani and Chen (2008, p. 899). A similar result which applies to the case of multiple density forecasts for a continuous quantity was proved by Hora (2004). This uses a very different mode of calibration, and there is no apparent way of deducing our result from Hora's, or vice versa.

The result in part (a) is an immediate consequence of the stronger statement in part (b), since the latter implies that p cannot equal q with probability 1. Part (b) hints at the nature of the departure from the recalibrated forecast, q, and is expressed in terms of the expected squared deviation from the climatological baseline probability, p_0. For a sharp forecast, the forecast probabilities are close to zero or one, so the larger this deviation the sharper the forecast (Murphy and Winkler 1992; Gneiting et al. 2008; Winkler and Jose 2008). As a result, the linear opinion pool is underconfident. Part (c) demonstrates the superiority of the recalibrated forecast in terms of strictly proper scoring rules (Gneiting and Raftery 2007) and is akin to Theorem 6.3 of Schervish (1989). Proper scoring rules address calibration and sharpness simultaneously, so in view of parts (a) and (b) this is an unsurprising result.

We proceed to discuss related results in the literature. Dawid, DeGroot and Mortera (1995) studied the problem of the coherent combination of probability forecasts. Briefly, a *coherent combination formula* is a function $f : [0,1]^k \to [0,1]$ such that

$$f(p_1, \ldots, p_k) = \mathbb{P}(Y = 1 | p_1, \ldots, p_k) \tag{3.2}$$

under some joint distribution of the random vector $(Y, p_1, \ldots, p_k)$. The conditional probability property (3.2) implies calibration. Thus, part (a) of Theorem 3.2.1 shows, for any $k \geq 2$, that any nontrivial linear combination formula with nonnegative coefficients is incoherent. This generalizes a result of Dawid, DeGroot and Mortera (1995, p. 275) which applies in the case of $k = 2$ sources.

Theorem 2 of Wallsten and Diederich (2001) considers the combination of expert probability judgements, assuming that the assessments are conditionally independent and that each expert's expressed (overt) opinion is a monotone stochastic transform of a hidden (covert) opinion, which is calibrated. Then the arithmetic mean of the expert opinions becomes increasingly diagnostic of the future event as the number of experts grows to infinity, roughly in the sense that if the mean exceeds $\frac{1}{2}$ the true conditional event probability converges to 1, and otherwise converges to 0. Consequently, the calibration curve for the arithmetic mean of the expert opinions becomes sigmoidal with a fixed point at $\frac{1}{2}$. In contrast to our Theorem 3.2.1, which is a finite sample result and does not make any assumptions on the dependence structure, Wallsten and Diederich (2001) rely critically on the asymptotic scenario and conditional independence.

Another related result is Theorem 4.1 of Genest and Schervish (1985), which adopts a Bayesian point of view and derives a formula for the posterior opinion of a decision maker. Similar to Wallsten and Diederich's (2001) findings, the true posterior opinion converges to 1, or 0, if the individual judgements lie above, or below, $\frac{1}{2}$. This result also depends on the conditional independence of the individual probability assessments.

Despite Theorem 3.2.1 being critical of the linear opinion tool, there is overwhelming empirical evidence that linearly combined probability forecasts outperform individual forecasts. This is not a contradiction and can readily be explained, by noting that linear opinion pools outperform individual forecasts, but are suboptimal themselves, and can potentially be improved upon by using nonlinear recalibration methods.

3.2.2 Simulation study

We now illustrate our theoretical findings in a simulation study. First we describe a statistical model that gives rise to a joint distribution for the binary event Y and probability forecasts p_1 and p_2, which represent forecasters with access to independent sources of information. Then we define linearly combined forecasts and assess calibration.

Specifically, let

$$p = \Phi(a_1 + a_2),$$

where $a_1 \sim \mathcal{N}(0, 1)$ and $a_2 \sim \mathcal{N}(0, 2)$ are independent random variables and Φ denotes the standard normal cumulative distribution function. Suppose that Y is a Bernoulli random variable with conditional success probability

$$\mathbb{P}(Y = 1 |\, p) = \mathbb{E}(Y \,|\, p) = p.$$

Forecaster 1 has access to a_1 only. This assessor's **probability forecast $\boldsymbol{p_1}$** is the conditional event probability

$$p_1 = \mathbb{P}(Y = 1|a_1) = \mathbb{E}(Y|a_1) = \mathbb{E}(p|a_1) = \mathbb{E}\left[\Phi(a_1 + a_2)|a_1\right] = \Phi\left(\frac{a_1}{\sqrt{3}}\right). \tag{3.3}$$

The second forecaster has knowledge of source a_2 only, whence **probability forecast $\boldsymbol{p_2}$** becomes

$$p_2 = \mathbb{P}(Y = 1|a_2) = \Phi\left(\frac{a_2}{\sqrt{2}}\right). \tag{3.4}$$

The final equality stems from the fact that if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbb{E}\,\Phi(X) = \Phi(\mu/\sqrt{\sigma^2 + 1})$, which is proved in the Appendix. Evidently, p, p_1 and p_2 are calibrated.

We take p_1 and p_2 as the individual forecasts from which we form combinations, namely the **equally weighted linear opinion pool (ELP)**, that is, the equally weighted average of p_1 and p_2, and an **optimally weighted linear opinion pool (OLP)**. The OLP weights for p_1 and p_2 are estimated on a training sample of size 10,000, using the maximum likelihood method and the special case of the log likelihood function (3.10) below, in which $\alpha = \beta = 1$. Table 3.1 shows the OLP estimates and their standard errors. The second individual forecast, p_2, which resolves events and non-events more successfully, obtains a substantially higher OLP weight, w_2, of about $\frac{3}{4}$.

In the simulation experiment, we consider an independent test sample of size 10,000 from the joint distribution of Y, p_1 and p_2 and generate the combined ELP and OLP forecasts. Figure 3.1 shows empirical calibration curves or reliability diagrams (Sanders 1963; Pocernich 2009) for the four types of forecasts, which plot the conditional empirical event frequency versus the forecast probability. The circles show the conditional empirical frequency; the broken lines give pointwise 95% lower and upper critical values under the null hypothesis of calibration, obtained with the bootstrap technique of Bröcker and Smith (2007). Significant deviations from the diagonal suggest that a forecast is uncalibrated. The inset histograms show the frequency distribution of the forecast probabilities and can be used diagnostically to assess sharpness.

The calibration curves for the individual forecasts, p_1 and p_2, show that they are empirically well calibrated, and the inset histograms confirm that p_2 is the sharper forecast, with forecast probabilities that are further away from the climatological event frequency, $p_0 = \frac{1}{2}$. The linearly pooled ELP and OLP forecasts are empirically uncalibrated. The direction of departure is as anticipated, towards underconfidence, and the extent of the lack of calibration is startling, even for the optimally weighted OLP forecast.

3.3 Recalibration

We have seen that the linear opinion pool yields a suboptimal combined probability forecast, in that it is uncalibrated even in the ideal case in which the individual sources are calibrated. If the individual forecasts are uncalibrated, the need for recalibration typically is even more pronounced. Before proposing a method that addresses these issues by applying a recalibration transform to the linear opinion pool, we digress to discuss a theoretically optimal approach to forecast aggregation.

We have chosen to work in a probabilistic setting that considers the joint distribution of the binary event and the individual probability forecasts. In this framework, the theoretically optimal combined forecast, $\hat{p}$, is the conditional probability (CP), or conditional

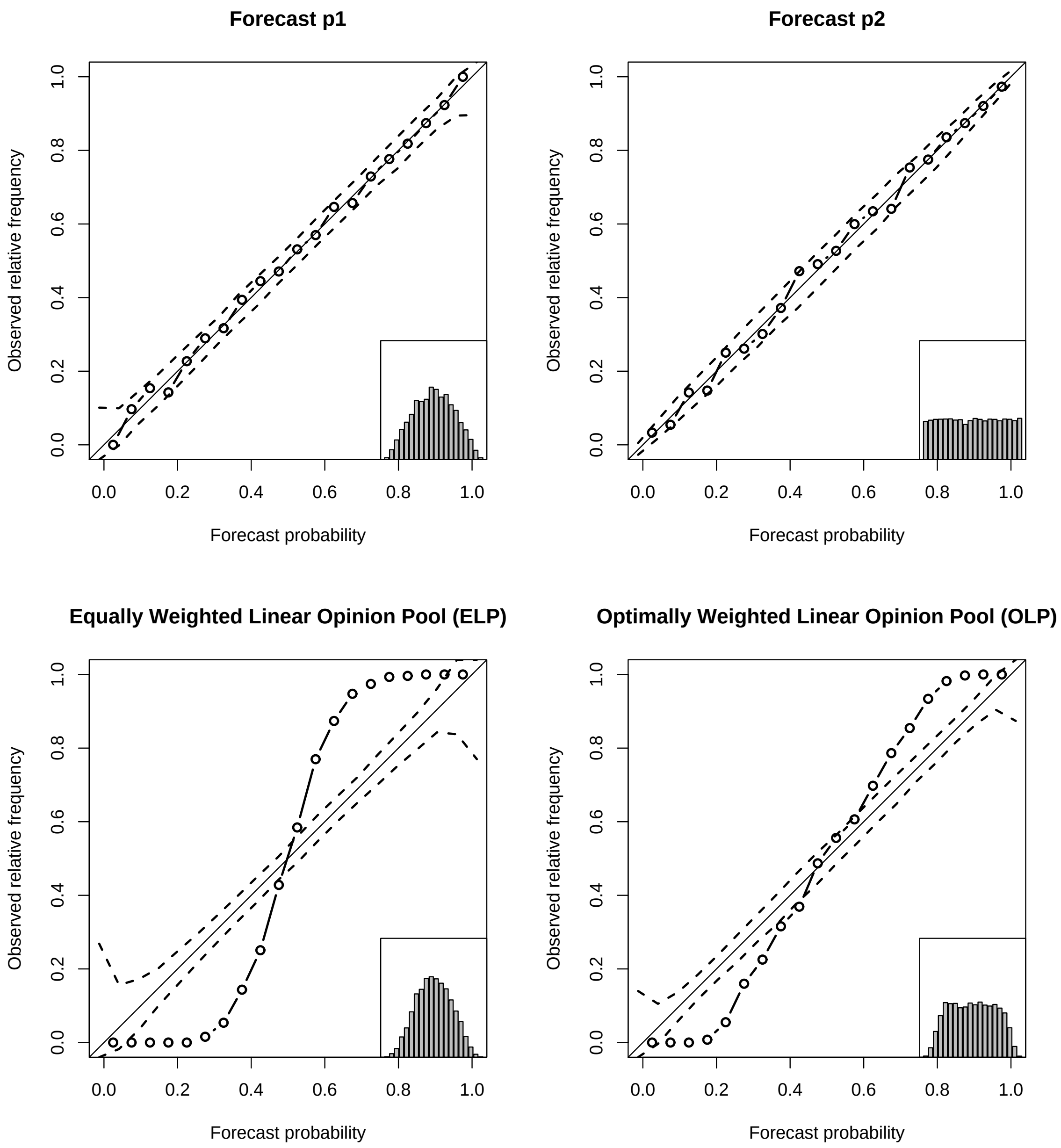

Figure 3.1: Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the individual and linearly combined forecasts in the simulation example of Section 3.2.2. The histograms show the empirical distribution of the forecast values over the unit interval.

expectation of the binary event Y, given the individual forecasts $p_1, \ldots, p_k$, that is,

$$\hat{p} = \mathbb{P}(Y = 1|p_1, \ldots, p_k) = \mathbb{E}(Y|p_1, \ldots, p_k). \tag{3.5}$$

By definition, this is the best approximation of the binary random variable Y in terms of the individual probability forecasts, $p_1, \ldots, p_k$, in the sense that $\mathbb{E}(\hat{p} - Y)^2 \leq \mathbb{E}(p - Y)^2$ for all functions p that are measurable with respect to the σ-algebra generated by $p_1, \ldots, p_k$. Hence, $\hat{p}$ minimizes the expected Brier score and, indeed, the expectation of any strictly proper scoring rule S, in that

$$\begin{aligned}
\mathbb{E}\,\mathrm{S}(\hat{p}, Y) &= \mathbb{E}\mathbb{E}\,[\mathrm{S}(\hat{p}, Y)|p_1, \ldots, p_k] \\
&= \mathbb{E}\,[\hat{p}\mathrm{S}(\hat{p}, 1) + (1 - \hat{p})\mathrm{S}(\hat{p}, 0)] \\
&\leq \mathbb{E}\,[\hat{p}\mathrm{S}(p, 1) + (1 - \hat{p})\mathrm{S}(p, 0)] \\
&= \mathbb{E}\mathbb{E}\,[\mathrm{S}(p, Y)|p_1, \ldots, p_k] \\
&= \mathbb{E}\,\mathrm{S}(p, Y),
\end{aligned}$$

with equality if and only if $p = \hat{p}$ almost surely. Under the conditions of Theorem 3.2.1, the conditional probability $\hat{p}$ is a necessarily nonlinear function of the individual forecasts, except for some special cases in which it is linear with at least one coefficient being negative (Dawid, DeGroot and Mortera 1995). In the simulation study of Section 3.2.2 there are two individual forecasts, p_1 and p_2, and the conditional probability (3.5) equals

$$\begin{aligned}
\hat{p} &= \mathbb{P}(Y = 1|p_1, p_2) \\
&= \mathbb{P}(Y = 1|a_1, a_2) \\
&= \Phi(a_1 + a_2) \\
&= \Phi(\sqrt{3}\,\Phi^{-1}(p_1) + \sqrt{2}\,\Phi^{-1}(p_2)). \qquad (3.6)
\end{aligned}$$

This is of the generalized linear form (24) of Dawid, DeGroot and Mortera (1995) with a normal quantile link function.

3.3.1 The beta-transformed linear opinion pool (BLP)

In the practice of forecasting, the functional form of the conditional probability (3.5) is unknown and needs to be estimated from training data. Nonparametric approaches can be attempted; however, we have chosen parsimonious, yet flexible parametric approximations. Nonparametric approaches are likely to suffer from the curse of dimensionality issue because, as the number of forecasters grows the difficulty in estimating conditional probability accurately will grow exponentially. For the parametric method we only need one additional parameter corresponding to a new forecaster. So, the number of parameters grows linearly and thus we will avoid the curse of dimensionality issue. Specifically, our preferred approach to aggregating individual probability forecasts, $p_1, \ldots, p_k$, is to first form a linear opinion pool, and then to apply a beta transform to achieve calibration. We call this the beta-transformed linear opinion pool (BLP), which takes the form

$$p = H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_i\right), \tag{3.7}$$

where $w_1, \ldots, w_k \geq 0$ and $w_1 + \cdots + w_k = 1$, and

$$H_{\alpha,\beta}(x) \quad = \quad B(\alpha,\beta)^{-1} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t \qquad \text{for} \qquad x \in [0,1],$$

is the cumulative distribution function of the beta density with shape parameters $\alpha > 0$ and $\beta > 0$. Note that the BLP model nests the traditional linear opinion pool that arises in the special case when $\alpha = \beta = 1$. If furthermore $w_1 = \cdots = w_k = \frac{1}{k}$ we recover the equally weighted linear opinion pool (ELP). While the use of the beta transform for the purpose of calibration dates back at least to Graham (1996), the statistical model (3.7) that merges the linear opinion pool with a parametric recalibration transformation appears to be new. It applies very generally and can be used to aggregate calibrated as well as uncalibrated sources.

