Creating and Evaluating Combinations of Density Forecasts

Stephen G. Hall

James Mitchell

This Presentation Summarises 3 Papers

- Density Forecast Combinations
- Optimal Combinations of Density Forecasts
- Evaluating, Comparing and Combining Density Forecasts using the KLIC with an Application to the Bank of England and NIESR fan charts of inflation

Introduction
Density Forecast Combinations

- A number of Institutions have started to regularly publish density forecasts (eg the Bank of England's Fan charts, NIESR main forecast)

Chart 6.3
Current CPI inflation projection based on market interest rate expectations

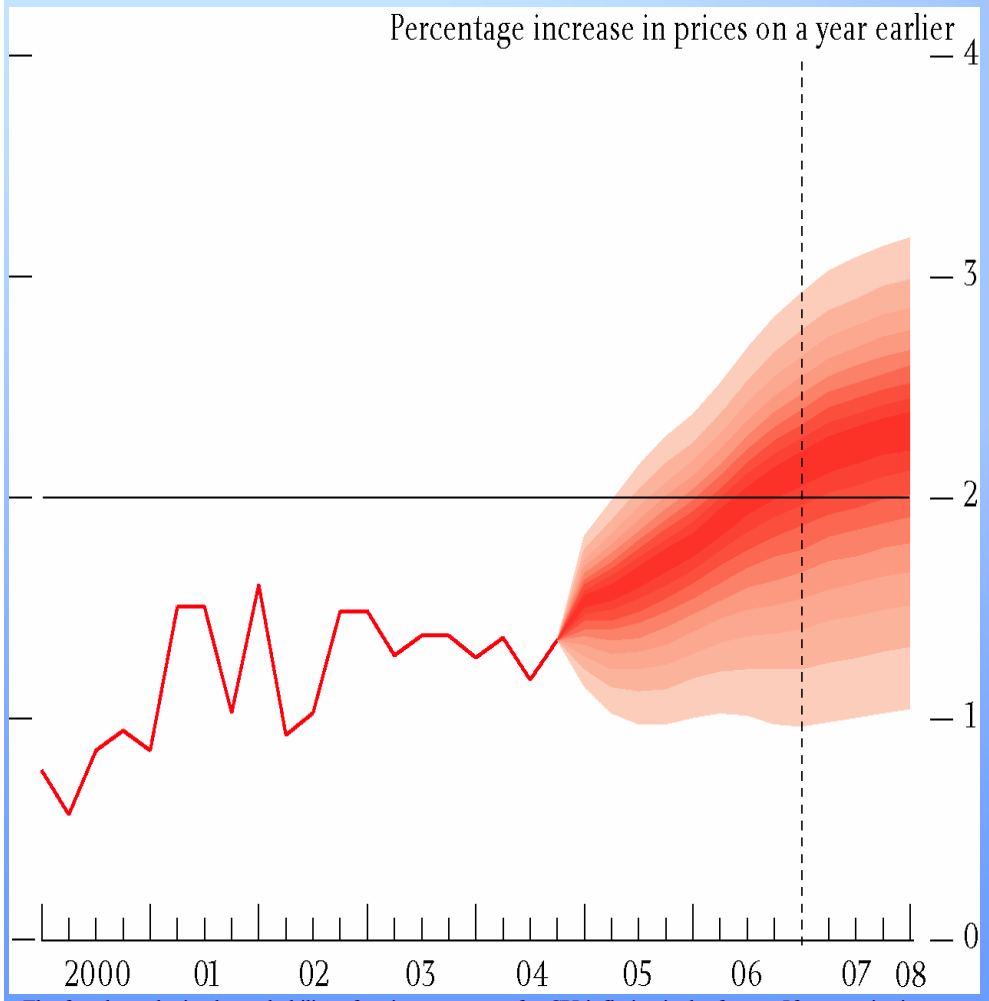
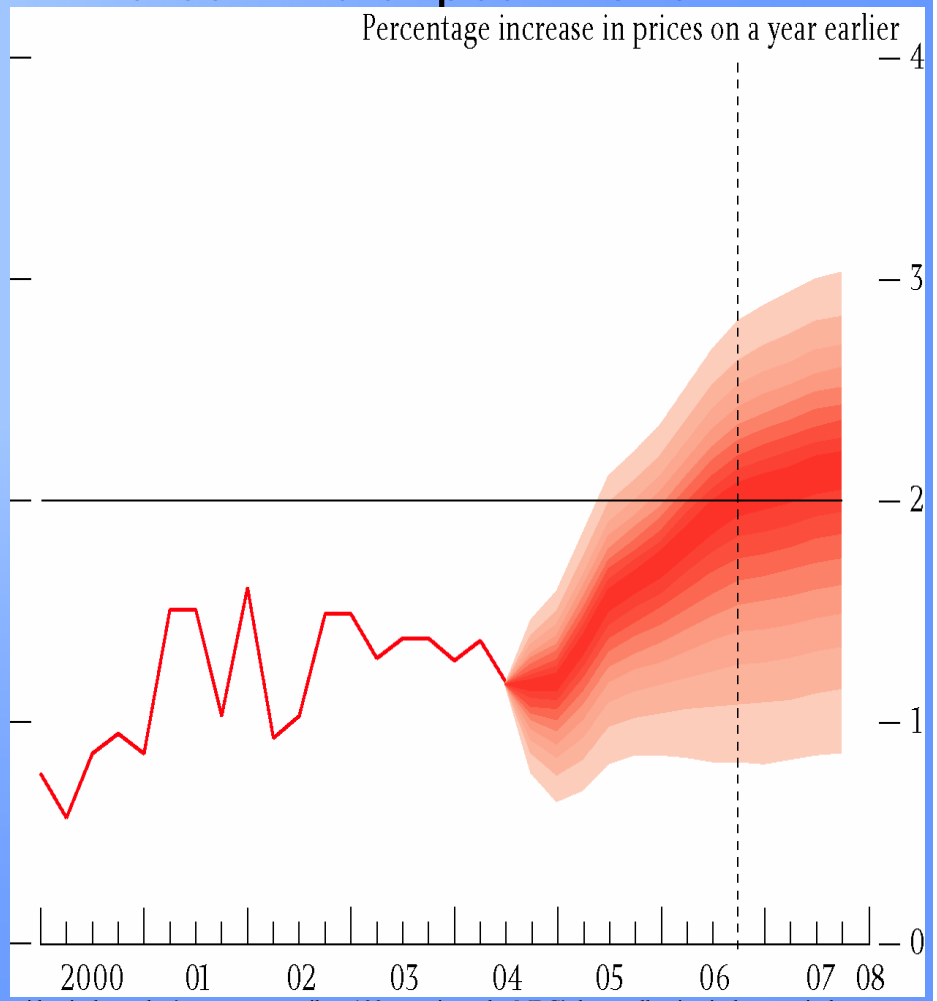


