Generalised Density Forecast Combinations

Nicholas Fawcett*, George Kapetanios†, James Mitchell‡ and Simon Price*,**

*Bank of England

**City University London

†Queen Mary, University of London

‡University of Warwick

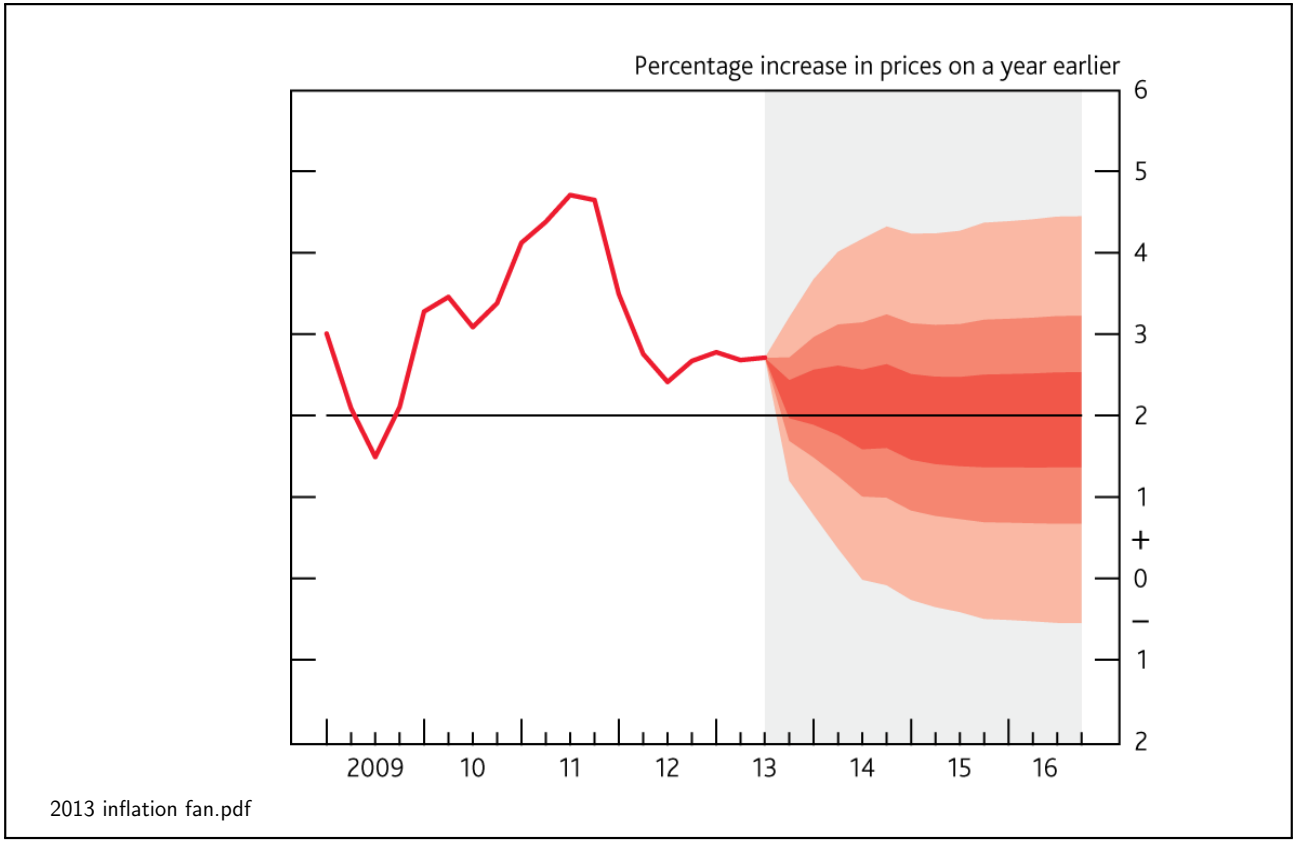
June 2014

Outline

Inflation Report density forecasts

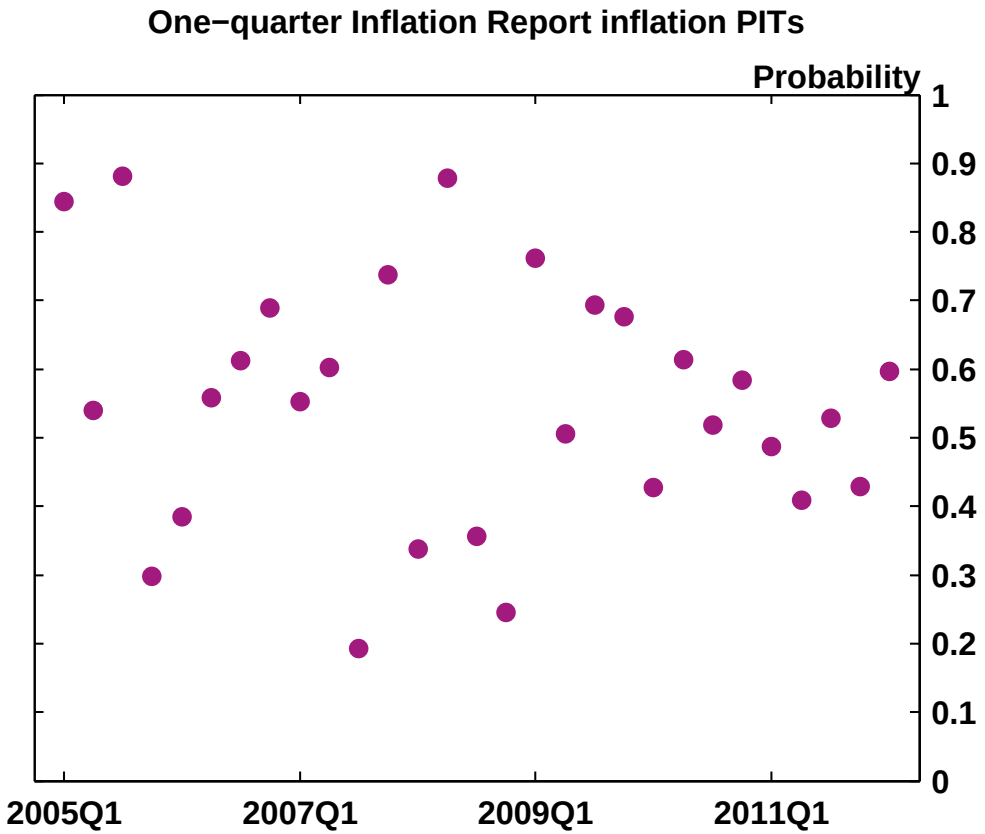
- Fan charts of various types centre-stage for Bank of England - inflation, growth, and now unemployment.
- Invariably emphasise uncertainty in policy discussions.
- BoE density forecasts - used to be, could do better, but not too bad (see eg Wallis 2004).
- But how have they done post crisis, a period of structural change and instability?
- Report period-by-period outcome by decile - cumulate to the PITS.