In many cases, full generality in (3.7) may not be needed or desirable. For instance, it is often useful to assume that the recalibration transform, $H_{\alpha,\beta}$, satisfies

$$H_{\alpha,\beta}(x) \leq x \quad \text{for} \quad x \leq x_0 \qquad \text{and} \qquad H_{\alpha,\beta}(x) \geq x \quad \text{for} \quad x \geq x_0 \tag{3.8}$$

for some $x_0 \in (0,1)$. This can be enforced by putting conditions on α and β. For example, if the individual forecasts are calibrated, Theorem 3.2.1 suggests that the linear opinion pool is underconfident, in the sense that its calibration curve lies under the diagonal for small forecast probabilities, and above the diagonal for high probabilities, with a fixed point at some $x_0 \in (0,1)$. The aforementioned results of Wallsten and Diederich (2001) support the choice of $x_0 = \frac{1}{2}$, under which (3.8) can be enforced by requiring that

$$\alpha = \beta \geq 1. \tag{3.9}$$

If we aim to address the hard-easy effect that has been described in the psychological literature (Lichtenstein, Fischhoff and Phillips 1982; Kynn 2008, p. 253), the fixed point in (3.8) can be taken to be $x_0 = \frac{3}{4}$.

We now describe how we go about parameter estimation for the BLP model in (3.7). Suppose that $y_1, \ldots, y_n$ are binary observations in the training set. Let $p_{i1}, \ldots, p_{in}$ denote the respective individual probability forecasts, for sources $i = 1, \ldots, k$. The aggregated BLP forecast then takes the form

$$p_t = H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_{it}\right) \qquad \text{for} \qquad t = 1, \ldots, n,$$

where the index ranges over the instances in the training set. Assuming independence, the log likelihood function for the BLP model (3.7) can be expressed as

$$\begin{aligned} \ell(w_1, \ldots, w_k; \alpha, \beta) \; &= \; \sum_{t=1}^{n} \left(y_t \log p_t + (1-y_t) \log(1-p_t)\right) \\ &= \sum_{t=1}^{n} y_t \log H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_{it}\right) + \sum_{t=1}^{n} (1-y_t) \log\left(1 - H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_{it}\right)\right). \end{aligned} \tag{3.10}$$

We obtain maximum likelihood estimates of the weights $w_1, \ldots, w_k$ and the recalibration

Table 3.1: Maximum likelihood estimates of OLP and BLP parameters in the simulation example of Sections 3.2.2 and 3.3.2, with standard errors in brackets.

Method	w_1	w_2	α
OLP	0.246 (0.014)	0.754 (0.014)	
BLP	0.519 (0.005)	0.481 (0.005)	9.55 (0.35)

parameters α and β by numerically optimizing the log likelihood function (3.10) under the constraints that $w_1, \ldots, w_k \geq 0$, $w_1 + \cdots + w_k = 1$, $\alpha > 0$ and $\beta > 0$. As noted above, it is often appropriate to enforce further constraints, with (3.9) being one such example. The traditional, non-transformed linear opinion pool arises when $\alpha = \beta = 1$. Estimated standard errors can be obtained in the usual way, by inverting a numerical approximation to the Hessian of the log likelihood function at the maximum likelihood estimates. We believe this would be a correct estimate of standard error as long as the parameter estimates lie in the open interior of the parameter space. The estimates can also be interpreted as optimum score estimates based on the logarithmic scoring rule, in the sense described by Gneiting and Raftery (2007, p. 375). This latter interpretation does not rely on any assumption of independence, and our results in concert with those of Wilks (1991) suggest robustness to non-independence.

3.3.2 Simulation study revisited

We return to the simulation study in Section 3.2.2 and fit the **beta-transformed linear opinion pool (BLP)** to combine the individual probability forecasts, p_1 and p_2. Then we compare to the theoretically optimal forecast, the **conditional probability (CP)** forecast (3.5) which here has the closed form solution (3.6).

Recall that both p_1 and p_2 are calibrated, so, as we explain in the previous section, we estimate the BLP model (4.7) under the constraint in (3.9), that is, we assume that $\alpha = \beta \geq 1$. Table 3.1 shows maximum likelihood estimates for the BLP parameters and compares to the respective OLP estimates. The individual forecasts, p_1 and p_2, get approximately equal

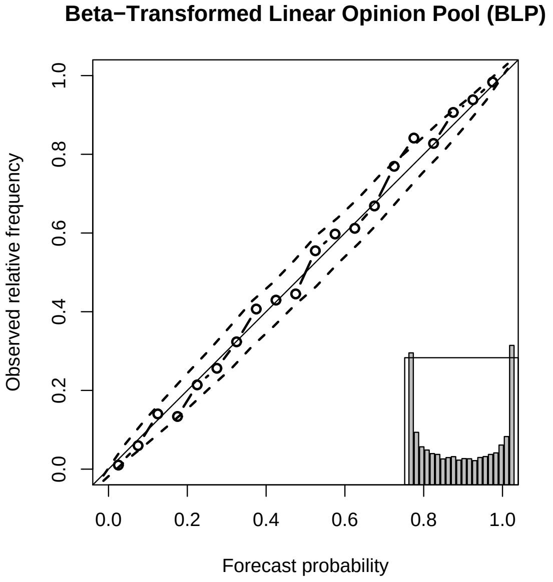

Figure 3.2: Calibration curve and 95% bootstrap intervals under the null hypothesis of calibration for the BLP forecast in the simulation example of Sections 3.2.2 and 3.3.2, for the same independent test sample as that used before. The histogram shows the empirical distribution of the forecast values over the unit interval.

BLP weights, much in contrast to the OLP model. The estimate for the BLP recalibration parameter, α, is far from the identity transform that arises when $\alpha = 1$, reflecting the striking lack of calibration of the traditional linear opinion pool.

Have we succeeded in our goal of approximating the theoretically optimal CP forecast (3.6) by the estimated, nonlinearly aggregated BLP model (3.7)? The empirical calibration curve for the BLP forecast in Figure 3.2 does not show any systematic departure from the diagonal, and the inset histogram shows that it is much sharper than any of the individual or linearly combined forecasts. A more detailed analysis reveals that if $0 < p_1 = p_2 < 1$ the maximal difference between the CP forecast and the fitted BLP model is 0.0215.

Table 3.2 shows the mean Brier or quadratic score and its reliability, resolution and uncertainty components for the various forecasts (Murphy 1973; Dawid 1986). Suppose that the probability forecasts p_t for the binary event y_t, where $t = 1, \ldots, n$, take discrete values $f_i \in [0,1]$, where $i = 1, \ldots, I$. Let n_i be the number of times that the forecast value f_i occurs, so that $n = n_1 + \cdots + n_I$, and let q_i be the respective empirical conditional event frequency, that is, the ex post recalibrated forecast. Let

$$\bar{q} = \frac{1}{n} \sum_{i=1}^{I} n_i q_i = \frac{1}{n} \sum_{t=1}^{n} y_t$$

denote the marginal event frequency. Then the mean Brier score,

$$\text{BS} = \frac{1}{n} \sum_{t=1}^{n} (p_t - y_t)^2,$$

decomposes as

$$\text{BS} = \underbrace{\frac{1}{n} \sum_{i=1}^{I} n_i (f_i - q_i)^2}_{\text{REL}} - \underbrace{\frac{1}{n} \sum_{i=1}^{I} n_i (q_i - \bar{q})^2}_{\text{RES}} + \underbrace{\bar{q}(1 - \bar{q})}_{\text{UNC}}.$$

The reliability term (REL) quantifies calibration and is negatively oriented, that is, the smaller the better. The resolution component (RES) equals the variance of the ex post recalibrated forecast and is positively oriented. For a calibrated forecast, it quantifies sharpness; for an uncalibrated forecast, it measures potential sharpness. As noted above, we generally seek a forecast which is as sharp as possible subject to it being calibrated (Murphy and Winkler 1987; Gneiting, Balabdaoui and Raftery 2007). The uncertainty term (UNC) is computed from the observations alone and is independent of the forecast. If the probability forecast is a continuous variable, the decomposition depends on a binning of the forecast values and is approximate only. It can be made exact by considering two additional components in the decomposition, as proposed by Stephenson, Coelho and Jolliffe (2008). In our case, the extra terms make very little difference, and we consider the classical decomposition only.

From Table 3.2 we see that the linearly combined ELP and OLP forecasts have lower Brier score than any of the individual forecasts. In both cases, the improvement stems from the resolution component, which is high, because the ex post recalibrated forecast is sharp,

Table 3.2: Out-of-sample mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for the probability forecasts in the simulation example of Sections 3.2.2 and 3.3.2.

Forecast	BS	REL	RES	UNC
p_1	0.2094	0.0002	0.0408	0.2500
p_2	0.1657	0.0004	0.0847	0.2500
ELP	0.1563	0.0418	0.1354	0.2500
OLP	0.1531	0.0120	0.1088	0.2500
BLP	0.1137	0.0005	0.1368	0.2500
CP	0.1126	0.0003	0.1377	0.2500

even though the forecast itself is uncalibrated and lacks sharpness, as reflected in Figure 3.1. The BLP forecast is much better calibrated, and simultaneously more successful in resolving events and non-events, than the ELP and OLP forecasts, resulting in a hugely improved Brier score. As anticipated, the theoretically optimal CP forecast shows the lowest Brier score. However, the BLP forecast is a very close competitor; it is nearly as well calibrated and nearly as sharp as the CP forecast.

3.3.3 Uncalibrated components

In the above simulation experiment, each individual source was calibrated, and we fitted the BLP model (4.7) under the constraint that $\alpha = \beta$. However, the BLP approach is more general, and applies equally in situations in which one or more of the component forecasts are uncalibrated. Furthermore, it can be beneficial to allow for the full BLP model with general parameters $\alpha > 0$ and $\beta > 0$.

In the remainder of the section, we give an example that covers these two types of situations. Specifically, we consider forecast combinations of $p_1 = \Phi(\frac{a_1}{\sqrt{3}})$, defined as previously, and of the **probability forecast $\boldsymbol{p_2^*}$**, which we take to be

$$p_2^* = \Phi(\tfrac{1}{5} + \tfrac{a_2}{2}). \tag{3.11}$$

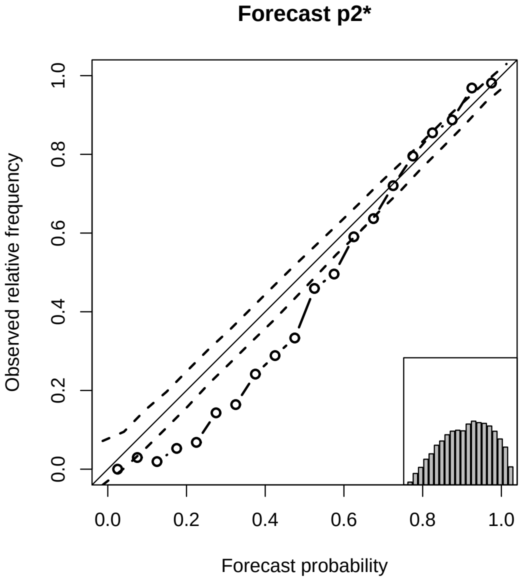

Figure 3.3: Calibration curve and 95% bootstrap intervals under the null hypothesis of calibration for the uncalibrated source p_2^* in the simulation example of Section 3.3.3. The histogram shows the empirical distribution of the forecast values over the unit interval.

As illustrated in Figure 3.3, the source p_2^* is uncalibrated and its marginal distribution is skewed. Table 3.3 shows maximum likelihood estimates for OLP and BLP models which have been fit on the same training sample (that is, using the same random seed for a_1 and a_2) as in Sections 3.2.2 and 3.3.2. Note that we refer to the constrained BLP model (with $\alpha = \beta$) as symmetric, and to the full model (with general parameters $\alpha > 0$ and $\beta > 0$) as asymmetric.

Figure 3.4 and Table 3.4 show performance results for the same independent test sample of size 10,000 as that used before. The calibration curves in Figure 3.4 demonstrate that the symmetric BLP forecast is uncalibrated, while the asymmetric BLP forecast is empirically calibrated. The Brier scores in Table 3.4 confirm that the linearly combined ELP and OLP forecasts outperform each of the individual sources, p_1 and p_2^*. However, the nonlinearly combined BLP forecasts show much better predictive performance. The asymmetric, general version of the BLP forecast outperforms the symmetric version, which is uncalibrated, and

Table 3.3: Maximum likelihood estimates of OLP and BLP parameters in the simulation example of Section 3.3.3, with standard errors in brackets.

Method	w_1	w_2	α	β
OLP	0.265 (0.017)	0.735 (0.017)		
BLP (symmetric)	0.473 (0.005)	0.527 (0.005)	10.11 (0.36)	10.11 (0.36)
BLP (asymmetric)	0.454 (0.005)	0.546 (0.005)	13.72 (0.49)	11.66 (0.42)

Table 3.4: Out-of-sample mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for the probability forecasts in the simulation example of Section 3.3.3.

Forecast	BS	REL	RES	UNC
p_1	0.2094	0.0002	0.0408	0.2500
p_2^*	0.1740	0.0084	0.0844	0.2500
ELP	0.1671	0.0514	0.1343	0.2500
OLP	0.1641	0.0321	0.1179	0.2500
BLP (symmetric)	0.1215	0.0084	0.1369	0.2500
BLP (asymmetric)	0.1132	0.0005	0.1373	0.2500
CP	0.1126	0.0003	0.1377	0.2500

performs nearly as well as the theoretically optimal CP forecast.

3.4 Case Study: Probability of Precipitation Forecasts

We turn to a data example on statistical and National Weather Service probability of precipitation forecasts in the continental US. With some one-third of the US economy being weather sensitive, and severe weather causing billions of dollars in damage and hundreds of deaths annually, there is a critical need for calibrated and sharp probabilistic weather forecasts, to allow for optimal decision making under inherent uncertainty (Dutton 2002; Regnier 2008).

Baars and Mass (2005) consider probability of precipitation forecasts for 29 meteoro-

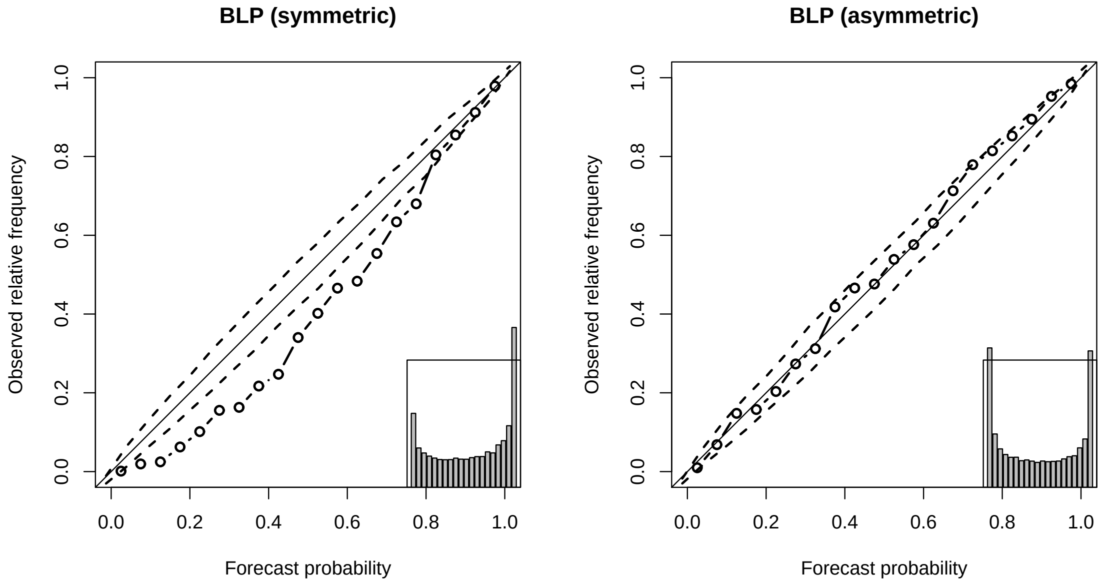

Figure 3.4: Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the BLP forecasts in the simulation example of Section 3.3.3. The histograms show the empirical distribution of the forecast values over the unit interval.

logical stations at major urban centers spread across the continental US. They compare the performance of individual and linearly combined model output statistics (MOS) and National Weather Service (NWS) forecasts, and conclude that a linear opinion pool of the machine generated MOS forecasts is competitive or superior to the NWS forecast at nearly all locations. Here we consider the aggregate performance of individual and combined forecasts at all 29 stations, based on the automated **GMOS**, **EMOS** and **NMOS** forecasts, and the human generated, operational **NWS** forecast. The MOS probability forecasts are statistical forecasts that apply regression techniques to the output of a numerical weather prediction model and recent weather observations (Glahn and Lowry 1972; Wilks 2006). The MOS forecasts are recorded in multiples of a hundredth; the NWS forecasts come in multiples of a tenth, except that a forecast probability of 0.05 is issued occasionally.