Chart 6.4
CPI inflation projection in November based on market interest rate expectations



The fan charts depict the probability of various outcomes for CPI inflation in the future. If economic circumstances identical to today's were to prevail on 100 occasions, the MPC's best collective judgement is that inflation over the subsequent three years would lie within the darkest central band on only 10 of those occasions. The fan charts are constructed so that outturns of inflation are also expected to lie within each pair of the lighter red areas on 10 occasions. Consequently, inflation is expected to lie somewhere within the entire fan charts on 90 out of 100 occasions. The bands widen as the time horizon is extended, indicating the increasing uncertainty about outcomes. See the box on pages 48–49 of the May 2002 Inflation Report for a fuller description of the fan chart and what it represents. The dotted lines are drawn at the respective two-year points.

Chart 6.7
Current projection for CPI inflation in 2007 Q1 [a] based on market interest rate expectations

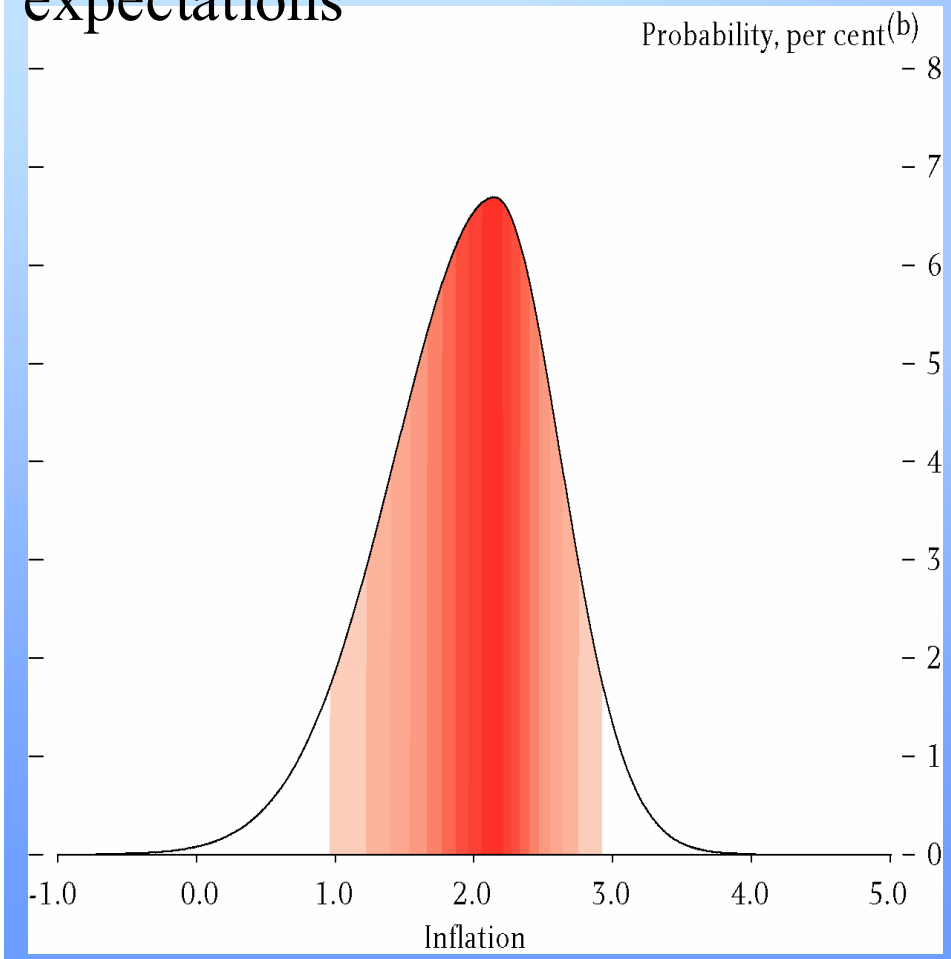
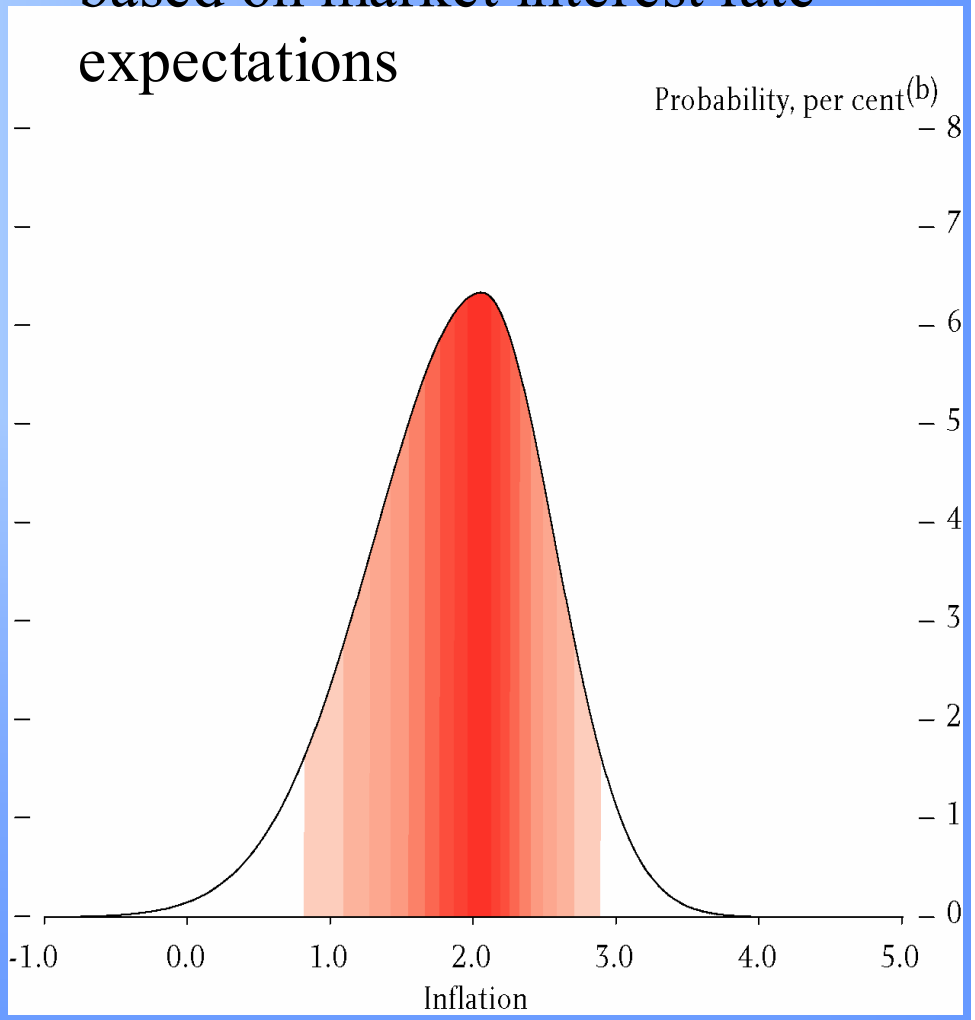