Inflation fan



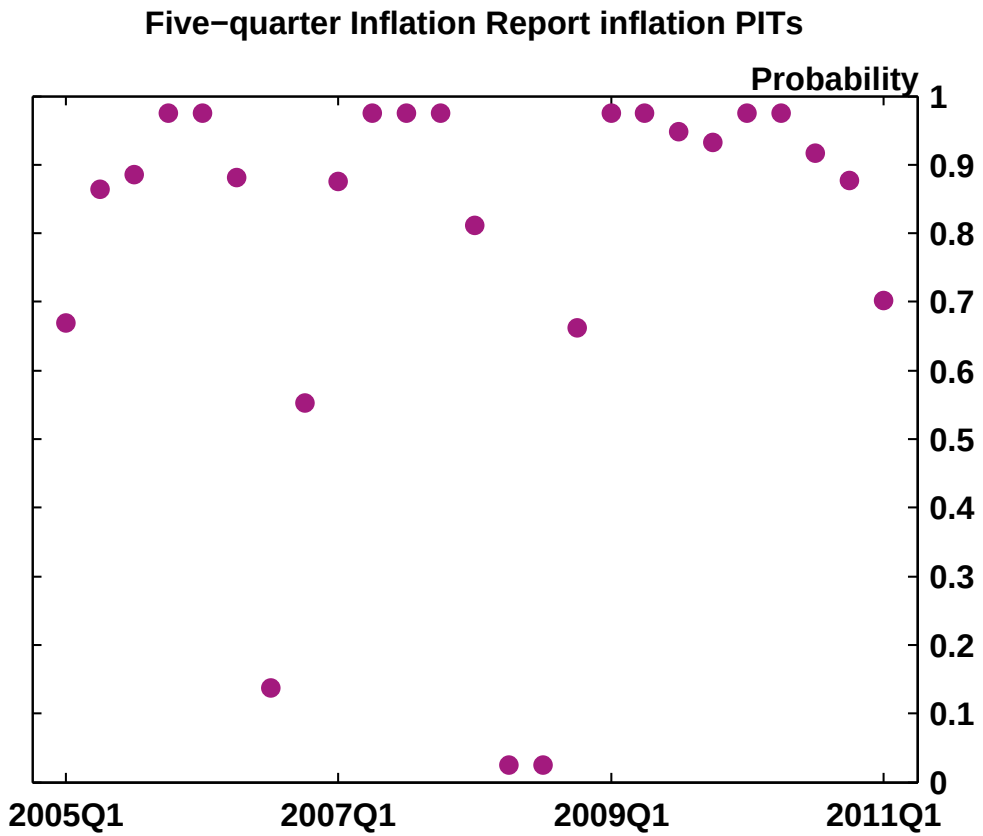
Market interest rate expectations and £375 billion asset purchases

CPI forecasts



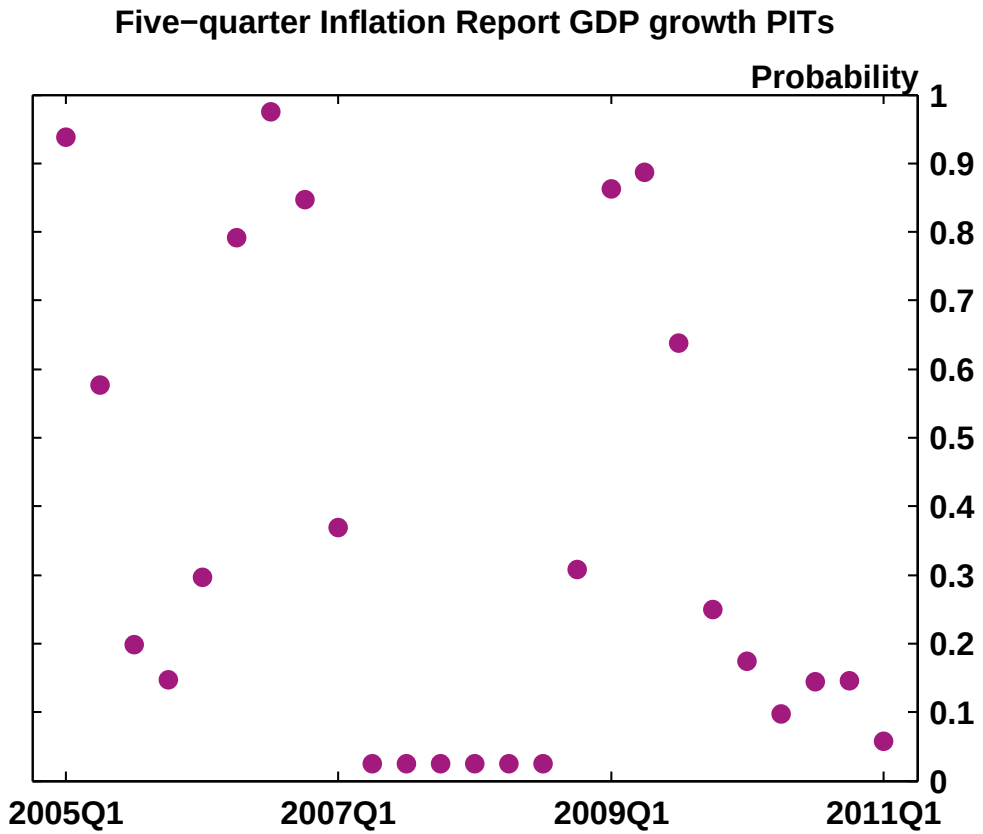
One-step inflation forecasts - density too wide (not enough large errors).