We consider 2-days ahead probability of precipitation forecasts for the 12-hour term called period 2 by Baars and Mass (2005), with our data ranging from July 1, 2003 to

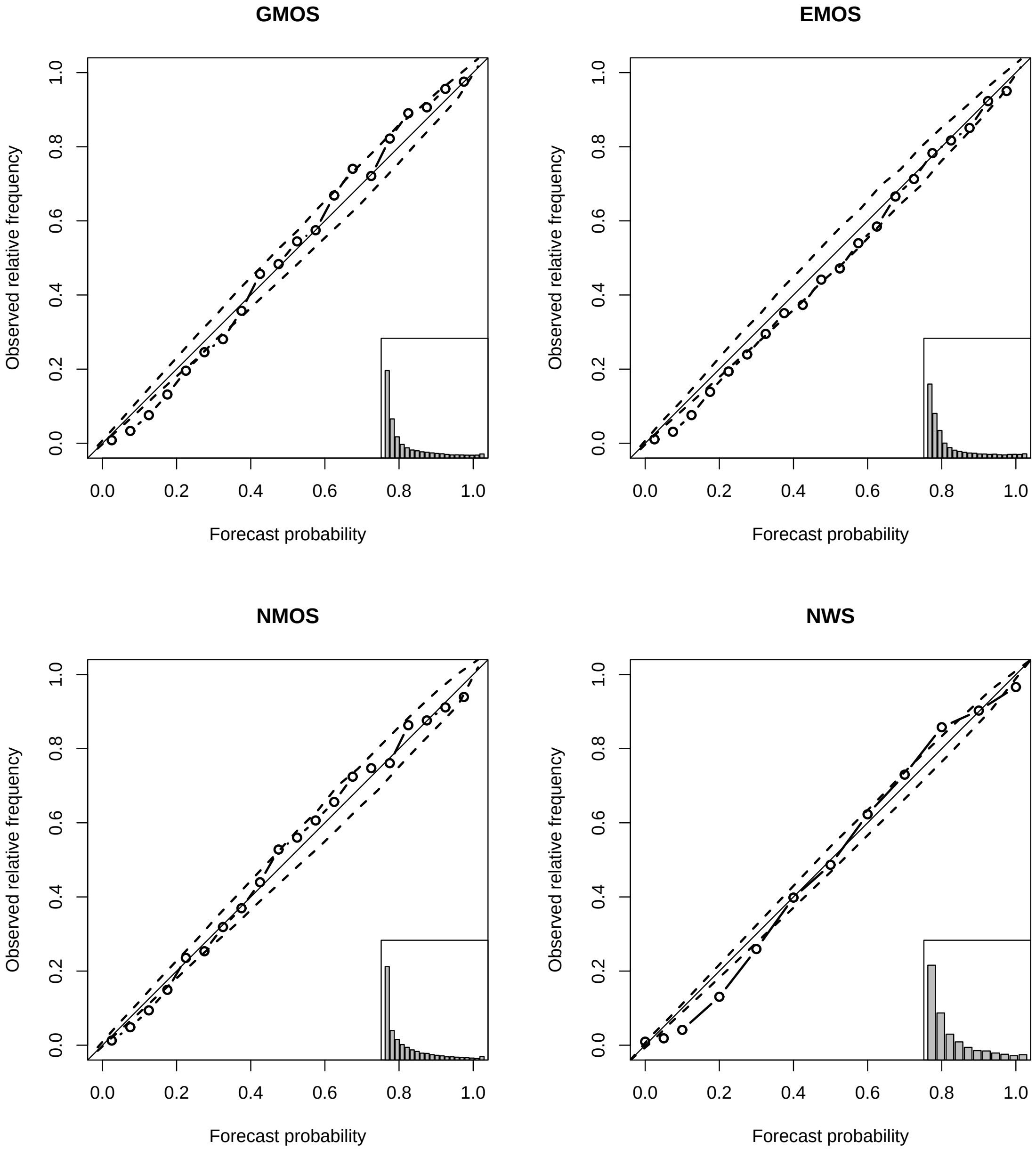

Figure 3.5: Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the four individual probability of precipitation forecasts in the test period. The histograms show the empirical distribution of the forecast values over the unit interval.

March 3, 2008. This includes but is not limited to the one year record studied by Baars and Mass (2005). We use the first two years (July 1, 2003 to June 30, 2005) as training data, on which we fit OLP and BLP models that apply at all stations simultaneously. The balance of the record (July 1, 2005 to March 3, 2008) is used as test data on which we evaluate the forecasts. All results are aggregated over the test period and the 29 stations.

Figure 3.5 shows calibration curves for the four individual forecasts over the test period. We are in the desirable situation in which the calibration curves show only minor deviations from the diagonal, and so we fit the BLP model (3.7) under the constraint (3.9). Hence, the BLP model has a single additional recalibration parameter, $\alpha \geq 1$, when compared to the traditional linear opinion pool.

3.4.1 Combining statistical forecasts

Following Baars and Mass (2005), we consider combined probability forecasts that use the three statistical probability forecasts, namely the **GMOS**, **EMOS** and **NMOS** forecasts. As previously, the **equally weighted linear opinion pool (ELP)** is obtained as the simple average of the three forecasts. Table 3.5 shows maximum likelihood (ML) estimates for the **optimally weighted linear opinion pool (OLP)** and the **beta transformed linear opinion pool (BLP)**, which we fit on the training data. For both methods, the GMOS and EMOS weights are about equal and nearly reach $\frac{1}{2}$, with the NMOS weight being much smaller. This is unsurprising, because NMOS is the oldest system and is well known to be the least accurate of the forecasts considered. The ML estimate of the BLP recalibration parameter, α, is 1.48.

Reliability diagrams for the combined forecasts are shown in Figure 3.6. The calibration curve for the OLP forecast deviates significantly from the diagonal; the effect is stronger than for any of the individual forecasts, and the direction of the departure agrees with our theoretical results, in that the linearly combined forecast is underconfident. The calibration curve for the ELP forecast is very similar and so it is not shown here. The nonlinearly recalibrated BLP forecast is empirically well calibrated and sharper than the OLP forecast.

Table 3.6 shows the Brier score and its reliability, resolution and uncertainty components

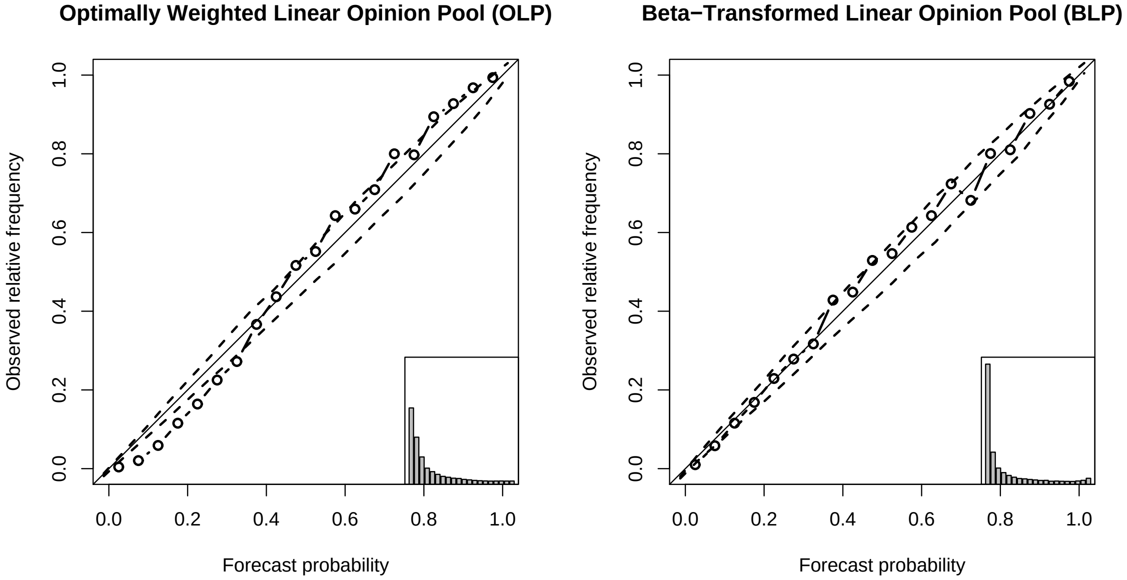

Figure 3.6: Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the OLP and BLP probability of precipitation forecasts in the test period, using the statistical forecasts only. The histograms show the empirical distribution of the forecast values over the unit interval.

Table 3.5: Combined probability forecasts in the precipitation example, using the statistical forecasts only. Maximum likelihood estimates for the OLP and BLP parameters from the training period with standard errors in brackets.

Method	GMOS	EMOS	NMOS	α
OLP	0.485 (0.026)	0.465 (0.027)	0.050 (0.020)	
BLP	0.462 (0.022)	0.447 (0.022)	0.091 (0.021)	1.48 (0.03)

for the individual and combined forecasts. The BLP forecast performs the best, both in terms of the Brier score, the reliability or calibration component, and the resolution component. The improvement of the nonlinear BLP method over the linear OLP forecast is about the same as that of the OLP forecast over the best individual source, the GMOS forecast.

3.4.2 Combining statistical and National Weather Service forecasts

We turn to combined probability forecasts that are based on all four individual sources, now including the NWS forecast, in addition to the GMOS, EMOS and NMOS forecasts. This possibility was not explored by Baars and Mass (2005), who aimed to compare the automated MOS forecasts to the subjective, human generated NWS forecast.

Table 3.7 shows ML estimates for the OLP and BLP models, which we fit on the training data. For both methods, the GMOS and EMOS forecasts receive weights that are nearly equal, at about 0.37. The NWS forecast receives weights of 0.27 and 0.22, respectively; the weights for the NMOS forecast are negligible. The ML estimate of the BLP recalibration parameter, α, is 1.49.

Calibration curves for the OLP and BLP forecasts are shown in Figure 3.7. We see the now familiar pattern, in that the linearly combined OLP forecast lacks calibration. The BLP forecast is empirically well calibrated, and considerably sharper than the OLP forecast. The Brier scores in Table 3.8 echo these results. The BLP forecast outperforms the OLP and ELP forecasts, which perform better than any of the individual forecasts. If we compare to Table 3.6, we see that the combined probability forecasts benefit from the inclusion of the

Table 3.6: Mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for individual and combined probability of precipitation forecasts in the test period, using the statistical forecasts only.

Forecast	BS	REL	RES	UNC
GMOS	0.0815	0.0011	0.0739	0.1543
EMOS	0.0866	0.0011	0.0688	0.1543
NMOS	0.0934	0.0005	0.0614	0.1543
ELP	0.0814	0.0023	0.0751	0.1543
OLP	0.0800	0.0021	0.0764	0.1543
BLP	0.0783	0.0004	0.0764	0.1543

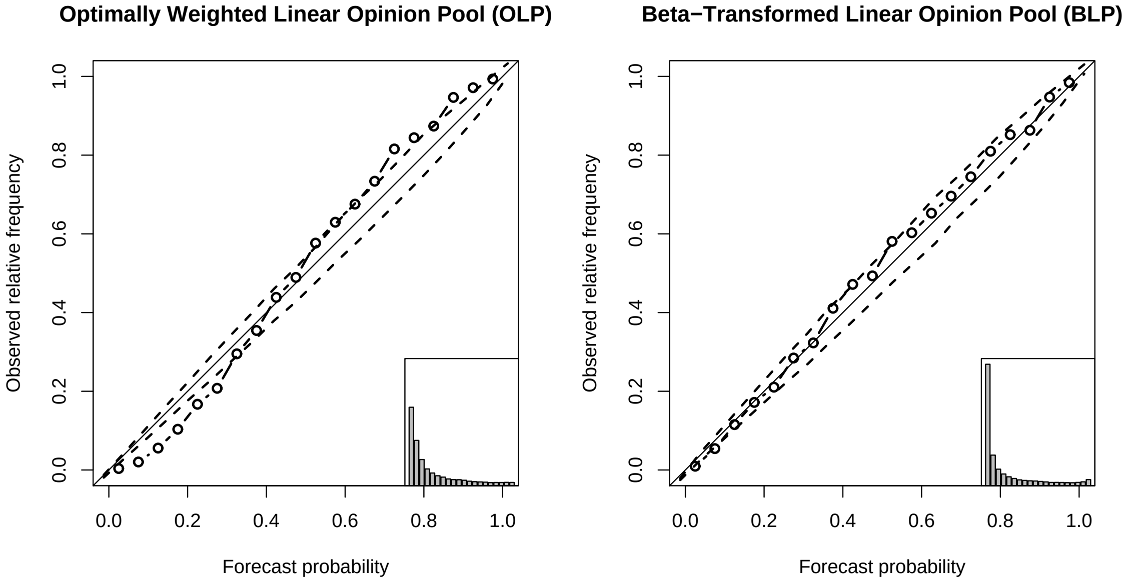

Figure 3.7: Same as Figure 3.6 but now using all four individual forecasts, including the NWS forecast.

human generated NWS forecast, with the improvement due to an increase in resolution.

3.5 Discussion

Our aim in this chapter is to provide theoretical and applied guidance in combining probability forecasts from distinct, calibrated or uncalibrated sources. Historically, the linear opinion pool has been the preferred method for doing this. Indeed, there is overwhelming empirical evidence that linearly combined probability forecasts perform better than individual forecasts, and our results make no exception. That said, the chapter demonstrates theoretically and empirically that the linear opinion pool is suboptimal, lacking both calibration and sharpness. To address these shortcomings, we propose the use of the nonlinearly recalibrated, beta transformed linear opinion pool (BLP) that nests the traditional, linearly combined probability forecast.

Theorem 3.2.1 is our analytic key result; it shows that the linear opinion pool is uncalibrated, even in the desirable case in which the individual sources are calibrated. This is a finite sample result that does not make any restrictive assumptions about the joint dependence structure of the individual forecasts, and complements the asymptotic results of Wallsten and Diederich (2001) that rely on an assumption of conditional independence. It would be of great interest to bridge the finite sample and asymptotic scenarios, and to establish a more general result, roughly to the extent that linearly combined probability forecasts are uncalibrated and underconfident, resulting in probability statements that are closer to the naive climatological forecast than necessary. A result of this type could perhaps be formulated for a general class of averaging operators and under a minimal assumption

Table 3.7: Same as Table 3.5 but now using all four individual forecasts, including the NWS forecast.

Method	GMOS	EMOS	NMOS	NWS	α
OLP	0.362 (0.031)	0.368 (0.030)	0.000 (0.026)	0.270 (0.032)	
BLP	0.371 (0.024)	0.377 (0.023)	0.032 (0.022)	0.220 (0.024)	1.49 (0.03)

of marginal consistency, in lieu of calibration.