Chart 6.8
November projection for CPI inflation in 2007 Q1 [a] based on market interest rate expectations



(a) These charts represent a cross-section of the respective fan charts in 2007 Q1 for the market interest rate projections. The coloured bands have a similar interpretation to those on the fan charts. For further details on how the fan charts are constructed see the box on pages 48–49 in the May 2002 Inflation Report.

(b) Probability of inflation being within ±0.05 percentage points of any given inflation rate, specified to one decimal place. For example, the probability of inflation being 2.0% (between 1.95% and 2.05%) in the current projection is around 7%.

- The point forecast literature has long appreciated that combination forecasts normally outperform any single forecast

- There are debates about why this happens
  - All forecasts are wrong but in different ways
  - Simple averaging may help

- A natural question then is to ask, would a combined density forecast also work better.

- This raises a number of issues
  - How should we combine densities
  - How should we evaluate the combined density
  - How should we test individual densities against each other

Various proposals for combining densities

The early OR approaches

Consider N forecasts made by N experts of a variable y, if the N forecasters make density forecasts g_i i=1,N then the linear opinion pool is

$$p(y) = \sum_{i=1}^{N} w_i g_i(y),$$

W_i sum to one

The Logarithmic opinion pool is

$$p(y) = k \prod_{i=1}^{N} g_i(y)^{w_i},$$

However

How are the weights determined?

if all the experts agree that g is normal the combined density will be mixed normal and may be very strange

The Bayesian Approach (combining the means)

The experts densities are combined by a decision maker. Then Bayes Theorem is used to update the decision makers prior distribution

Consider a group of experts who each forecast the mean (m) and variance (v) of an event. The forecast error is,

$$s_i = m_i - y,$$

Assume the vector of forecast errors has covariance matrix $\sum$ which is assumed **KNOWN**

Then the decision maker combines the point forecasts as follows

$$\begin{aligned} h(y|\mathbf{m},\mathbf{v}) &= \frac{h(\mathbf{m}|y).h(\mathbf{v}|\mathbf{m},y).h(y)}{h(\mathbf{m},\mathbf{v})}, \\ &\propto h(\mathbf{m}|y).h(\mathbf{v}|\mathbf{m},y)h(y), \end{aligned}$$

And assuming normality

$$h(\mathbf{m}|y) \propto \exp\left[-\tfrac{1}{2}(\mathbf{m}-y\mathbf{e})'\boldsymbol{\Sigma}^{-1}(\mathbf{m}-y\mathbf{e})\right]$$

The combined forecast is then given by

$$h(y|\mathbf{m}) \propto \phi\left[(y - m^*)/\sigma_m^*\right],$$

Where ϕ is the standard normal density function

$$\begin{aligned} m^* &= \mathbf{e}'\boldsymbol{\Sigma}^{-1}\mathbf{m}/\mathbf{e}'\boldsymbol{\Sigma}^{-1}\mathbf{e}, \\ \sigma_m^{*2} &= 1/\mathbf{e}'\boldsymbol{\Sigma}^{-1}\mathbf{e}. \end{aligned}$$

Although this derivation is unusual the weights are simply those given by a regression of the forecasts on the outturn

Extending the Bayesian approach to higher moments

Let $\mu_t \; and \; \bar{\sigma}_t^2$ be the mean and variance of y

And let m_{it} and v_{it} be expert i's forecast of the mean and variance, and let

$$s_{it}^2 = (y_t - m_{it})^2; \; (i = 1, \ldots, N).$$

we now also have an error in the forecast of the variance

$$v_{it} - s_{it}^2 = u_{it}.$$

Then, with some normality assumptions spelt out in the paper

Where

$$h(\overline{\sigma}_t^2|\mathbf{v}) \propto \phi\left[(\overline{\sigma}_t^2 - v_t^*)/\sigma_v^*\right],$$

$$\begin{aligned} v_t^* &= \mathbf{e}'\Omega_t^{-1}\mathbf{v}_t/\mathbf{e}'\Omega_t^{-1}\mathbf{e}, \\ \sigma_{vt}^{*2} &= 1/\mathbf{e}'\Omega_t^{-1}\mathbf{e}. \end{aligned}$$

v_t^* is the optimal combined variance

And Ω_t can be based on historical data

Combining the complete density

So far we have considered combining the mean and variance, but we can of course combine a complete set of density forecasts.

2 approaches

Indirect - combining moments

Direct – complete combination of densities

Indirect method

Assume a distribution for the combined density, then estimate its moments by combining the individual moments of the forecasters distributions.