CPI forecasts



Five-step inflation forecasts - density too thinly populated in centre - too many large errors, especially on high inflation side.

GDP forecasts



Five-step growth forecast - density too thinly populated in centre - too many large errors at low growth outcomes (mean wrong).

Could do better

- Important to have good benchmarks for density forecasts.

Outline

Density forecasts

- Increasing interest in density forecasts.
- One way to get good density forecasts is model combination.
- An advantage is then you can then forecast effectively even with simple models.
- Seems very intuitive as well (all models are wrong, etc).
- Especially useful with instabilities and uncertainty about the preferred model; eg see Jore *et al.* (2010), Geweke & Amisano (2012) and Rossi (2012).
- Geweke & Amisano (2011) contrast Bayesian model averaging with linear opinion pools, where the weights on the component density forecasts are optimised to maximise the score, typically the logarithmic score, of the combination, as suggested in Hall & Mitchell (2007).

Our generalisation

- Existing literature on optimal density combinations tends to treat weights as fixed, like optimal point-forecast combinations à la Bates-Granger.
- One generalisation is to let weights follow flexible schemes.
- Specifically, let combination weights depend on the forecast variable or region of density.
- Allow for possibility that while one model may be particularly useful (and receive a high weight in the combination) when (eg) the economy is in recession, another model may be more informative when (eg) output growth is positive.
- Wide range of relevant applications - we apply to density forecasts of S&P500 daily return used in Geweke & Amsisano (2011).

Flexibility

- Might be an alternative to large N combinations.
- Contrasts with two recent suggestions where weights
  - follow a Markov-switching structure (Waggoner & Zha 2012)
  - evolve over time according to a Bayesian learning mechanism (Billio *et al.* 2013).
- Accommodating time variation in the combination weights mimics our approach to the extent that over time one moves into different regions of the forecast density.

Outline

The general approach

- We wish to provide a general scheme for combining density forecasts.
- We start by noting that we consider a stationary stochastic process of interest y_t, $t = 1, ..., T$ and a vector of predictor variables x_t, $t = 1, ..., T$.
- Our aim is to forecast the density of y_{t+1} conditional on the data available, formally $\mathfrak{F}_t = \sigma\left(x_{t+1}, (y_t, x_t')', ..., (y_1, x_1')'\right)$.

Conditional weights

- Assume the existence of N density forecasts $q_i\left(y|\mathfrak{F}_t\right) = q_{it}\left(y\right)$, $i = 1, ..., N$, $t = 1, ...T$.
- We propose a generic combined density forecast, given by the *generalised linear opinion pool*

$$p_t\left(y\right) = \sum_{i=1}^{N} w_i\left(y\right) q_{it}\left(y\right)$$

such that

$$\int p_t\left(y\right) dy = 1$$

where $w_i\left(y\right)$ are the weights on the individual density forecasts which themselves depend on y.

- Thus generalises existing approaches where $w_i\left(y\right) = w_i$.

Which weights?

- Define a predictive loss function given by

$$L_T = \sum_{t=1}^{T} l\left(p_t\left(y_t\right); y_t\right).$$

- We assume that there exist $w_i^0\left(y\right)$ in the space of q_i-integrable functions Ψ_{q_i} where

$$\Psi_{q_i} = \left\{ w\left(.\right) : \int w\left(y\right) q_i\left(y\right) dy < \infty \right\}, i = 1, .., N,$$

such that

$$E\left(l\left(p_t\left(y_t\right); y_t\right)\right) \equiv E\left(l\left(p_t\left(y_t; w_1^0, .., w_N^0\right); y_t\right)\right) \leq$$

$$E\left(l\left(p_t\left(y_t; w_1, .., w_N\right); y_t\right)\right)$$

for all $\left(w_1, .., w_N\right) \in \prod_i \Psi_{q_i}$.

- Ie, at least one minimising set of weights exist.

Choosing the weights by minimising a loss function

- We determine w_i by minimising L_T, ie

$$\{\hat{w}_{1T}, ..., \hat{w}_{NT}\} = \arg \min_{w_i, i=1,...,N} L_T. \quad (1)$$

- Problem impossible to solve without restrictions on the space searched over (Ψ_{q_i}).