Empirically, the shortcomings of the linear opinion pool have been well documented in an interdisciplinary strand of literature that includes the works of Clemen and Winkler (1987), Winkler and Poses (1993), Vislocky and Fritsch (1995), Ariely et al. (2000), Wallsten and Diederich (2001) and Johnson et al. (2001). Despite their ubiquity, these issues have frequently been overlooked, with Sloughter et al. (2007) being one such example. Indeed, Figure 7 of Sloughter et al. (2007) shows the typical S-shaped calibration curve for a linearly combined, underconfident probability forecast, even though the effect is comparably small.

With a view toward applied forecasting problems, we recommend a transition from the traditional linear opinion pool to the nonlinearly recalibrated, beta-transformed linear opinion pool (BLP). The BLP model (3.7) has at most two, and typically only one, additional parameters when compared to the linear opinion pool, and it is easy to fit, using the maximum likelihood method or related optimum score techniques. More general and more complex parametric or nonparametric approaches to the aggregation of probability forecasts can easily be envisioned, including but not limited to copula models (Jouini and Clemen 1996), and might provide effective approximations to the hypothetical, ideally combined forecast, namely the conditional probability (CP) forecast (3.5). However, more complex statistical models bear the danger of overfitting, and the resulting gains in predictive performance, if any, are likely to be incremental.

Appendix: Mathematical Details

Proof of Theorem 3.2.1

From the basic properties of Bernoulli random variables and conditional expectations,

$$\mathbb{P}(Y=1)=\mathbb{E}Y^2=\mathbb{E}Y=\mathbb{E}\mathbb{E}[Y|p]=\mathbb{E}p=\mathbb{E}q,$$

which will be used repeatedly in what follows. We first prove part (a). For a contradiction, suppose that p is calibrated, that is, $p=q$ almost surely. Then we can condition on p to see that

$$\mathbb{E}(Y-p)^2=\mathbb{E}[p(1-p)]. \tag{3.12}$$

We proceed to show that under the conditions of the theorem equality in (3.12) is violated. Toward this end, note that

$$\begin{aligned}
\mathbb{E}(Y-p)^2 &= \mathbb{E}\left(\sum_{i=1}^k w_i(Y-p_i)\right)^2\\
&= \sum_{i=1}^k\sum_{j=1}^k w_iw_j\,\mathbb{E}[(Y-p_i)(Y-p_j)]\\
&= \sum_{i=1}^k\sum_{j=1}^k w_iw_j\,\mathbb{E}[Y-p_iY-p_jY+p_ip_j]\\
&= \sum_{i=1}^k\sum_{j=1}^k w_iw_j\,\mathbb{E}[\mathbb{E}(Y|p_i)-\mathbb{E}(p_iY|p_i)-\mathbb{E}(p_jY|p_j)+p_ip_j]\\
&= \sum_{i=1}^k\sum_{j=1}^k w_iw_j\,\mathbb{E}[p_i-p_i^2-p_j^2+p_ip_j]\\
&= \sum_{i=1}^k\sum_{j=1}^k w_iw_j\,\mathbb{E}[p_i(1-p_j)]-\sum_{i=1}^k\sum_{j=1}^k w_iw_j\,\mathbb{E}(p_i-p_j)^2
\end{aligned}$$

and

$$\mathbb{E}[p(1-p)]=\sum_{i=1}^k\sum_{j=1}^k w_iw_j\,\mathbb{E}[p_i(1-p_j)],$$

so that

$$\mathbb{E}(Y-p)^2 = \mathbb{E}[p(1-p)] - \sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j \mathbb{E}(p_i - p_j)^2. \tag{3.13}$$

The double sum on the right-hand side of (3.13) is strictly positive, whence (3.12) is violated, for the desired contradiction.

We now prove part (b). From (3.13) we see that $\mathbb{E}(Y-p)^2 < \mathbb{E}[p(1-p)]$. A straightforward conditioning argument shows that

$$\mathbb{E}(Y-p)^2 = \mathbb{E}(Y-q)^2 + \mathbb{E}(q-p)^2 > \mathbb{E}[q(1-q)].$$

Hence,

$$\mathbb{E}[q(1-q)] < \mathbb{E}(Y-p)^2 < \mathbb{E}[p(1-p)],$$

which implies that $\mathbb{E}p^2 < \mathbb{E}q^2$. From this, part (b) follows.

As for part (c),

$$\begin{aligned} \mathbb{E}\mathrm{S}(q,Y) &= \mathbb{E}\mathbb{E}[\mathrm{S}(q,Y)|p] \\ &= \mathbb{E}[q\mathrm{S}(q,1) + (1-q)\mathrm{S}(q,0)] \\ &< \mathbb{E}[q\mathrm{S}(p,1) + (1-q)\mathrm{S}(p,0)] \\ &= \mathbb{E}\mathbb{E}[\mathrm{S}(p,Y)|p] \\ &= \mathbb{E}\mathrm{S}(p,Y) \end{aligned}$$

with the inequality being strict, because S is a negatively oriented strictly proper scoring rule and $q = \mathbb{E}[Y|p] \neq p$ with positive probability. □

Details for (3.3) and (3.4)

The final equality in (3.3) stems from the fact that if $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\begin{aligned}
\mathbb{E}\Phi(X) &= \int_{-\infty}^{\infty} \Phi(x)\, \frac{1}{\sigma}\, \phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{x} \phi(y)\, \mathrm{d}y\right) \frac{1}{\sigma}\, \phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x \\
&= \int_{y \le x} \phi(y)\, \frac{1}{\sigma}\, \phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x\, \mathrm{d}y \\
&= \mathbb{P}(Y \le X) \;=\; \mathbb{P}(Y - X \le 0) \;=\; \Phi\left(\frac{\mu}{\sqrt{\sigma^2+1}}\right),
\end{aligned}$$

where ϕ denotes the standard normal density function and Y is standard normal and independent of X, so that $Y - X \sim \mathcal{N}(-\mu, \sigma^2+1)$. The conditional distribution of $a_1 + a_2$ given a_1 is normal with mean a_1 and variance 2, whence

$$\mathbb{E}\left[\Phi(a_1 + a_2) | a_1\right] = \Phi\left(\frac{a_1}{\sqrt{3}}\right).$$

An almost identical calculation applies to (3.4).

Table 3.8: Same as Table 3.6 but now using all four individual forecasts, including the NWS forecast.

Forecast	BS	REL	RES	UNC
GMOS	0.0815	0.0011	0.0739	0.1543
EMOS	0.0866	0.0011	0.0688	0.1543
NMOS	0.0934	0.0005	0.0614	0.1543
NWS	0.0827	0.0009	0.0725	0.1543
ELP	0.0800	0.0026	0.0770	0.1543
OLP	0.0789	0.0024	0.0778	0.1543
BLP	0.0770	0.0004	0.0777	0.1543

Chapter 4

DENSITY FORECAST COMBINATION, CALIBRATION, AND RECALIBRATION

4.1 Introduction

Probabilistic forecasts aim to provide calibrated and sharp predictive distributions for future quantities or events of interest. For a continuous outcome, probabilistic forecasts take the form of a predictive density or density forecast. As they admit the assessment of forecast uncertainty and allow for optimal decision making (Granger and Pesaran 2000; Gneiting 2008b), density forecasts continue to gain prominence in a wealth of applications, ranging from economics and finance to climatology and meteorology (Tay and Wallis 2000; Timmermann 2000; Gneiting 2008a). The general goal is to maximize the sharpness of the density forecasts subject to calibration (Murphy and Winkler 1987; Hora 2004; Gneiting et al. 2007).

In many situations, complementary or competing density forecasts from dependent or independent information sources are available. For example, the individual density forecasts might stem from distinct experts, organizations or statistical models. The most prevalent method for aggregating the individual forecasts into a single, combined density forecast is the linear opinion pool, or simply linear pool (Stone 1961), that is, a weighted linear combination of the individual density forecasts. While other methods for combining density forecasts are available (Genest and Zidek 1986; Clemen and Winkler 1999; 2007), the linear pool is typically the method of choice, with the pioneering work of Winkler (1968) and Zarnowitz (1969) and recent papers by Mitchell and Hall (2005), Wallis (2005), Hall and Mitchell (2007), Jore et al. (2008), Geweke and Amisano (2008), Kascha and Ravazzolo (2008) and Österholm (2009) being examples.

In this context, Hora (2004) demonstrated an interesting and disconcerting result, by proving that any nontrivial linear combination of two calibrated density forecasts is uncalibrated. Here we extend Hora's findings in various directions. In Section 4.2 we prove an analogous result under weaker conditions and for an arbitrary number of component forecasts, and we expose the nature of the forecast deficiency, in that the linearly combined density forecast is overdispersed. To address these shortcomings, we propose two recalibration techniques, namely the deflated linear pool (DLP), which adapts the spread of the component densities, and the beta-transformed linear pool (BLP), which applies a nonlinear recalibration transform to the traditional linear opinion pool. These methods can be used effectively to aggregate calibrated as well as uncalibrated sources. Sections 4.3 and 4.4 turn to case studies on density forecasts for daily maximum temperature at Seattle-Tacoma Airport and S&P 500 returns. The chapter ends in Section 4.5, where we summarize our findings and hint at their relevance in the closely related problem of the fusion of expert opinions that are expressed in terms of probability densities.

4.2 Theory and methods

4.2.1 Calibration and dispersion

In a seminal paper, Murphy and Winkler (1987) proposed a distribution oriented framework for the evaluation of point forecasts. Here, we generalize to density forecasts and consider the joint distribution of a random variable Y, which represents the future quantity of interest, and a finite family $\{f_i : i = 1, \ldots, k\}$ of density-valued random quantities, which stand for the forecasts. We denote the cumulative distribution functions that correspond to the density forecasts by $\{F_i : i = 1, \ldots, k\}$ and require them to be right-continuous. Two density forecasts then are distinct if there is a positive probability of the corresponding cumulative distribution functions being such.

In this setting, the probability integral transform (PIT) for the density forecast f is the random variable

$$Z = F(Y),$$

which takes values in the closed unit interval. Thus, the PIT is the value that the predictive cumulative distribution function attains at the verifying observation. This notion includes the traditional PIT (Rosenblatt 1952), in which the cumulative distribution function F is deterministic, rather than a random quantity, as a special case.

With this, we are ready to define the notions of calibration and dispersion.

Definition 4.2.1. Suppose that the density-valued random quantity f is a density forecast for the random variable Y. Let F denote the corresponding cumulative distribution function.

(a) The density forecast f is *calibrated* if its PIT, $Z = F(Y)$, is uniformly distributed on the unit interval.

(b) The density forecast f is *overdispersed* if its PIT, $Z = F(Y)$, satisfies $\text{var}(Z) < \frac{1}{12}$, *neutrally dispersed* if $\text{var}(Z) = \frac{1}{12}$, and *underdispersed* if $\text{var}(Z) > \frac{1}{12}$.

(c) A density forecast f is *regular* if the distribution of its PIT, $Z = F(Y)$, is supported on the unit interval.

The following result then is immediate.

Proposition 4.2.2. *A calibrated density forecast is neutrally dispersed and regular.*

The converse is not necessarily true, in that a density forecast which is neutrally dispersed and regular need not be calibrated.

Dawid (1984), Diebold et al. (1998) and Gneiting et al. (2007), among others, have argued powerfully that calibration is a critical requirement for a density forecast. Consequently, checks for the uniformity of the PIT have formed a cornerstone of density forecast evaluation. In practice, one observes a sample $\{(f_{1j}, \ldots, f_{kj}, y_j) : j = 1, \ldots, l\}$ from the joint distribution of the density forecasts and the observation, and the uniformity of the PIT is assessed empirically. The prevalent way of doing this is by plotting PIT histograms for the various density forecasting methods, which show the frequency distribution of the corresponding PIT values over an evaluation or test period. U-shaped PIT histograms correspond to underdispersed density forecasts that are too narrow on average, while hump or inverse U-shaped histograms indicate overdispersed forecasts. Formal tests of uniformity can also be employed; for a review, see Corradi and Swanson (2006).

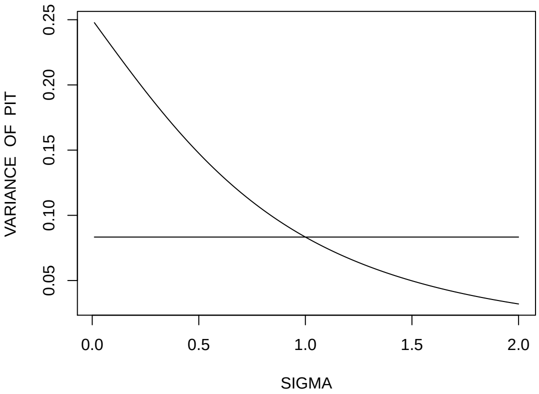

Figure 4.1: The variance (4.2) of the PIT, $Z_\sigma = f_\sigma(Y)$, for the density forecast f_σ in Example 4.2.3 as a function of the predictive standard deviation, σ. The horizontal line is at $\frac{1}{12}$ and indicates a neutrally dispersed forecast.

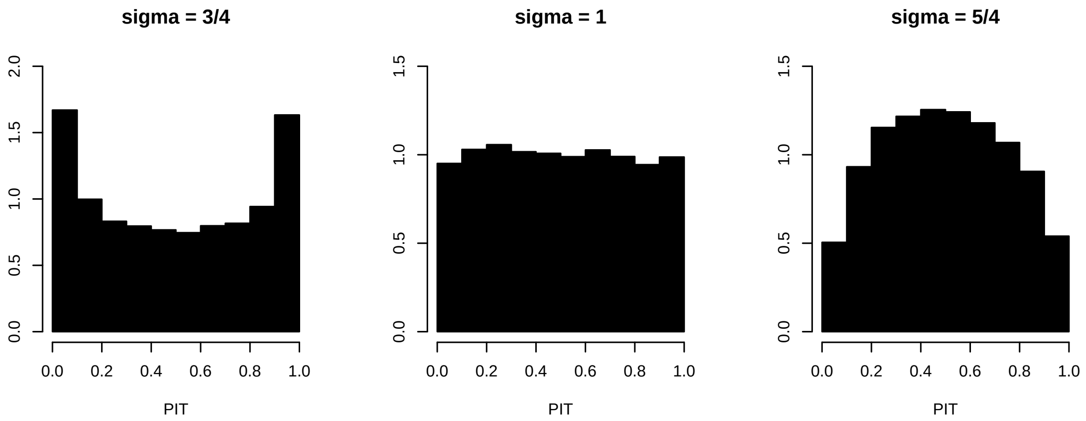

Figure 4.2: PIT histograms for the density forecast f_σ in Example 4.2.3 where $\sigma = \frac{3}{4}$ (underdispersed), $\sigma = 1$ (neutrally dispersed and calibrated) and $\sigma = \frac{5}{4}$ (overdispersed). Details are given in the text.