Advantage; If every forecaster thinks the distribution is normal the combined distribution will be normal

Direct method

Linear opinion pool

Here we propose a full bayesian combination method

$$h(y_t|\mathbf{m}_t, \mathbf{v}_t) \propto h(\mathbf{m}_t|y_t).h(\mathbf{v}_t|\mathbf{m}_t, y_t).h(y_t),$$

Given that $h(y_t)$ is uniform

$$h(y_t|\mathbf{m}_t, \mathbf{v}_t) \propto h(\mathbf{m}_t|y_t).h(\mathbf{v}_t|\mathbf{m}_t, y_t).$$

And so

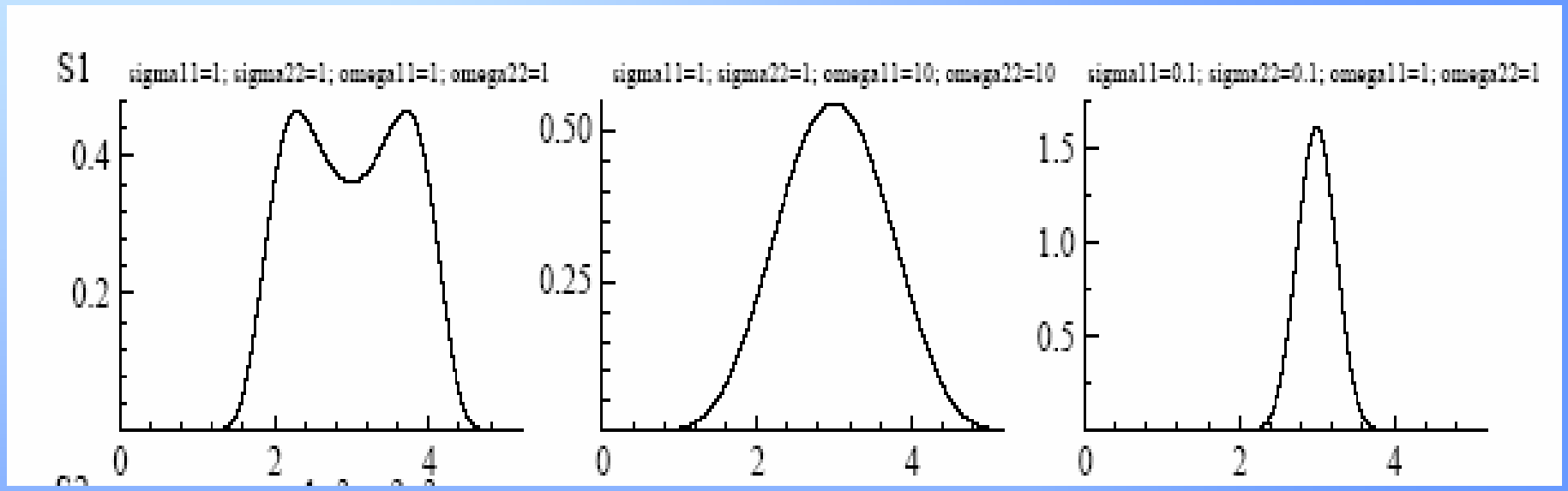
$$h(y_t|\mathbf{m}_t, \mathbf{v}_t) \propto \exp\left[\begin{array}{c} -\frac{1}{2}(\mathbf{m}_t - y_t\mathbf{e})'\boldsymbol{\Sigma}_t^{-1}(\mathbf{m}_t - y_t\mathbf{e}) - \\ \frac{1}{2}\left(\mathbf{v}_t - (y_t\mathbf{e} - \mathbf{m}_t)^2\right)'\boldsymbol{\Omega}_t^{-1}\left(\mathbf{v}_t - (y_t\mathbf{e} - \mathbf{m}_t)^2\right) \end{array}\right]$$

The densities are combined according to the reliability of their first two moments

However with this method even if all the forecasters agree that the distribution is normal the combined distribution may be far from normal.