Statistical properties

- In general cannot derive asymptotic properties.
- But with reasonable restrictions can prove that the method delivers the true density asymptotically, in the sense that it has the maximum log score over all potential densities.
- With a further set of assumptions we can establish aymptotic normality.
- This allows us to test whether it is useful to allow for the weights to depend on y (ie test $w(y) = w$).

Using boundary conditions

- Could use indicator functions, ie $\eta_s = I\left(r_{s-1} \leq y < r_s\right)$ where boundaries $r_0 < r_1 < ... < r_s$ with $s = 1, ..., p$; r and p either known *a priori* (unlikely) or estimated.
- Easily motivated: eg, some models might forecast better in recessions than booms, or when inflation < 1 *vs* > 3.
- In this case

$$p_t\left(y\right) = \sum_{i=1}^{N} \sum_{s=1}^{p_t} \nu_{is} q_{it}\left(y\right) I\left(r_{s-1} \leq y < r_s\right)$$

where ν_{is} are constants to be estimated and

$$\kappa_{is} = \int_Y I\left(r_{s-1} \leq y < r_s\right) q_{it}\left(y\right) dy = \int_{r_{s-1}}^{r_s} q_{it}\left(y\right) dy.$$

Determining p by cross validation

- In practice number of boundaries or regions p unknown.
- We use cross-validation (CV) to determine p.
- Choose p in range $1, ..., p^{\max}$, to min average loss associated with the series of recursive density forecasts over out-of-sample period $t_0, ..., T$.

$$\widehat{p} = \arg \min_{1 \leq p \leq p^{\max}} \sum_{t=t_0}^{T} l\left(p_t\left(y_{t+1}, \hat{\vartheta}_{t,p}\right); y_t\right),$$

$\hat{\vartheta}_{t,p}$ recursively computed estimate of ϑ_p (parameters determining density) for given value p at time t; loss function evaluated at outcome, y_t

- Likely that CV has desirable properties.

Evaluation - the log score

- Loss function - score assigned based on predictive density at t and value of y_t that emerges at $t+1$.
- Common choice L_T logarithmic scoring rule.
- If user's loss function unknown, by max log score still min Kullback-Leibler IC relative to true but unknown density.
- When zero, know from Diebold *et al.* (1998) all loss functions are minimised.
- Loss function:

$$L_T = \sum_{j=1}^{T} -\log p_j\left(y_{j+1}\right) = \sum_{j=1}^{T} -\log\left(\sum_{i=1}^{N}\sum_{s=1}^{p_t} \nu_{is} q_{it}\left(y_{j+1}\right) I\left(r_{s-1} \leq y_{j+1} < r_s\right)\right).$$

- Minimise this with respect to ν_{is} subject to $\sum_{i=1}^{N}\sum_{s=1}^{p_t} \nu_{is}\kappa_{is} = 1$ (density proper).

Estimating thresholds

- Unlikely thresholds known *a priori*.
- We use grid estimation - optimising on loss for each value in grid.
- Requires some judgement on range of y.
- Practically feasible for relevant examples.
- We are able to show subsampling allows asymptotically valid inference.

Weighted log score?

- Our weights hard to interpret - not simply related to that region of density, as they are restricted across the density.
- In an evaluation context, Diks *et al.* (2010) discuss weighted logarithmic scoring rule, $w_t(y_{t+1}) \log q_{it}(y_{t+1})$.
- Weight function $w_t(y_{t+1})$ emphasises regions of the density of interest.
- One possibility that $w_t(y_{t+1}) = I(r_{s-1} \leq y_{t+1} < r_s)$.
- But weighted logarithmic score rule is improper and can systematically favour misspecified densities even when the candidate densities include the true density.

Properties - summary

- Given reasonable assumptions, the method delivers the true density asympotically.
- With further assumptions eg the log score and piecewise linear weights, we can establish asymptotic normality enabling inference.