Example 4.2.3. Let

$$Y = X + \epsilon$$

where X and ϵ are independent, standard normal random variables. Let ϕ denote the standard normal density function. The Gaussian density forecast

$$f_\sigma(y) = \frac{1}{\sigma}\,\phi\!\left(\frac{y - X}{\sigma}\right)$$

has mean X and standard deviation σ, and is a random quantity, because its mean depends on the random variable X. For short, we write

$$f_\sigma \sim \mathcal{N}(X, \sigma^2) \tag{4.1}$$

to indicate that f_σ is a normal density with mean X and variance σ^2. A stochastic domination argument, the details of which we give in Appendix A, shows that f_σ is underdispersed if $\sigma < 1$, neutrally dispersed if $\sigma = 1$ and overdispersed if $\sigma > 1$. If $\sigma = 1$ then f_σ is furthermore calibrated. A more detailed calculation, which is also given in Appendix A, shows that the PIT $Z_\sigma = f_\sigma(Y)$ satisfies

$$\mathrm{var}(Z_\sigma) = 2\int_0^1 z\left(1 - \Phi(\sigma(\Phi^{-1}(z)))\right)\mathrm{d}z - \left(\int_0^1 \left(1 - \Phi(\sigma(\Phi^{-1}(z)))\right)\mathrm{d}z\right)^2, \tag{4.2}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. In Figure 4.1, we plot the variance (4.2) as a function of the predictive standard deviation, σ. Figure 4.2 shows PIT histograms for a sample of size $10,000$ from the joint distribution of the observation Y and the density forecasts f_σ, where $\sigma = \frac{3}{4}$, $\sigma = 1$ and $\sigma = \frac{5}{4}$. The PIT histograms are U-shaped, uniform, and inverse U-shaped, reflecting underdispersion, neutral dispersion and calibration, and overdispersion, respectively.

Similarly, the unconditional density forecast $g_\sigma \sim \mathcal{N}(0, \sigma^2)$ is underdispersed if $\sigma < \sqrt{2}$, neutrally dispersed and calibrated if $\sigma = \sqrt{2}$ and overdispersed if $\sigma > \sqrt{2}$. While both f_1 and $g_{\sqrt{2}}$ are calibrated, the conditional forecast f_1 is much sharper.

The uniform density forecast u_δ on the interval $[-\delta, \delta]$ is underdispersed if $\delta < \sqrt{6}$, neutrally dispersed but not calibrated if $\delta = \sqrt{6}$ and overdispersed if $\delta > \sqrt{6}$. All density forecasts in this example are regular. □

The current setting differs from but relates to the approach of Gneiting et al. (2007), who studied notions of calibration for density forecasts. Specifically, if the density forecast f is calibrated and $\{(f_j, Y_j) : j = 1, 2, \ldots\}$ is a sample from the joint distribution of the density forecast and the observation, then this sequence is probabilistically calibrated in the sense described by Gneiting et al. (2007).

4.2.2 Properties of linearly combined density forecasts

We proceed to state and prove our key result, in that linear combinations of neutrally dispersed density forecasts are uncalibrated and overdispersed.

Theorem 4.2.4. *Let $f_1, \ldots, f_k$ be neutrally dispersed density forecasts, at least two of which are regular and distinct. We consider the linearly combined density forecast $f = \sum_{i=1}^k w_i f_i$ with weights $w_1, \ldots, w_k$ that are strictly positive and add to 1. Then f is overdispersed. In particular, if the density forecasts $f_1, \ldots, f_k$ are calibrated, the linearly combined density forecast f is uncalibrated and overdispersed.*

Proof. For $i = 1, \ldots, k$, let Z_i denote the PIT for the density forecast f_i. By assumption, $\text{var}(Z_i) = \frac{1}{12}$ and $\text{cov}(Z_i, Z_j) \leq \frac{1}{12}$ with strict inequality for at least one pair $i \neq j$, because the density forecasts f_i and f_j are regular and distinct. The PIT for the linear pool $f = \sum_{i=1}^k w_i f_i$ is $Z = \sum_{i=1}^k w_i Z_i$, whence

$$\begin{aligned} \text{var}(Z) &= \sum_{i=1}^k \sum_{j=1}^k w_i w_j \, \text{cov}(Z_i, Z_j) \\ &< \frac{1}{12} \sum_{i=1}^k w_i \sum_{j=1}^k w_j \;\; = \frac{1}{12}. \end{aligned}$$

Thus, the linearly combined density forecast f is overdispersed. The final part of the statement then is immediate from Proposition 4.2.2. □

Despite its elementary and potentially surprising proof, Theorem 4.2.4 generalizes the key result of Hora (2004) in various ways. Hora (2004) applied Fourier analytic tools to show that if the density forecasts f_1 and f_2 are calibrated and distinct, then any linearly combined density forecast is uncalibrated. Our result goes considerably further, by allowing for any number $k \geq 2$ of component densities, by substituting the weaker condition of neutral dispersion for the assumption of calibration, and by exposing the nature of the forecast deficiency, in that the linearly combined density forecast is overdispersed. A similar result in the case of probability forecasts for a binary outcome was proved by Ranjan and Gneiting (2009). This uses a very different mode of calibration, and there is no apparent way of deducing their result from ours, or vice versa.

In practice, the weights $w_1, \ldots, w_k$ in the linear pool are fitted on training data, to satisfy some sort of optimality criterion. Our preferred technique for doing this is the maximum likelihood method (see, for example, Ferguson 1996). Let $\{(f_{1j}, \ldots, f_{kj}, y_j) : j = 1, \ldots, l\}$ denote the training data. Under the assumption of independence between the training cases, the log likelihood function for the linear opinion pool is

$$\ell(w_1, \ldots, w_k) = \sum_{j=1}^{l} \log \left(\sum_{i=1}^{k} w_i f_{ij}(y_j) \right). \tag{4.3}$$

An alternative interpretation of the optimality criterion (4.3), which does not depend on any assumption of independence, is that of the mean logarithmic score (Matheson and Winkler 1976; Gneiting and Raftery 2007) for the training data. The logarithmic score is simply the logarithm, $\log f(x)$, of the value that the density forecast, f, attains at the realizing observation, x. It is positively oriented, that is, the higher the score, the better, and it is proper, in the sense that truth telling is an expectation maximizing strategy. Like all proper scoring rules, the logarithmic score rewards calibrated and sharp predictive distributions.

The optimization in (4.3) is carried out numerically using the method of scoring (see, for example, Ferguson 1996), for which we give details in Appendix B. Approximate standard errors for the estimates can be obtained in the usual way, by evaluating and inverting the Hessian matrix for the log likelihood function. However, the weights $w_1, \ldots, w_k$ need to be nonnegative. Thus, if unconstrained optimization results in negative weights, we turn to the active barrier algorithm implemented in the constrained optimization routine CONSTROPTIM in R (R Development Core Team 2009).

4.2.3 Recalibration

In practice, we employ the linear pool,

$$f(y) = \sum_{i=1}^{k} w_i f_i(y), \tag{4.4}$$

with maximum likelihood estimates for the weights from the training period. Like all linearly combined density forecasts, the resulting **optimal linear pool (OLP)** is subject to the overdispersion described in Theorem 4.2.4 and thus generally is uncalibrated. To address these shortcomings, we propose two recalibration approaches.

The first approach is the **deflated linear pool (DLP)**, which introduces a deflation parameter to reduce the spread of each individual density component. Consider the situation in which each component forecast, $f_i(y) = f_i(y; \mu_i, \sigma_i)$, comes from a location-scale family with location parameter μ_i and scale parameter σ_i, for $i = 1, \ldots, k$. In this setting, the deflated linear pool has density

$$f(y) = \sum_{i=1}^{k} w_i f_i(y; \mu_i, c\sigma_i), \tag{4.5}$$

where $w_1, \ldots, w_k$ are nonnegative weights that add to 1, and c is a strictly positive deflation parameter. If the components are normal we recover the setting in Berrocal et al. (2007) and Glahn et al. (2008).

We use the maximum likelihood method to estimate the weights and the deflation parameter, c, from training data. For calibrated or overdispersed components, we expect estimates for c below 1. If the individual density components are underdispersed, values above 1 might be appropriate. When $c = 1$ the DLP reduces to the standard linear opinion pool. While the DLP model (4.5) can be generalized to allow for distinct deflation parameters for the individual density components, such an extension has not been beneficial in our experience. The assumption of a common deflation parameter yields a more parsimonious model and stabilizes the estimation.

As an alternative, we consider the **beta-transformed linear pool (BLP)**, which composites the traditional linear pool with a beta transform. This method of nonlinearly aggregating density forecasts is best described in terms of cumulative distribution functions. Let $F_1, \ldots, F_k$ denote the cumulative distribution functions for the component densities $f_1, \ldots, f_k$. The BLP cumulative distribution function then is

$$F(y) = B_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_i(y)\right), \tag{4.6}$$

where $w_1, \ldots, w_k$ are nonnegative weights that sum to 1, and $B_{\alpha,\beta}$ denotes the cumulative distribution function of the beta distribution with parameters $\alpha > 0$ and $\beta > 0$. The BLP density forecast then is

$$f(y) = \left(\sum_{i=1}^{k} w_i f_i(y)\right) b_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_i(y)\right), \tag{4.7}$$

where $b_{\alpha,\beta}$ denotes the beta density with parameters $\alpha > 0$ and $\beta > 0$. Like the DLP, the BLP nests the traditional linear opinion pool, which arises as the special case in which $\alpha = \beta = 1$, so that the beta term in (4.7) becomes a constant. For every fixed threshold $y \in \mathbb{R}$ the BLP transform (4.6) acts on a set of probability forecasts for the binary event $\{Y \leq y\}$, which is the setting in which the transform was introduced by Ranjan and Gneiting (2009). Here we consider all threshold values simultaneously and thus transform a cumulative distribution function, rather than just a probability forecast for a dichotomous event.

The weights $w_1, \ldots, w_k$ and the transformation parameters $\alpha > 0$ and $\beta > 0$ are estimated from training data, using the maximum likelihood method. Generalizing (4.3), the log likelihood function for the BLP model (4.7) is

$$
\begin{aligned}
\ell(w_1, \ldots, w_k; \alpha, \beta) &= \sum_{j=1}^{J} \log(f(y_j)) \qquad (4.8) \\
&= \sum_{j=1}^{J} \log\left(\sum_{i=1}^{k} w_i f_{ij}(y_j)\right) + \sum_{j=1}^{J} \log\left(b_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_{ij}(y_j)\right)\right) \\
&= \sum_{j=1}^{J} \left((\alpha - 1) \log\left(\sum_{i=1}^{k} w_i F_{ij}(y_j)\right) + (\beta - 1) \log\left(1 - \sum_{i=1}^{k} w_i F_{ij}(y_j)\right)\right) \\
&\quad + \sum_{j=1}^{J} \log\left(\sum_{i=1}^{k} w_i f_{ij}(y_j)\right) - J \log \mathrm{B}(\alpha, \beta),
\end{aligned}
$$

where B denotes the classical beta function. As before, the optimization is carried out numerically by the method of scoring, for which we give details in Appendix B, or by an active barrier algorithm that honors the linear constraints on the parameters.

4.2.4 Simulation example: Calibrated components

We now present a simulation example, in which three calibrated density forecasts are to be aggregated. The data generating process for the observation, Y, is the regression model

$$Y = X_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + \epsilon, \qquad (4.9)$$

where a_1, a_2 and a_3 are real constants, and X_0, X_1, X_2, X_3 and ϵ are independent, standard normal random variables. The individual forecasts rest on partial knowledge of the data generating process, in that density forecast f_1 has access to covariate X_0 and X_1, but not to X_2 or X_3, and similarly for f_2 and f_3. Thus, we seek to combine the density forecasts

$$f_1 \sim \mathcal{N}(X_0 + a_1 X_1, 1 + a_2^2 + a_3^2),$$

$$f_2 \sim \mathcal{N}(X_0 + a_2 X_2, 1 + a_1^2 + a_3^2) \quad \text{and} \quad f_3 \sim \mathcal{N}(X_0 + a_3 X_3, 1 + a_1^2 + a_2^2),$$

where X_0 stands for shared, public information, while X_1, X_2 and X_3 represent proprietary information sets. The density forecasts represent the true conditional distributions under the regression model (4.9), given the corresponding partial information, and thus they are calibrated. We estimate OLP, DLP and BLP models for combining the component forecasts on a training sample $\{(f_{1j}, f_{2j}, f_{3j}, Y_j) : j = 1, \ldots, l\}$ of size $l = 500$ from the joint distribution of the density forecasts and the observation, and assess the aggregation methods on an independent test sample of the same size. The regression coefficients in the data generating model (4.9) are taken to be $a_1 = a_2 = 1$ and $a_3 = 1.1$.

Table 4.1 shows maximum likelihood estimates, along with approximate standard errors, for the OLP, DLP and BLP parameters. For all three methods, the weight estimate is highest for the component density f_3, which is sharper than f_1 or f_2. The DLP deflation parameter is estimated at 0.78, and the BLP transformation parameters at 1.49 and 1.44, respectively.

The PIT diagrams for the various types of forecasts in the test period are shown in Figure 4.3. The component forecasts f_1, f_2 and f_3 are calibrated and have uniform PIT histograms, up to sample fluctuations. The OLP forecast is overdispersed, as expected. The DLP and BLP forecasts show nearly uniform PIT histograms and thus are empirically calibrated.

Table 4.2 shows the mean logarithmic score for the various forecasts. The best individual forecast is f_3, because it is sharper than f_1 and f_2, while all three forecasts are calibrated. The linearly combined OLP forecast outperforms the individual density forecasts, even though it is overdispersed. The nonlinearly aggregated DLP and BLP forecasts show higher scores than any of the individual or linearly combined density forecasts.

The OLP, DLP and BLP techniques are methods for aggregating density forecasts into a single, combined predictive distribution. Engle et al. (1984, p. 160) make the argument that

> "The best forecast is obtained by combining information sets, not forecasts from information sets. [...] one should combine the information that goes into the models, not the forecasts that come out of the models."

Of course, this is correct, even though it may not be feasible in practice, when individual sources of expertise reveal density forecasts, rather than information sets. In the current simulation setting, however, we can readily combine the forecasters' information sets, re-

Table 4.1: Maximum likelihood estimates with approximate standard errors (in brackets) for the parameters of the combined density forecasts in the simulation example of Section 4.2.4.

	w_1	w_2	w_3	c	α	β
OLP	0.212 (0.083)	0.254 (0.084)	0.534 (0.080)	—	—	—
DLP	0.257 (0.060)	0.283 (0.061)	0.460 (0.059)	0.783 (0.032)	—	—
BLP	0.256 (0.057)	0.293 (0.057)	0.451 (0.054)	—	1.492 (0.062)	1.440 (0.059)

Table 4.2: Mean logarithmic score for the individual and combined density forecasts in the simulation example of Section 4.2.4, for the training set and an independent test set.

	Training	Test
f_1	−2.025	−2.018
f_2	−2.017	−2.022
f_3	−1.956	−1.992
OLP	−1.907	−1.922
DLP	−1.871	−1.892
BLP	−1.865	−1.886

Table 4.3: Maximum likelihood estimates with approximate standard errors (in brackets) for the parameters of the combined density forecasts in the simulation example of Section 4.2.5.

	w_1	w_2	w_3	c	α	β
OLP	0.276 (0.042)	0.294 (0.043)	0.430 (0.043)	—	—	—
DLP	0.254 (0.060)	0.280 (0.060)	0.466 (0.059)	1.380 (0.054)	—	—
BLP	0.257 (0.065)	0.262 (0.067)	0.481 (0.065)	—	0.670 (0.032)	0.643 (0.031)

Table 4.4: Mean logarithmic score for the individual and combined density forecasts in the simulation example of Section 4.2.5, for the training set and an independent test set.