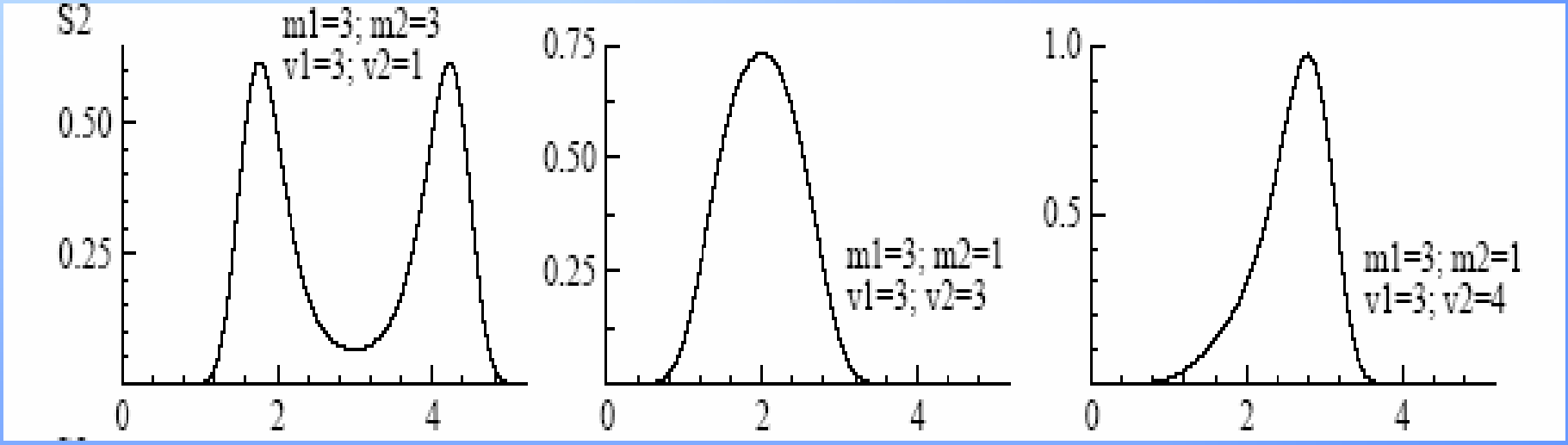
We illustrate this below

Both forecasters agree on the mean and variance, just the uncertainty differs

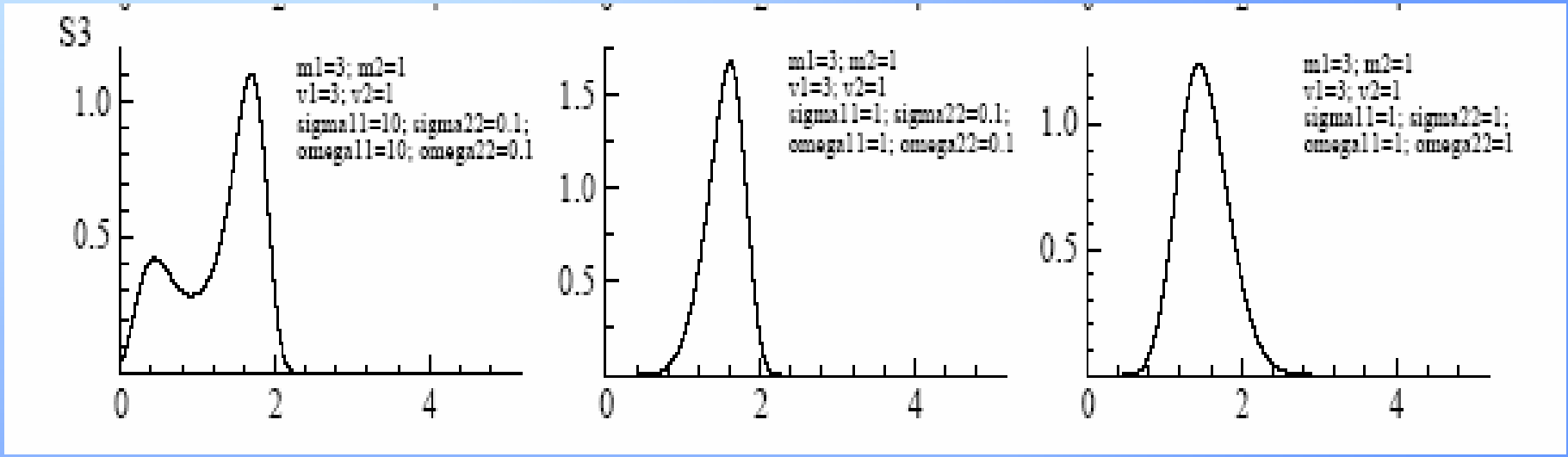


Even in this case the combined forecast can be non normal

Forecasters disagree on mean and variance but decision maker gives them equal weight



Forecasters disagree on mean and variance decision maker gives them different weights



Evaluation of Density Forecasts

This is done by calculating the probability integral transform (PIT) following Diebold et al(1998)

Given a sequence of estimated density forecasts $P(y_t)$ then

$$z_t = \int_{-\infty}^{y_t} p_t(u)du; \ (t = 1, ..., T).$$

Z will be both iid and uniform(0,1) if the estimated density functions are correct

It is often more convenient to take the inverse normal cumulative density function of z (say z*) which should then be standard normal and test this for normality.

There are then many tests either for iid or normality which can be used, we use 10 such tests in our appication

Here we compare the Bank of England and the NIESR density forecasts of inflation and compare them with a combination forecast

The Bank has an asymmetric density forecast based on a two piece normal distribution

NIESRs forecast is a normal density

Table 1 in the paper summarises these forecasts

We also compare these results with a benchmark forecast which is the assumed gaussian with mean equal to the previous years inflation and varaince estimated from the sample

We also estimate the tests and combination in sample and recursively (that is the combination is performed recursively rather than using the whole sample weights)

Ind – combination of bank/NIESR indirect method

Ind2- combination bank and benchmark indirect

	In-sample: 93q1-03q2					
	Bank	NIESR	Bench	Ind	Ind2	Dir
Bias	-0.104	-0.187	-0.067	-0.110	-0.086	-0.048
RMSE	0.5312	0.8384	0.4833	0.5389	0.4581	0.6386
KS	0.156	0.298	0.184	0.223	0.128	0.101
KS:cv	0.205	0.205	0.205	0.205	0.205	0.205
AD	1.723	5.374	2.545	3.670	0.887	0.344
DH	0.702	0.008	0.888	0.345	0.387	0.750
BN	0.316	0.508	0.603	0.175	0.494	0.830
Ber	0.000	0.000	0.000	0.000	0.000	0.302
JB	0.663	0.019	0.807	0.411	0.463	0.880
DM	0.002	0.000	0.072	0.000	0.202	0.246
mean	-0.057	-0.093	-0.190	-0.070	-0.087	-0.005
sd	0.661	0.440	1.377	0.488	0.837	0.887
Q_T	2.115	5.903	2.310	2.790	2.305	0.670
cvm	0.236	0.982	0.269	0.566	0.144	0.043
cvm1	1.467	2.601	1.712	1.628	1.302	0.367
cvm2	0.412	2.320	0.328	0.596	0.859	0.261
H: M1	0.002	0.005	0.011	0.004	0.005	0.000
H: iid	7.414	16.290	9.318	8.526	6.569	1.311
LB1	0.003	0.000	0.000	0.001	0.004	0.203
LB2	0.058	0.000	0.047	0.095	0.002	0.447
LB3	0.005	0.000	0.000	0.002	0.000	0.086

Recursive Out-of-Sample: 97q3-03q2				
Bank	NIESR	Bench	Ind	Dir
-0.020	0.053	-0.036	-0.011	-0.051
0.4044	0.4543	0.458	0.4105	0.4744
0.143	0.317	0.237	0.224	0.171
0.269	0.269	0.269	0.269	0.269
0.675	4.181	1.760	2.122	0.999
0.752	0.986	0.241	0.782	0.460
0.653	0.861	0.327	0.491	0.619
0.007	0.000	0.002	0.000	0.022
0.825	0.871	0.381	0.677	0.795
0.056	0.000	0.195	0.000	0.009
-0.009	0.028	-0.103	-0.017	-0.032
0.721	0.305	1.325	0.484	0.683
1.281	1.435	1.514	1.510	0.888
0.089	0.768	0.269	0.329	0.145
0.677	0.450	0.924	0.636	0.424
0.516	0.218	0.321	0.545	0.320
0.001	0.002	0.010	0.001	0.001
2.997	1.634	4.580	2.729	1.329
0.061	0.205	0.011	0.074	0.144
0.054	0.372	0.020	0.174	0.652
0.062	0.502	0.008	0.049	0.404