Outline

Performance metric

- Relative performance generalised *vs* linear pool assessed by tests for equal predictive accuracy using logarithmic scoring rule.
- Giacomini and White (2004) Wald-type test statistic.

$$T\left(T^{-1}\sum_{t=1}^{T}\Delta L_t\right)'\Sigma^{-1}\left(T^{-1}\sum_{t=1}^{T}\Delta L_t\right),$$

- ΔL_t difference in logarithmic scores of the generalised and linear pools, equals KLIC
- Σ robust estimate of asymptotic covariance matrix.
- Under null equal accuracy $E(\Delta L_t) = 0 \sim \chi^2_1$ as $T \to \infty$.
- Two-sided tests at nominal size of 10%.
- Report proportion of rejections in favour of both generalised and linear pools.

First Monte Carlo experiment

- True density 2-part-normal, which splices two separate normals at the mean using a normalising constant.

$$f(Y) = \begin{cases} A \exp(-y-\mu)^2/2\sigma_1^2 \text{ if } y < \mu \\ A \exp(-y-\mu)^2/2\sigma_2^2 \text{ if } y \geq \mu \end{cases}$$

$$A = \left(\sqrt{2\pi}(\sigma_1 + \sigma_2)/2\right)^{-1}$$

- Assume combine with two normals common mean.
- Max log score with fixed weights (linear pool, LP) and with our generalized method (GP):
  - known boundaries, ie $p = 2$ regions, split at 0
  - unknown boundaries, with 11 equally spaced points.
- Fix $\sigma_1^2 = 1$ and set $\sigma_2^2 = 1.5, 2, 4, 8$.
- In practice would use estimated densities.

First Monte Carlo experiment

- In this case there are true thresholds and Gaussian elements, so we examine properties more thoroughly than the other experiments.

1. G&W tests of relative performance, comparing effect of successively dropping the (unrealistic) assumptions that r and p are known.
2. Absolute performance (IMSE).
3. Size and power of the linearity test
4. Choice of p.

Setting 1 - two-part normal - relative performance

- Regardless of p and r_s known or estimated, GP preferred with high rejection probabilities.
- Proportion approaches 1 as σ_2^2 and T increase.
- Even for low σ_2^2 (less skew in true density) GP preferred with probabilities > 0.9 for larger T.
- Estimation of p and r_s reduces performance but GP still preferred.

Table 1 - two-part normal - relative performance

σ_2^2	T	$r_1 = 0$, $p = 2$		Est $r_1 \mid p = 2$		Est r_1, p	
		G/L	L/G	G/L	L/G	G/L	L/G
1.5	100	0.302	0.012	0.188	0.008	0.138	0.022
	200	0.514	0.000	0.380	0.000	0.254	0.008
	400	0.718	0.000	0.616	0.000	0.536	0.000
	1000	0.968	0.000	0.966	0.000	0.936	0.000
2	100	0.620	0.000	0.508	0.002	0.342	0.010
	200	0.872	0.000	0.752	0.000	0.688	0.004
	400	0.976	0.000	0.968	0.000	0.964	0.000
	1000	1.000	0.000	1.000	0.000	1.000	0.000
4	100	0.966	0.000	0.920	0.004	0.748	0.012
	200	1.000	0.000	0.996	0.002	0.932	0.004
	400	1.000	0.000	1.000	0.000	0.996	0.000
	1000	1.000	0.000	1.000	0.000	1.000	0.000
8	100	0.988	0.006	0.874	0.002	0.690	0.006
	200	1.000	0.000	0.986	0.000	0.872	0.004
	400	1.000	0.000	1.000	0.000	0.972	0.000
	1000	1.000	0.000	1.000	0.000	1.000	0.000

Rejection probabilities in favour of the Generalised (G) and Linear (L) pools using the Giacomini-White test for equal density forecast performance

Setting 1 - two-part normal - absolute performance

- Report integrated mean square errors as metric of performance.
- LP beats either component density.
- When threshold r_1 assumed known GP outperforms LP in all cases.
- Estimation of either r or both p and r reduces performance quite markedly.
- But for sufficiently large T and variance σ_2^2 the GP remains considerably better.