	Training	Test
g_1	−2.598	−2.575
g_2	−2.572	−2.587
g_3	−2.383	−2.490
OLP	−1.951	−1.990
DLP	−1.871	−1.892
BLP	−1.887	−1.914

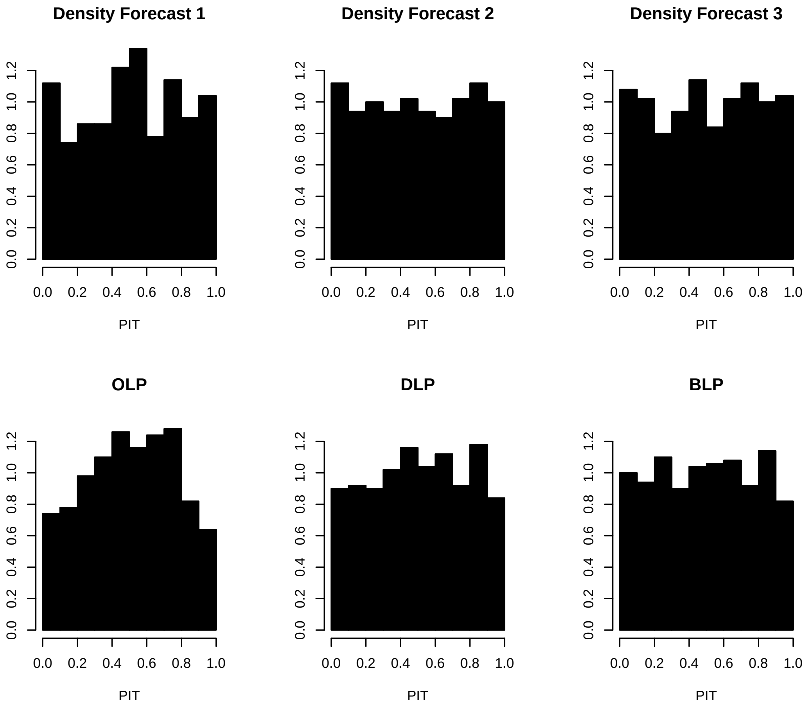

Figure 4.3: PIT histograms for the individual and combined density forecasts in the simulation example of Section 4.2.4, for the test set.

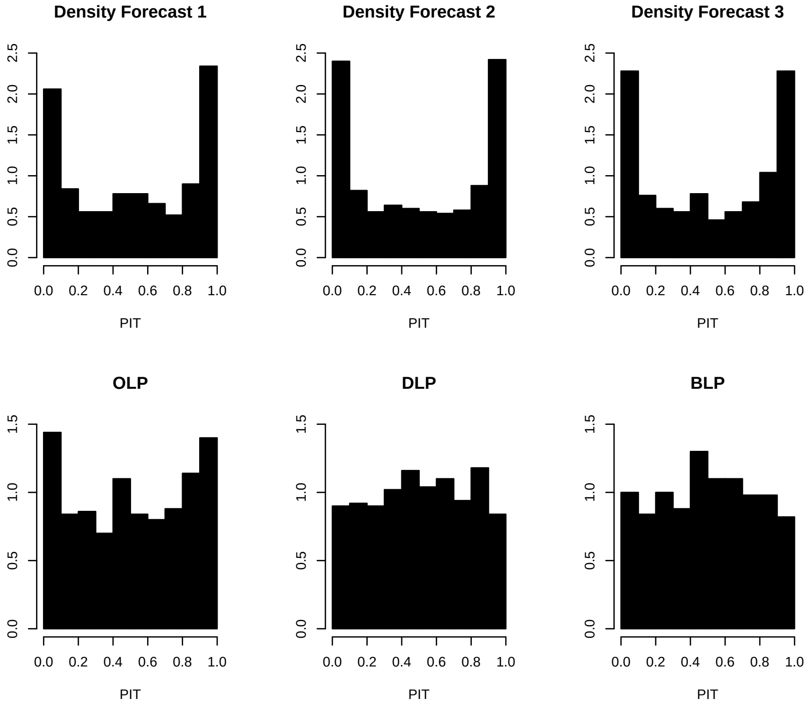

Figure 4.4: PIT histograms for the individual and combined density forecasts in the simulation example of Section 4.2.5, for the test set.

sulting in the density forecast

$$f \sim \mathcal{N}(X_0 + a_1 X_1 + a_2 X_2 + a_3 X_3, 1),$$

which is the true predictive density under the data generating process and complete information about the covariates. This ideal density forecast obtains a mean logarithmic score of -1.432 in the training period and -1.487 in the test period.

4.2.5 Simulation example: Uncalibrated components

In the above simulation example, the individual density forecasts were calibrated. However, the DLP and BLP recalibration techniques apply in the case of uncalibrated components as well. To illustrate this, we retain the setting of the previous section, but consider the density forecasts

$$g_1 \sim \mathcal{N}(X_0 + a_1 X_1, 1), \qquad g_2 \sim \mathcal{N}(X_0 + a_2 X_2, 1), \qquad g_3 \sim \mathcal{N}(X_0 + a_3 X_3, 1),$$

where $a_1 = a_2 = 1$ and $a_3 = 1.1$, as before. These forecasts are underdispersed and thus show U-shaped PIT histograms, as illustrated in the upper row of Figure 4.4. Table 4.3 shows maximum likelihood estimates for the OLP, DLP and BLP parameters. The estimate for the DLP deflation parameter is 1.380, and the estimates for the BLP recalibration parameters α and β are 0.670 and 0.643, in line with the underdispersion of the component forecasts.

The lower row of Figure 4.4 shows PIT histograms for the OLP method, which is still underdispersed, because of the severe underdispersion of the component forecasts. The DLP and BLP methods result in empirically calibrated density forecasts.

The logarithmic score for the various types of forecasts is shown in Table 4.4. The DLP and BLP forecasts provide substantial improvement over the individual density forecasts as well as the linearly combined OLP forecast.

4.3 Density forecasts for daily maximum temperature at Seattle-Tacoma Airport

With some one-third of the US economy being weather sensitive, there is a critical need for calibrated and sharp probabilistic weather forecasts, to allow for optimal decision making under inherent environmental uncertainty (Dutton 2002; Regnier 2008).

In practice, probabilistic weather forecasts rely on ensemble prediction systems. An ensemble system comprises multiple runs of a numerical weather prediction model, with the runs differing in the initial conditions and/or the details of the mathematical representation of the atmosphere (Palmer 2002; Gneiting and Raftery 2005). Here we consider two-days ahead forecasts of daily maximum temperature at Seattle-Tacoma Airport, based on the University of Washington Mesoscale Ensemble (UWME; Eckel and Mass 2005), which employs a regional numerical weather prediction model over the Pacific Northwest, with initial and lateral boundary conditions supplied by eight distinct weather centers. A brief description of the ensemble members is given in Table 4.5.

Our training period ranges from January 1, 2006 to August 12, 2007, with a few days missing in the data record, for a total of 500 training cases. The test period extends from August 13, 2007 to June 30, 2009, for a total of 559 cases. We first use the maximum likelihood method on the training data to estimate, for each ensemble member $i = 1, \ldots, 8$ individually, a Gaussian predictive density of the form

$$f_i \sim \mathcal{N}(a_i + b_i x_i, \sigma_i^2), \tag{4.10}$$

where x_i is the point forecast from the ith ensemble member, a_i and b_i are member specific linear bias correction parameters, and σ_i is the member specific predictive standard deviation. From Table 4.6 we see that the estimates for $\sigma_1, \ldots, \sigma_8$ range from 1.958 to 2.214.

Next we combine the individual density forecasts. Table 4.7 shows maximum likelihood estimates for the OLP, DLP and BLP parameters on the training period. For all three methods, the GFS member, f_1, obtains the highest weight and the ETA member, f_3, the lowest weight. This can readily be explained, in that both members have a common institutional origin, and thus are highly correlated, whence the more competitive GFS member

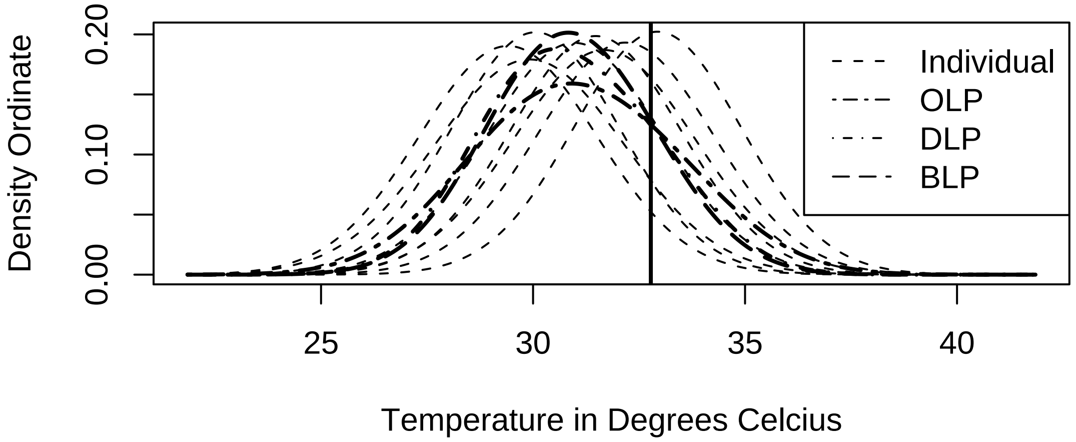

Figure 4.5: Two-day ahead density forecasts for the maximum temperature at Seattle-Tacoma Airport on June 28, 2008. The vertical line is at the verifying maximum, at 32.8 degrees Celsius or 91 degrees Fahrenheit.

acquires the weight of the ETA member as well. The DLP deflation parameter is estimated at 0.768, and the BLP transformation parameters at 1.467, in line with the overdispersion of the linear pool.

Figure 4.5 illustrates the various density forecasts for June 28, 2008, an unusually hot day at Seattle-Tacoma Airport with a verifying maximum temperature of 32.8 degrees Celsius or 91 degrees Fahrenheit. The member specific individual density forecasts are shown by the dotted lines, and the OLP forecast by the dash-dotted line. The nonlinearly aggregated DLP and BLP densities, which are shown by the solid and dashed lines, are sharper than the OLP density.

PIT histograms for the test period are shown in Figure 4.6. The individual, member specific density forecasts $f_1, \ldots, f_8$ are empirically calibrated, showing nearly uniform PIT histograms. The linearly aggregated OLP forecast is overdispersed, as reflected by an inverse U-shaped and skewed PIT histogram. The DLP and BLP forecasts show somewhat rough and skewed, yet more nearly uniform PIT histograms. These results are corroborated by

Table 4.5: Composition of the eight-member University of Washington Mesoscale Ensemble (UWME; Eckel and Mass 2005), with member acronyms and organizational sources for initial and lateral boundary conditions. The United States National Centers for Environmental Prediction supply two distinct sets of initial and lateral boundary conditions, namely, from its Global Forecast System (GFS) and Limited-Area Mesoscale Model (ETA).

Index	Acronym	Source of Initial and Lateral Boundary Conditions
1	GFS	National Centers for Environmental Prediction
2	CMCG	Canadian Meteorological Centre
3	ETA	National Centers for Environmental Prediction
4	GASP	Australian Bureau of Meteorology
5	JMA	Japanese Meteorological Agency
6	NGPS	Fleet Numerical Meteorology and Oceanography Center
7	TCWB	Taiwan Central Weather Bureau
8	UKMO	United Kingdom Met Office

Table 4.6: Maximum likelihood estimates for the predictive standard deviation, σ_i, for the individual, member specific density forecasts in the temperature example.

σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
1.966	2.051	2.119	2.214	1.958	2.055	2.084	1.995

Table 4.7: Maximum likelihood estimates for the parameters of the combined density forecasts in the temperature example.

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	c	α	β	σ
OLP	0.394	0.005	0.000	0.000	0.317	0.030	0.144	0.109	—	—	—	—
DLP	0.304	0.080	0.000	0.085	0.216	0.051	0.172	0.090	0.768	—	—	—
BLP	0.295	0.079	0.000	0.083	0.230	0.062	0.173	0.076	—	1.467	1.467	—
BMA	0.305	0.075	0.000	0.081	0.216	0.056	0.170	0.098	—	—	—	1.566

Table 4.8: Mean logarithmic score for the individual and combined density forecasts in the temperature example, for the training period and the test period.

	Training	Test
f_1	−2.091	−2.088
f_2	−2.134	−2.071
f_3	−2.167	−2.093
f_4	−2.211	−2.172
f_5	−2.088	−2.043
f_6	−2.136	−2.143
f_7	−2.150	−2.131
f_8	−2.107	−2.041
OLP	−2.027	−2.010
DLP	−1.990	−1.961
BLP	−1.988	−1.960
BMA	−1.992	−1.963

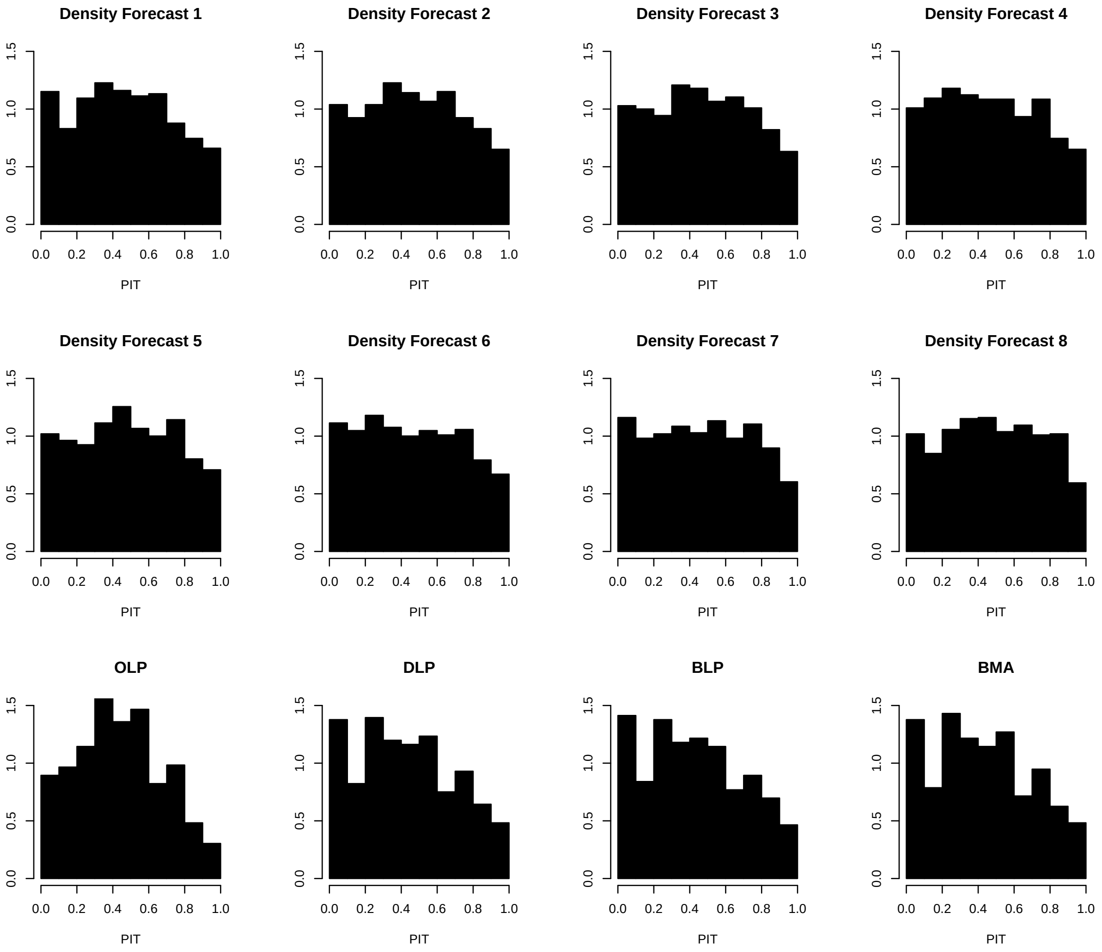

Figure 4.6: PIT histograms for the individual and combined density forecasts in the temperature example, for the test period.