Summary of results

The Bank passes most tests

NIESR does not do so well, failing many distributional tests and independence

Benchmark does quite well

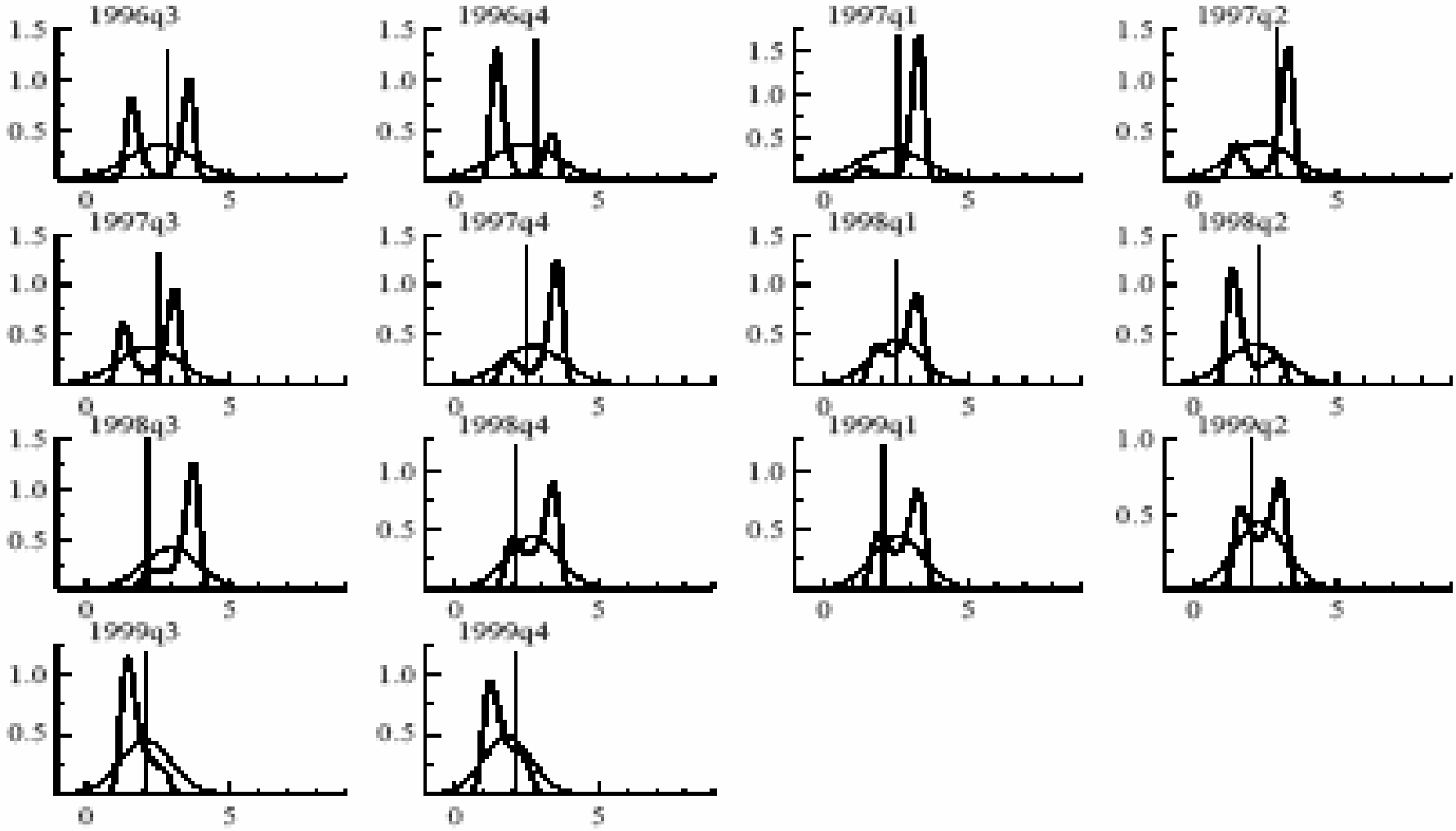
Indirect combination of Bank and NIESR does not improve over the bank (NIESR gets a very low weight)

Indirect combination of Bank and Benchmark is better

Direct combination of Bank and NIESR is better

Much the same is true of out of sample tests

Examples of the direct and indirect combined densities



Optimal Combinations of Density Forecasts

- Any of the direct forecast combination methods require (either explicitly or implicitly) a set of weights to combine the densities, e.g. the linear opinion pool

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

This paper makes a new suggestion as to how these weights may be calculated

Traditional point forecast combinations usually works by the regression method which forms the ‘optimal’ combinations so as to minimise the root mean square error of the combined forecast.

This is not possible for the complete density as we never observe the true density only a single realisation.

We extend the point forecast approach by analogy to chose the combination which gives the most accurate combined density

Diebold Gunther and Tay propose the idea that a density forecast is optimal if the model for the density is correctly specified. Using the PIT on a sequence of forecasts p(y) then if p is optimal

$$z_t = \int_{-\infty}^{y_t} p_t(u)du, \ (t = 1, \ldots, T).$$

Z is uniform and iid

The distribution p(y) is then optimal

There are a range of tests available for testing the properties of z, call a suitable test s(z)

We propose the optimal combination weights

$$\widehat{\mathbf{w}}, \text{ where } \mathbf{w} = (w_1, .., w_N),$$

As those weights that minimise the test statistic s(z)