Table 2 - two-part normal - absolute performance

σ_2^2	T	Generalised			Linear	Component Densities	
		$r_1 = 0$	est $r_1 \mid p = 2$	Est p		Comp.1	Comp. 2
1.5	100	0.343	2.293	1.910	0.773	1.654	1.103
	200	0.167	0.883	0.926	0.715	1.654	1.103
	400	0.084	0.389	0.456	0.691	1.654	1.103
	1000	0.032	0.149	0.197	0.673	1.654	1.103
2	100	0.297	1.520	1.562	1.592	4.421	2.211
	200	0.139	0.679	0.772	1.531	4.421	2.211
	400	0.072	0.293	0.445	1.505	4.421	2.211
	1000	0.030	0.115	0.203	1.486	4.421	2.211
4	100	0.189	0.619	0.872	2.649	12.728	3.182
	200	0.090	0.310	0.414	2.595	12.728	3.182
	400	0.045	0.139	0.218	2.569	12.728	3.182
	1000	0.019	0.062	0.103	2.556	12.728	3.182
8	100	0.101	0.291	0.373	2.221	19.413	2.427
	200	0.053	0.139	0.196	2.190	19.413	2.427
	400	0.026	0.073	0.100	2.174	19.413	2.427
	1000	0.011	0.033	0.048	2.164	19.413	2.427

IMSE estimates for the Generalised and Linear Combinations

Setting 1 - two-part normal - linearity tests

- Test generally slightly undersized.
- But power increases rapidly with T and σ_2^2.
- In general, very powerful test.

Table 3 - two-part normal - size of linearity test

T/σ_2^2	1.1	1.25	1.5	2
100	0.021	0.021	0.014	0.014
200	0.017	0.014	0.016	0.012
400	0.042	0.019	0.014	0.023
1000	0.070	0.042	0.021	0.019

Rejection probabilities for linearity test under null of linearity

Table 4 - two-part normal - power of linearity test

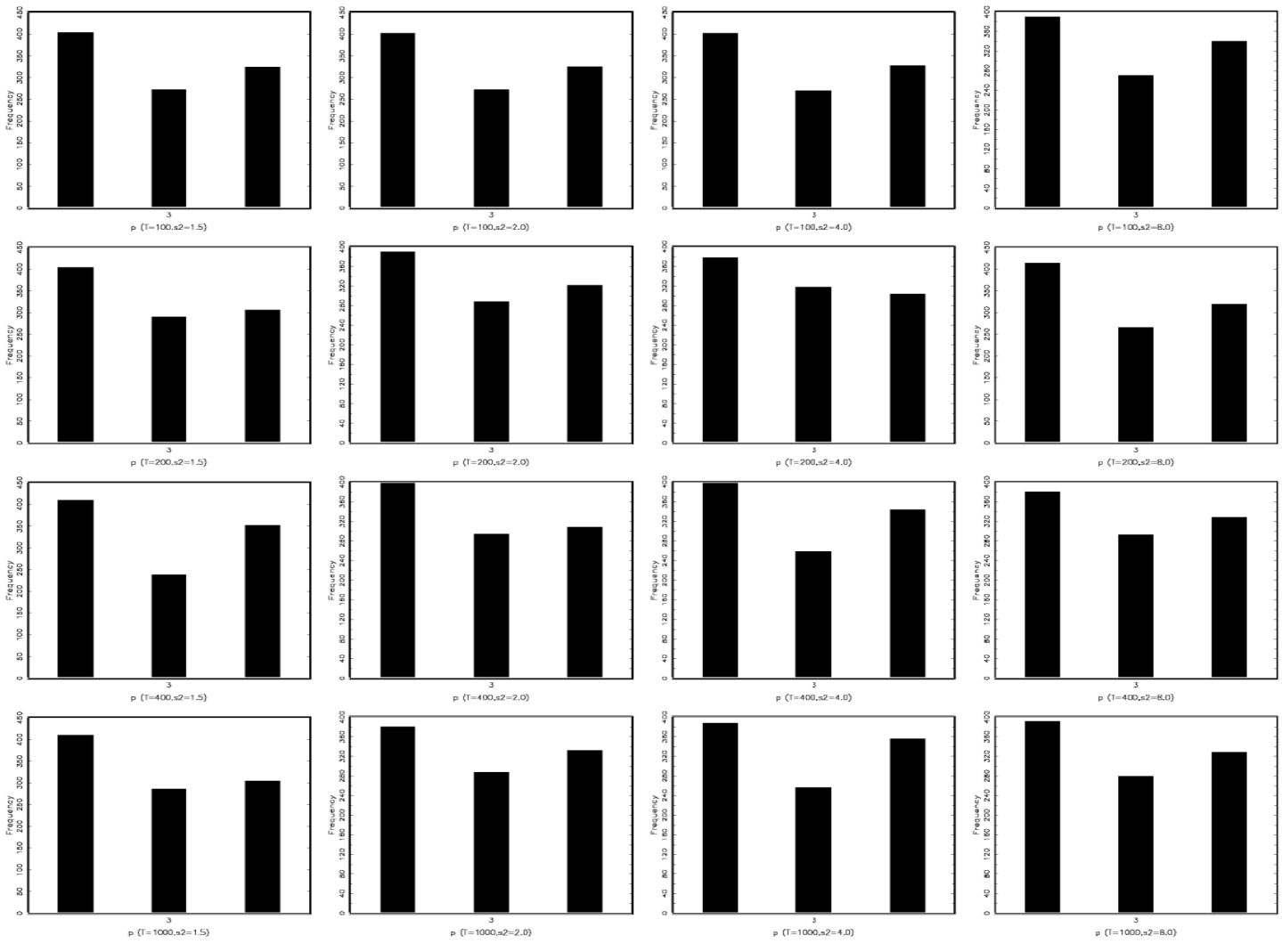
T/σ_2^2	1.1	1.25	1.5	2
100	0.024	0.134	0.611	0.998
200	0.028	0.267	0.915	1.000
400	0.081	0.552	0.997	1.000
1000	0.420	0.950	1.000	1.000