Table 4.8, which shows the mean logarithmic score for the various density forecasts, both for the training period and the test period. The linearly combined OLP forecast shows a higher score than any of the individual density forecasts, which attests to the benefits of forecast aggregation. Nevertheless, the linear pool is suboptimal, because it is overdispersed, and thus is outperformed by the nonlinearly aggregated DLP and BLP density forecasts.

Finally, we compare to the Bayesian model averaging (BMA; Raftery et al. 2005) technique, which is a state of the art approach to generating density forecasts from forecast ensembles. The BMA density forecast is of the form

$$f \sim \sum_{i=1}^{8} w_i \, \mathcal{N}(a_i + b_i x_i, \, \sigma^2), \tag{4.11}$$

with BMA weights, $w_1, \ldots, w_8$, that are nonnegative and sum to 1, member specific bias parameters a_i and b_i for $i = 1, \ldots, 8$, and a common variance parameter, σ^2. In view of our individual density forecasts $f_1, \ldots, f_8$ being Gaussian, the OLP density and the BMA density are of the same functional form. However, there is a conceptual difference, in that the OLP weights are fitted conditionally on the individual density forecasts. Thus, a two-stage procedure is used, in which the member specific component densities are estimated first, and only then the weights, with the component forecasts held fixed. In contrast, the BMA method estimates the weights, $w_1, \ldots, w_8$, and the common spread parameter, σ, for the component forecasts in the Gaussian mixture model (4.11) simultaneously. While the BMA method can be employed with member specific spread parameters, the assumption of a common spread parameter stabilizes the estimation algorithm and does not appreciably deteriorate the predictive performance (Raftery et al. 2005).

Table 4.7 shows maximum likelihood estimates for the BMA parameters, obtained with the R package ENSEMBLEBMA (Fraley et al. 2009). The BMA weights echo the DLP weights. The BMA spread parameter σ is estimated at 1.566 and differs from the predictive standard deviations for the member specific density forecasts by factors ranging from 0.707 to 0.800, much in line with our estimate of 0.768 for the DLP deflation parameter, c. Thus, the DLP and BMA density forecasts are very much alike, which is confirmed by the PIT histograms in Figure 4.6 and logarithmic scores in Table 4.8. In Figure 4.5 the graphs for the DLP and BMA density forecasts are nearly identical and lie essentially on top of each other, and so we refrain from plotting the BMA density.

4.4 Density forecasts for S&P 500 returns

In this second data example, we follow Diebold et al. (1998) in considering S&P 500 log returns for the period of July 3, 1962 to December 29, 1995. The data record through December 1978 is used as training period, for a total of 4,133 training cases. All estimates reported are maximum likelihood fits on the training period obtained with the R package fGarch (Wuertz and Rmetrics Core Team 2007). The balance of the record is used as test period, for a total of 4,298 one-day ahead density forecasts.

The first component forecast, f_1, is based on a generalized autoregressive conditional heteroscedasticity (GARCH; Bollerslev 1986) specification for the variance structure. With r_t denoting the log return on day t, our GARCH(1,1) model assumes that $r_t = \sigma_t \epsilon_t$, where ϵ_t is Student-t distributed with ν degrees of freedom, while σ_t evolves dynamically as

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2.$$

The maximum likelihood estimates for the GARCH parameters are $\omega = 0.000$, $\alpha = 0.089$, $\beta = 0.903$ and $\nu = 9.25$.

The second component forecast, f_2, is based on a standard moving average (MA) model for the mean dynamics, which assumes that $r_t = Z_t + \theta Z_{t-1}$, where $\{Z_t\}$ is a Gaussian white noise process with mean zero and variance σ^2. The maximum likelihood estimates for the MA parameters are $\theta = 0.252$ and $\sigma = 0.00736$.

Our goal now is to combine the density forecasts f_1 and f_2. Table 4.9 shows maximum likelihood estimates for the OLP, DLP and BLP parameters. For all three methods, the conditionally heteroscedastic forecast f_1 obtains a much higher weight than the simplistic density forecast f_2. The DLP deflation parameter is estimated at 0.940, and the BLP recalibration parameters α and β at 1.100 and 1.081. This suggests that the overdispersion of the OLP forecast is quite mild, which is confirmed by the corresponding PIT histogram in Figure 4.7. Table 4.10 shows the mean logarithmic score for the various types of forecasts.

Table 4.9: Maximum likelihood estimates of the parameters for the combined density forecasts in the S&P 500 example.

	w_1	w_2	c	α	β
OLP	0.821	0.179	—	—	—
DLP	0.756	0.244	0.940	—	—
BLP	0.758	0.242	—	1.100	1.081

Table 4.10: Mean logarithmic score for the individual and combined density forecasts in the S&P 500 example, for the training period and the test period.

	Training	Test
f_1	3.606	3.458
f_2	3.492	3.247
OLP	3.612	3.469
DLP	3.614	3.470
BLP	3.614	3.470

The OLP forecast performs slightly better than the component forecast f_1, with a score that is very slightly lower than for the nonlinearly aggregated DLP and BLP density forecasts, both for the training and the test period.

Finally, we consider the predictive performance of a more comprehensive model, which addresses both the first and the second order dynamics, in that $r_t = \mu_t + \epsilon_t$ where $\{\mu_t\}$ and $\{\epsilon_t\}$ are MA(1) and t-GARCH(1,1) processes, respectively. The maximum likelihood estimates in this mixed specification are $\theta = 0.269$ and $\sigma = 0.00736$ for the MA parameters, and $\omega = 0.000$, $\alpha = 0.098$, $\beta = 0.892$ and $\nu = 8.284$ for the GARCH parameters. The resulting forecast can be thought of as combining information sets with respect to the first and second order dynamics, as opposed to combining the corresponding component forecasts f_1 and f_2. It outperforms the other types of density forecasts and achieves a mean logarithmic score of 3.638 for the training period and 3.473 for the test period.

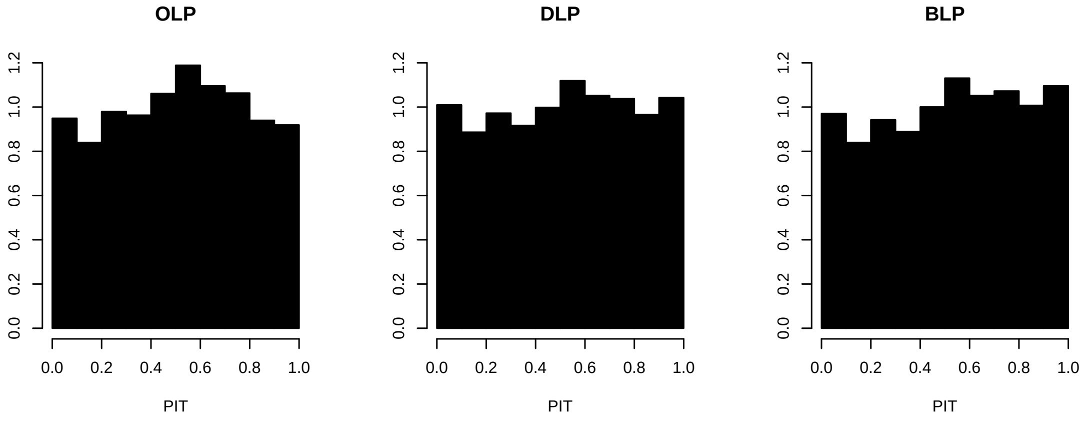

Figure 4.7: PIT histograms for the combined density forecasts in the S&P 500 example, for the test period.

4.5 Discussion

We have demonstrated theoretically and in simulation and data studies that linear combinations of calibrated density forecasts are uncalibrated. Our key result, Theorem 4.2.4, generalizes the extant finding of Hora (2004), which applied to two forecasts, to the case of multiple density forecasts, and identifies the direction of the departure, in that the linear pool is overdispersed. Thus, linearly combined density forecasts tend to show hump or inverse U-shaped PIT histograms. Madigan and Raftery(2004) show that a weighted linear combination of distinct models tend to out perform individual models or density forecasts in terms of log score. Our results are no exception to that. However, we show that while density forecast combination by linear aggregation improves on the individual forecast densities it is in itself suboptimal and one might be able to do better by combining using non-linear techniques. This is also similar to our findings in chapter 2 where linear pooling was shown to be suboptimal for combining probability forecasts.

Also, it is interesting to note that the linear combination which is a uncalibrated forecast improves upon individual calibrated forecasts in terms of log score. This can be explained by the fact that the log score being a proper score, is composed of two components namely, calibration and sharpness. So, although the combined forecast has a smaller calibration component it can have a large sharpness component thus enabling it to have a larger log score as compared to individual forecasts.

We have proposed two nonlinear aggregation methods for density forecasts, namely the deflated linear pool (DLP) and the beta transformed linear pool (BLP). Both methods nest the traditional linear opinion pool and can be used effectively to combine calibrated as well as uncalibrated sources. In our simulation and data studies, the nonlinear methods gave empirically calibrated density forecasts, and outperformed the individual as well as the linearly combined density forecasts, to varying degrees. The parsimonious DLP technique, which generalizes the linear pool by allowing for a single deflation parameter that modifies the spread of the component densities, shows good predictive performance, in line with the well-established stylized fact that simple forecasting methods outperform overly sophisticated ones. The BLP method operates on the corresponding cumulative distribution functions, rather than the density forecasts themselves, and thus applies in the general case of probabilistic forecasts for ordered real-valued outcomes, including discrete, continuous and mixed discrete-continuous variables. Alternative approaches to density forecast aggregation that remain to be explored in practice include consensus models (Winkler 1981) and copula-based approaches (Jouini and Clemen 1996).

While forecast combination is beneficial, it is often preferable to aggregate information sets and derive a density forecast from the combined information basis. In our simulation setting, the benefits of this latter approach were huge; in the data examples, the out-of-sample improvement in the predictive performance was small. Of course, if the density forecasts are supplied by external expert sources, there might be no practical way of combining information sets, and one depends on forecast aggregation. Future work is called for to provide additional theoretically principled as well as applied guidance in doing this.

In addition to the ramifications in density forecasting, our findings bear on the related problem of the fusion of expert opinions that are expressed in the form of probability densities. Ha-Duong (2008) reviews methods for doing this, and applies them to combine expert opinions about the climate sensitivity constant, which is a key quantity in the study of the greenhouse effect. Our results suggest that if each individual expert is calibrated, linear aggregation methods result in combined assessments that are underconfident and show an unduly wide range of uncertainty, when in fact a sharper assessment could be made.

Appendix A: Details for Example 4.2.3

Let $Z_\sigma = f_\sigma(Y)$ denote the PIT for the density forecast f_σ. Then Z_σ has expectation $\frac{1}{2}$ and its cumulative distribution function is $F_\sigma(z) = \Phi(\sigma\,\Phi^{-1}(z))$. In particular, Z_1 has a uniform distribution. If $\sigma < 1$ then $|Z_\sigma - \frac{1}{2}|$ is stochastically larger than $|Z_1 - \frac{1}{2}|$ and therefore

$$\mathrm{var}(Z_\sigma) = \mathbb{E}(Z_\sigma - \mathbb{E}[Z_\sigma])^2 = \mathbb{E}|Z_\sigma - \tfrac{1}{2}|^2 > \mathbb{E}|Z_1 - \tfrac{1}{2}|^2 = \frac{1}{12}.$$

An analogous argument applies when $\sigma > 1$. To prove the variance formula (4.2), we use the fact that $\mathrm{var}(Z_\sigma) = \mathbb{E}[Z_\sigma^2] - (\mathbb{E}[Z_\sigma])^2$ and invoke the well-known expectation equality $\mathbb{E}[Z^r] = r\int_0^\infty z^{r-1}(1 - F(z))\,\mathrm{d}z$ for a nonnegative random variable Z with cumulative distribution function F, where $r > 0$.

Appendix B: Method of scoring

We give details for the method of scoring (see, for example, Ferguson 1996) for numerically maximizing the log likelihood function, $\ell = \ell_{\mathrm{BLP}}$, of the BLP model, as a function of the weights $w_1, \ldots, w_k$ and transformation parameters α and β. The OLP model arises in the special case in which $\alpha = \beta = 1$. Let Y denote a random variable that has a beta distribution with parameters α and β. Then

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha} &= \sum_{j=1}^{J} \log\left(\sum_{i=1}^{k} w_i F_{ij}(y_j)\right) - J\,\mathbb{E}[\log Y], \\
\frac{\partial \ell}{\partial \beta} &= \sum_{j=1}^{J} \log\left(1 - \sum_{i=1}^{k} w_i F_{ij}(y_j)\right) - J\,\mathbb{E}[\log(1 - Y)]
\end{aligned}$$

and

$$\frac{\partial \ell}{\partial w_i} = \sum_{j=1}^{J} \left(\frac{(\alpha - 1)(F_{ij}(y_j) - F_{kj}(y_j))}{\sum_{l=1}^{k} w_l F_{lj}(y_j)} - \frac{(\beta - 1)(F_{ij}(y_j) - F_{kj}(y_j))}{1 - \sum_{l=1}^{k} w_l F_{lj}(y_j)} + \frac{f_{ij}(y_j) - f_{kj}(y_j)}{\sum_{l=1}^{k} w_l f_{lj}(y_j)} \right)$$

for $i = 1, \ldots, k-1$. The second derivatives are

$$\frac{\partial^2 \ell}{\partial \alpha^2} = -\,J\,\mathrm{var}(\log(Y)), \quad \frac{\partial^2 \ell}{\partial \beta^2} = -\,J\,\mathrm{var}(\log(1-Y)), \quad \frac{\partial^2 \ell}{\partial \alpha\,\partial \beta} = -\,J\,\mathrm{cov}(\log(Y), \log(1-Y))$$

and

$$\frac{\partial^2 \ell}{\partial \alpha\,\partial w_i} = \sum_{j=1}^{J} \frac{F_{ij}(y_j) - F_{kj}(y_j)}{\sum_{l=1}^{k} w_l F_{lj}(y_j)}, \qquad \frac{\partial^2 \ell}{\partial \beta\,\partial w_i} = \sum_{j=1}^{J} \frac{F_{kj}(y_j) - F_{ij}(y_j)}{1 - \sum_{l=1}^{k} w_l F_{lj}(y_j)}$$

for $i = 1, \ldots, k-1$, while

$$\begin{aligned} \frac{\partial^2 \ell}{\partial w_{i_1} \partial w_{i_2}} &= -\sum_{j=1}^{J} \frac{(f_{i_1 j}(y_j) - f_{kj}(y_j))(f_{i_2 j}(y_j) - f_{kj}(y_j))}{(\sum_{l=1}^{k} w_l f_{lj}(y_j))^2} \\ -\sum_{j=1}^{J} &\left(\frac{\alpha - 1}{(\sum_{l=1}^{k} w_l F_{lj}(y_j))^2} + \frac{\beta - 1}{(1 - \sum_{l=1}^{k} w_l F_{lj}(y_j))^2} \right) (F_{i_1 j}(y_j) - F_{kj}(y_j))\,(F_{i_2 j}(y_j) - F_{kj}(y_j)) \end{aligned}$$

for $i_1 = 1, \ldots, k-1$ and $i_2 = 1, \ldots, k-1$. The method of scoring now applies Newton's algorithm to optimize the likelihood as a function of the parameter vector.