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

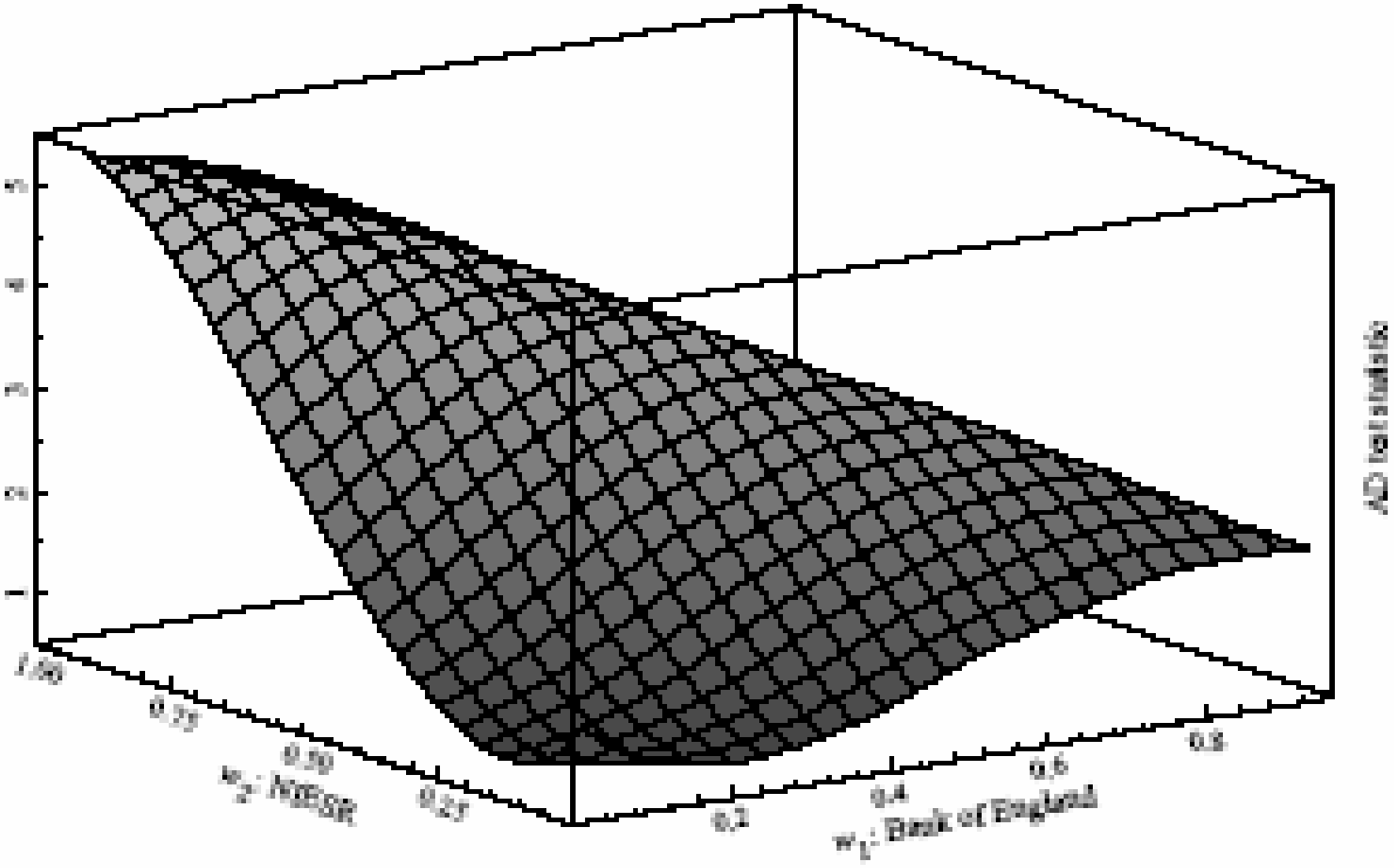
Here we use the Anderson and Darling test as s(z)

Example, Bank, NIESR and Benchmark (in sample and out of sample recursive)

w_1 weight on Bank

w_2 weight on NIESR

$(1-w_1-w_2)$ benchmark



Optimal weights bank 0.27 niesr 0.0 benchmark 0.73

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

Out of sample weights

Clearly combination brings about an improvement here

Evaluating, Comparing and Combining Density Forecasts using the KLIC with an Application to the Bank of England and NIESR fan charts of inflation

In this paper we consider the Kullback-Leibler information criterion (KLIC) as a natural metric to use when combining density forecasts

The KLIC provides a unifying framework which shows the equivalence of many of the existing tests of a density function.

It provides a formal framework for testing between density functions (a generalisation of Diebold and Mariano)

And optimally combining them

Berkowitz LR test

By taking the inverse normal cumulative density function of z say z* we can then test this for zero mean unit variance and independence within the following framework

$$z_{1t}^{*} = \mu + \rho z_{1t-1}^{*} + \varepsilon_t, \text{ where } Var(\varepsilon_t) = \sigma^2.$$

The test statistic LR_B is then

$$LR_B = -2\left[L(0,1,0) - L(\widehat{\mu},\widehat{\sigma}^2,\widehat{\rho})\right],$$

The KLIC and the Berkowitz LR test

The KLIC distance measure the distance between the true distribution f(y) and a density forecast g(y)

$$KLIC_{1t} = \int f_t(y_t) \ln \left\{ \frac{f_t(y_t)}{g_{1t}(y_t)} \right\} dy_t \text{ or}$$
$$KLIC_{1t} = \mathsf{E}\left[\ln f_t(y_t) - \ln g_{1t}(y_t)\right].$$

This can be consistently estimated by

$$KLIC_1 = \frac{1}{T} \sum_{t=1}^{T} \left[\ln f_t(y_t) - \ln g_{1t}(y_t)\right].$$

Which still requires us to know the truth

However

$$\ln f_t(y_t) - \ln g_{1t}(y_t) = \ln p_t(z^*_{1t}) - \ln \phi(z^*_{1t}),$$

And we know the inverse normal cumulative density function of the PIT using the true distribution is iid N(0,1). We then have to make some assumption regarding p to make the KLIC operational

For example if we assume p is a normal distribution testing the departure of Z* from iid N(0,1) is equivalent to evaluating the KLIC

To test the null hypothesis that f=g, consider

$$d_t = [\ln f_t(y_t) - \ln g_{1t}(y_t)] = [\ln p_t(z_{1t}^*) - \ln \phi(z_{1t}^*)] ;$$

The null hypothesis that g is correct is then

$$H_0 : \mathsf{E}(d_t) = 0 \Rightarrow KLIC_1 = 0$$

And

$$\overline{d} = KLIC_1 = \frac{1}{T} \sum_{t=1}^{T} [\ln p_t(z_{1t}^*) - \ln \phi(z_{1t}^*)] .$$

And

$$\sqrt{T}(\overline{d} - \mathsf{E}(d_t)) \xrightarrow{d} \mathsf{N}(0, \Omega),$$