In the special case of the OLP, $\alpha = \beta = 1$ are fixed and the above expressions reduce to

$$\frac{\partial \ell}{\partial w_i} = \sum_{j=1}^{J} \frac{f_{ij}(y_j) - f_{kj}(y_j)}{\sum_{l=1}^{k} w_l f_{lj}(y_j)}$$

for $i = 1, \ldots, k-1$, while

$$\frac{\partial^2 \ell}{\partial w_{i_1} \partial w_{i_2}} = -\sum_{j=1}^{J} \frac{(f_{i_1 j}(y_j) - f_{kj}(y_j))(f_{i_2 j}(y_j) - f_{kj}(y_j))}{(\sum_{l=1}^{k} w_l f_{lj}(y_j))^2}$$

for $i_1 = 1, \ldots, k-1$ and $i_2 = 1, \ldots, k-1$.

Chapter 5

CONCLUDING REMARKS

We have taken up the problem of combining and evaluating probabilistic forecasts. In Chapter 2, we considered the use of weighted scoring rules to evaluate density forecasts in different regions of interest. The weighted logarithmic scoring rule suggested by Amisano and Giacomini (2007) is not proper and hence can be hedged. To remedy the situation, weighted versions of continuous ranked probability scores (CRPS) are proposed. These scoring rules while emphasizing different regions of density also retain propriety. Threshold and quantile based decompositions of CRPS provide a diagnostic tool to assess the performance of the densities in different regions. We applied our methods on simulation examples and case studies on Bank of England inflation forecasts and wind speed forecasts in Pacific Northwest. A closely related work is Diks, Panchenko and van Dijk (2008). This paper also points out the impropriety of weighted logarithmic score. The authors propose two versions of weighted scoring rules, namely the conditional likelihood (CL) and censored likelihood (CSL) scoring rules. These scoring rules equal the logarithmic scoring rule on a collapsed sample space and hence are proper but not strictly proper.

In Chapter 3, the question of combining forecast probabilities for binary events was taken up. The traditional method of combining probability forecasts by taking a weighted linear combination was shown to be deficient in two important ways namely, calibration and sharpness, even when the individual forecasts are calibrated. The result is quite general and doesn‘t make any restrictive assumptions about the joint structure of observations and forecasts. Wallsten and Diederich (2001) establish a related result in the asymptotic scenario when the number of forecasters goes to infinity and the individual forecasters are conditionally independent given the observation. Their result showed that in the limit the conditional event probability given the weighted average will converge to 0 or 1 according to the average being below or above 0.5.

In order to come up with a calibrated as well as sharp combined forecast, the use of beta transformed linear opinion pool (BLP) was proposed. This combination procedure takes a weighted combination and applies a beta cumulative distribution function to it. The weights as well as the parameters of the beta transform are estimated simultaneously on the training data by maximizing the log score. The method was applied to simulation examples as well as a case study on the probability of precipitation forecasts, with good results.

Chapter 4 considers the problem of combining density forecasts. The use of linearly combined density forecasts has been recommended among others by Winkler (1968). While linearly combined density forecasts can improve upon the individual density forecasts, they are necessarily uncalibrated and hence suboptimal. This is a generalization of Hora (2004) who proves this result for the case of two forecasters. We also show that the linearly combined forecast is overdispersed. Thus, the linearly combined density forecast requires recalibration. We accomplish this by two different approaches by the use of deflated linear opinion pool (DLP) or by beta transformed linear opinion pool (BLP). The DLP method first deflates the scale of individual densities and then takes a weighted average whereas the BLP applies a beta cdf transform to the linearly combined cumulative distribution function. The parameters are obtained jointly by maximizing the log score over training data. The method has been illustrated on simulation examples and a case study on temperature forecasting at Sea-Tac Airport with good results.

As another way to combine density forecasts, Winkler (1981) suggests a normal consensus model. In this approach, the combined distribution is a normal distribution. The mean of the distribution equals a weighted average of the mean of individual density forecasts and variance is a function of the covariance between the forecasts of the individual forecasters. Although this method will work well when the true forecast distribution is normal, it will not be effective if the true distribution is non-normal. The BLP and the DLP approaches will work even when the true distribution is non-normal.

The present work points towards two general directions of further research. We have separately shown the lack of calibration and under confidence (or overdispersion) of linear pool for the case of density forecasts and binary forecasts. It would be of interest to use general notions of calibration and to establish the above result for any predictive distribution function (discrete, continuous, or mixed) and thus bridge a major theoretical gap. One could perhaps generalize these results to averaging operators other than linear pool, like the harmonic or the geometric mean. Another direction of future research is the use of aggregation techniques like copulas, to combine the different forecasts. This can potentially improve upon the combination procedures we have used.

Bibliography

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

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H. B., Wallsten, T. S. and Zauberman, G. (2000), The effects of averaging subjective probability estimates between and within judges, *Journal of Experimental Psychology: Applied*, **6**, 130–147.

Baars, J. A. and Mass, C. F. (2005), Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics, *Weather and Forecasting*, **20**, 1034–1047.

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

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2007), Combining spatial statistical and ensemble information for probabilistic weather forecasting, *Monthly Weather Review*, **135**, 1386–1402.

Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, **78**, 1–3.

Bröcker, J. and Smith, L. A. (2007), Increasing the reliability of reliability diagrams, *Weather and Forecasting*, **22**, 651–661.

Clemen, R. T. and Winkler, R. L. (1987), Calibrating and combining precipitation probability forecasts, in Viertl, R. (ed.), *Probability and Bayesian Statistics*, Plenum, New York, pp. 97–110.

Clemen, R. T. and Winkler, R. L. (2007), Aggregating probability distributions, in Ward, E., Miles, R. F. and von Winterfeldt, D. (eds.), *Advances in Decision Analysis: From Foundations to Applications*, Cambridge University Press, pp. 154–176.

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

Clemens, M. P. (2006), Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts. *Empirical Economics*, **31**, 49–64.

Croushore, D. (1993), Introducing: the Survey of Professional Forecasters, *Federal Reserve Bank of Philadelphia Business Review*, November/December 3–13.

Corradi, V., and Swanson, N. R. (2006a), Predictive density and conditional confidence interval accuracy tests, *Journal of Econometrics*, **135**, 187–228.

Corradi, V., and Swanson, N. R. (2006b), Predictive density evaluation, in *Handbook of Economic Forecasting*, ed. by C. W. J. Granger, G. Elliott and A. Timmermann, Amsterdam, North Holland, 197–286.

Dawid, A. P. (1984), Statistical theory: the prequential approach, *Journal of the Royal Statistical Society Series A*, **147**, 278–292.

Dawid, A. P. (1982), The well-calibrated Bayesian, *Journal of the American Statistical Association*, **77**, 605–610.

Dawid, A. P. (1986), Probability forecasting, in Kotz, S., Johnson, N. L. and Read, C. B. (eds.), *Encyclopedia of Statistical Sciences*, Vol. 7, Wiley, New York, pp. 210–218.

Dawid, A. P., DeGroot, M. H. and Mortera, J. (1995), Coherent combination of experts' opinions (with discussion and rejoinder), *Test*, **4**, 263–313.

DeGroot, M. H. and Fienberg, S. E. (1982), Assessing probability assessors: Calibration and refinement, in Gupta, S. S. and Berger, J. O. (eds.), *Statistical Decision Theory and Related Topics III*, Vol. 1, Academic Press, New York, pp. 291–314.

DeGroot, M. H. and Fienberg, S. E. (1983), The comparison and evaluation of forecasters, *Statistician*, **32**, 12–22.

DeGroot, M. H. and Mortera, J. (1991), Optimal linear opinion pools, *Management Science*, **32**, 546–558.

Diebold, F. X., and Mariano, R. S. (1995), Comparing predictive accuracy, *Journal of Business and Economic Statistics*, **13**, 253–263.

Diebold, F. X., and Rudebusch, G. D. (1989), Scoring the leading indicators, *Journal of Business*, **62**, 369–391.

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.

Diks, C., Panchenko, V. and van Dijk, D. (2008), Partial likelihood-based scoring rules for evaluating density forecasts in tails, Tinbergen Institute Discussion Paper TI 2008-050/4.

Elder, R., Kapetanios, G., Taylor, T. and Yates, T. Y. (2005), Assessing the MPC's fan charts, *Bank of England Quarterly Bulletin*, Autumn, 326–348.

Elliott, G., and Timmermann, A. (2008), Economic forecasting, *Journal of Economic Literature*, **46**, 1–53.

Engle, R. F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **45**, 987–1007.

Foster, D. P. and Vohra, R. V. (1998), Asymptotic calibration, *Biometrika*, **85**, 379–390.

Genest, C. and McConway, K. J. (1990), Allocating the weights in the linear opinion pool, *Journal of Forecasting*, **9**, 53–73.

Genest, C. and Schervish, M. J. (1985), Modeling expert judgements for Bayesian updating, *Annals of Statistics*, **13**, 1198–1212.

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

Geweke, J. and Amisano, G. (2008), Optimal prediction pools, Working paper, University of Iowa.

Giacomini, R. and Komunjer, I. (2005), Evaluation and combination of conditional quantile forecasts, *Journal of Business and Economic Statistics*, **23**, 416–431.

Giacomini, R. and White, H. (2006), Tests of conditional predictive ability, *Econometrica*, **74**, 1545–1578.

Glahn, H. R. and Lowry, D. A. (1972), The use of model output statistics (MOS) in objective weather forecasting, *Journal of Applied Meteorology*, **11**, 1203–1211.

Gneiting, T. (2008), Editorial: Probabilistic forecasting, *Journal of the Royal Statistical Society Series A*, **171**, 319–321.

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

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007), Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society Series B*, **69**, 243–268.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008), Rejoinder on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds, *Test*, **17**, 256–264.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006), Calibrated probabilistic forecasting at the Stateline wind energy center: The Regime-Switching Space-Time method, *Journal of the American Statistical Association*, **101**, 968–979.

Granger, C. W. J. (1989), Combining forecasts — Twenty years later, *Journal of Forecasting*, **8**, 167–173.

Granger, C. W. J., and M. H. Pesaran (2000), A decision theoretic approach to forecast evaluation, in Chan, W.-S., W. K. Li, and H. Tong, eds., *Statistics and Finance: An Interface*, London: Imperial College Press, 261–278.

Good, I. J. (1952), Rational decisions, *Journal of the Royal Statistical Society Series B*, **14**, 107–114.

Graham, J. R. (1996), Is a group of economists better than one? Than none?, *Journal of Business*, **69**, 193–232.

Hall, S. G. and Mitchell, J. (2007), Combining density forecasts, *International Journal of Forecasting*, **23**, 1–13.

Hoeting, J. A., Madigan, D. M., Raftery, A. E. and Volinsky, C. T. (1999), Bayesian model averaging: A tutorial, *Statistical Science*, **14**, 382–401.

Hora, S. C. (2004), Probability judgements for continuous quantities: Linear combinations and calibration, *Management Science*, **50**, 597–604.

Jouini, M. N. and Clemen, R. T. (1996), Copula models for aggregating expert opinions, *Operations Research*, **44**, 444–457.

Kynn, M. (2008), The 'heuristics and biases' bias in expert elicitation, *Journal of the Royal Statistical Society Series A*, **171**, 239–264.

Lehrer, E. (2001), Every inspection is manipulaple, *Econometrica*, **69**, 1333–1347.

Laio, F., and Tamea, S. (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrology and Earth System Sciences*, **11**, 1267–1277.

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

Madigan, D., and Raftery, A. E. (1994), Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of the American Statistical Association*, **89**, 1535–1546.

Matheson, J. E., and Winkler, R. L. (1976), Scoring rules for continuous probability distributions, *Management Science*, **22**, 1087–1096.

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*, **67S**, 995–1033.

Murphy, A. H. (1973), A new vector partition of the probability score, *Journal of Applied Meteorology*, **12**, 595–600.

Murphy, A. H. (1998), The early history of probability forecasts: Some extensions and clarifications, *Weather and Forecasting*, **13**, 5–15.

Murphy, A. H. and Winkler, R. L. (1987), A general framework for forecast verification, *Monthly Weather Review*, **115**, 1330–1338.

Murphy, A. H. and Winkler, R. L. (1992), Diagnostic verification of probability forecasts, *International Journal of Forecasting*, **7**, 435–455.

Pinson, P., Chevallier, C. and Kariniotakis, G. N. (2007), Trading wind generation from short-term probabilistic forecasts of wind power, *IEEE Transactions on Power Systems*, **22**, 1148–1156.

Pal, S. (2009), On a conjectured sharpness principle for probabilistic forecasting with calibration, *Biometrika*, in press.

Pocernich, M. (2009), Verification: Forecast verification utilities, R package, version 1.29.

Primo, C., Ferro, C. A. T., Jolliffe, I. T. and Stephenson, D. B. (2009), Combination and calibration methods for probabilistic forecasts of binary events, *Monthly Weather Review*, **137**, 1142–1149.

Regnier, E. (2008), Doing something about the weather, *Omega*, **36**, 22–32.

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

Sandroni, A., Smorodinsky, R. and Vohra, R. V. (2003), Calibration with many checking rules, *Mathematics of Operations Research*, **28**, 141–153.

Schervish, M. J. (1989), A general method for comparing probability assessors, *Annals of Statistics*, **17**, 1856–1879.

Selten, R. (1998), Axiomatic characterization of the quadratic scoring rule, *Experimental Economics*, **1**, 43–62.

Sloughter, J. M., Raftery A. E., Gneiting, T. and Fraley, C. (2007), Probabilistic quantitative precipitation forecasting using Bayesian model averaging, *Monthly Weather Review*, **135**, 3209–3220.

Spiegelhalter, D. J. (1986), Probabilistic prediction in patient management and clinical trials, *Statistics in Medicine*, **5**, 421–433.

Stephenson, D. B., Coelho, C. and Jolliffe, I. T. (2008), Two extra components in the Brier score decomposition, *Weather and Forecasting*, **23**, 752–757.

Stone, M. (1961), The linear pool, *Annals of Mathematical Statistics*, **32**, 1339–1342.

Tetlock, P. E. (2005), *Expert Political Judgement: How Good is It? How Can we Know?*, Princeton University Press.

Timmermann, A. (2000), Density forecasting in economics and finance, *Journal of Forecasting*, **19**, 231–234.

Vovk, V. and Shafer, G. (2005), Good randomized sequential probability forecasting is always possible, *Journal of the Royal Statistical Society Series B*, **67**, 747–763.

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*, **189**, 64–71.

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

Wallsten, T. S. and Diederich, A. (2001), Understanding pooled subjective probability estimates, *Mathematical Social Sciences*, **18**, 1–18.

Wallsten, T. S., Budescu, D. V., Erev, I. and Diederich, A. (1997), Evaluating and combining subjective probability estimates, *Journal of Behavioral Decision Making*, **10**, 243–268.

Wilks, D. S. (1991), Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models, *Monthly Weather Review*, **119**, 1640–1662.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd edition, Academic Press.

Winkler, R. L. (1968), The consensus of subjective probability distributions, *Management Science*, **15**, B 61 –75.

Winkler, R. L. (1981), Combining probability distributions from dependent information sources, *Management Science*, **27**, 479-488.

Winkler, R. L. (1996), Scoring rules and the evaluation of probabilities (with discussion and rejoinder), *Test*, **5**, 1–60.

Winkler, R. L. and Jose, V. R. R. (2008), Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds, *Test*, **17**, 251–255.

Winkler, R. L. and Poses, R. M. (1993), Evaluating and combining physicians' probabilities of survival in an intensive care unit, *Management Science*, **39**, 1526–1543.

VITA

Roopesh Ranjan was born in Bihar, India. He did his schooling from his village town. In the year 2000, he moved to Calcutta to do his Bachelors and Masters at the Indian Statistical Institute. He studied Mathematics and Statistics at the institute for next five years. In the year 2005, he traveled to the United States to attend University of Washington, Seattle and pursue a PhD degree. Currently he lives in India.