This is proportional (2T) to the Berkowitz LR test

A Test of equal accuracy of two density forecasts

If we have two forecasts g_1 and g_2, then

$$\begin{aligned} d_t &= [\ln f_t(y_t) - \ln g_{1t}(y_t)] - [\ln f_t(y_t) - \ln g_{2t}(y_t)], \\ d_t &= \ln g_{2t}(y_t) - \ln g_{1t}(y_t), \\ d_t &= [\ln p_t(z_{1t}^*) - \ln \phi(z_{1t}^*)] - [\ln p_t(z_{2t}^*) - \ln \phi(z_{2t}^*)]. \end{aligned}$$

And the null of equal accuracy is then

$$H_0 : \mathsf{E}(d_t) = 0 \Rightarrow KLIC_1 - KLIC_2 = 0.$$

And the sample estimate is

$$\overline{d} = \frac{1}{T} \sum_{t=1}^{T} [[\ln p_t(z_{1t}^*) - \ln \phi(z_{1t}^*)] - [\ln p_t(z_{2t}^*) - \ln \phi(z_{2t}^*)]].$$

And again

$$\sqrt{T}(\overline{d} - \mathsf{E}(d_t)) \xrightarrow{d} \mathsf{N}(0, \Omega).$$

Combination of density forecasts using KLIC weights

Again consider bayesian model averaging

$$p_t(y_t \mid \Omega_t) = \sum_{i=1}^{N} w_{it} g_{it}(y_t);\ (t = 1, ..., T),$$

We may determine the weights from a bayesian perspective

$$w_{it} = \Pr(S_{it} \mid \Omega_t) = \frac{\Pr(\Omega_t \mid S_{it}) \Pr(S_{it})}{\sum_{i=1}^{N} \Pr(\Omega_t \mid S_{it}) \Pr(S_{it})}.$$

And we can derive these weights from a KLIC perspective

$$w_i = \frac{\exp(-\Delta_i)}{\sum_{i=1}^{N} \exp(-\Delta_i)} \ (i = 1, ..., N),$$

Where

$$\Delta_i = KLIC_i - \min_{N}(KLIC),$$

w_i can be interpreted as the probability that forecast i is the most accurate forecast in the KLIC sense

2 Monte Carlo Experiments

1 tests the accuracy of the KLIC weights

2 considers the size and power of the test of equal accuracy

Experiment 1

The true density is

$$f_t(y_t) = w_1 g_{1t}(y_t) + (1 - w_1) g_{2t}(y_t),$$

And

$$\begin{aligned} g_{1t}(y_t) &= N(\alpha x_{1t}, 1), \\ g_{2t}(y_t) &= N(\beta x_{2t}, 1), \end{aligned}$$

2 conclusions

KLIC weights are more accurate for w_1=1 or 0

KLIC weights more accurate the more different the two distributions are

Experiment 2

DGP

$$y_t = \alpha x_{1t} + \beta x_{2t} + \varepsilon_t,$$

x_1 and x_2 are uncorrelated N(0,1)

Two incorrect density forecasts

$$\begin{aligned} y_t &= \alpha x_{1t} + \varepsilon_{1t}, \\ y_t &= \beta x_{2t} + \varepsilon_{2t}. \end{aligned}$$

Findings

1, the more different are the two forecasts the more power to the KLIC test

2, The power also depends on the accuracy of the individual forecasts

3, estimated KLIC weights biased towards 0.5 unless the best model is very bad

4, Combined forecast generally does better

Application, Bank and NIESR

Individual results

	1993q1-2003q2		p-values				1997q3-2003q2					
	RMSE	$\overline{S}$	KLIC_{IN}	KLIC_{N}	LR_{IN}	LR_{N}	RMSE	$\overline{S}$	KLIC_{IN}	KLIC_{N}	LR_{IN}	LR_{N}
Bank	0.53	-0.83	0.33	0.14	0.000	0.003	0.40	-0.61	0.25	0.10	0.007	0.097
Bank_{N}	0.53	-0.83	0.33	0.14	0.000	0.003	0.40	-0.62	0.25	0.10	0.008	0.095
NIESR	0.83	-1.57	0.91	0.43	0.000	0.000	0.45	-1.42	0.86	0.75	0.000	0.000
NIESR_{R}	0.83	-1.21	0.20	0.02	0.001	0.435	0.45	-0.90	0.12	0.03	0.127	0.534

Over the full sample we reject both density forecasts

Over the later sample the Banks and NIESR_{R} densities are accepted

Imposing normality on the Bank seems acceptable

Testing the density forecasts

DM: Bank=NIESR	1.189
KLIC equal: Bank=NIESR	10.007
KLIC equal: Bank=NIESR$_R$	4.551
KLIC equal: Bank=Bank$_N$	0.374

DM finds no real difference in the point forecast

Bank and NIESR densities are clearly different

Bank and NIESR$_r$ also different

Bank and Bank$_n$ almost the same

Combined KLIC weights

		w_1	w_2
Point		1.042	-0.106
	robust e.s.e.	(0.213)	(0.227)
Density		0.572	0.428
Density: Bank vs. NIESR_R		0.470	0.530

The KLIC weights give NIESR a relatively large weight

Combined forecast performance

	In-sample (1993q1-2003q2)			Out-of-sample (1997q3-)		
Point Forecasts	RMSE			RMSE		
Optimal	0.497			0.457		
Equal	0.642			0.397		
BMA	0.619			0.393		
Density Forecasts	KLIC_N	$\overline{S}$	LR_N	KLIC_N	$\overline{S}$	LR_N
BMA	0.345	-1.076	0.000	0.392	-0.894	0.000
$N(m_t^*, v_t^*)$	0.010	-0.719	0.654	0.131	-0.676	0.043
BMA: Equal	0.373	-1.125	0.000	0.422	-0.894	0.000
BMA: Bank vs NIESR_R	0.133	-0.995	0.004	0.090	-0.720	0.114

BMA does not improve on the Banks own forecast

Equal weights does a little worse than the BMA

Moment combinations does much better

Combining Bank and NIESR_R does well

conclusion

- KLIC is a useful unifying framework to combine and test density forecast.
- However BMA can lead to highly non-normal distributions which may perform badly