Rejection probabilities for linearity test under alternative

Choice of p in two-part normal

- Search over number of regions p restricted to between 2 and 4.
- Number of times a given p value was selected by cross validation for each of the T and σ_2^2 cases.
- $p = 2$ modal although larger p often selected.
- Reassuring: CV protects against over-fitting.
- Tables 1 and 2 showed that despite estimation uncertainty leading to incorrect choice of p, GP preferred to LP.

Choice of p by CV, cut by T and σ_2^2



Fourth Monte Carlo experiment

- Motivation: setup for a macro environment.
- UC model with SV Stock and Watson (2007) for US inflation.
- Allows variances of both the permanent and transitory component to evolve.

$$\begin{aligned} \pi_t &= \tau_t + \eta_t, \text{ where } \eta_t = \sigma_{\eta,t}\zeta_{\eta,t} \\ \tau_t &= \tau_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t = \sigma_{\varepsilon,t}\zeta_{\varepsilon,t} \\ \ln \sigma^2_{\eta,t} &= \ln \sigma^2_{\eta,t-1} + \upsilon_{\eta,t} \\ \ln \sigma^2_{\varepsilon,t} &= \ln \sigma^2_{\varepsilon,t-1} + \upsilon_{\varepsilon,t} \end{aligned}$$

- $\zeta_t = (\zeta_{\eta,t}, \zeta_{\varepsilon,t})$ is i.i.d. $N(0, I_2)$, $\upsilon_t = (\upsilon_{\eta,t}, \upsilon_{\varepsilon,t})$ is i.i.d. $N(0, \gamma I_2)$, ζ_t and υ_t are independently distributed, γ is 0.01 scalar.
- Component density forecasts UC models *sans* SV.
- First, calibrated according to a model S&W found fitted well for high inflation: $\sigma_\eta = 0.66$ and $\sigma_\varepsilon = 0.91$.
- Second, good fit Great Moderation: $\sigma_\eta = 0.61$, $\sigma_\varepsilon = 0.26$.

Setting 4 - UC SV - results

- Again, relative performance depends on T.
- For $T = 100$ LP preferred more frequently than GP.
- For larger T reverse is the case - GP preferred more frequently than LP.
- Suggests may be useful in macro applications.

Table 7 - UC SV

T	Unknown p	
	G/L	L/G
100	0.124	0.168
200	0.270	0.130
400	0.354	0.126
1000	0.476	0.180

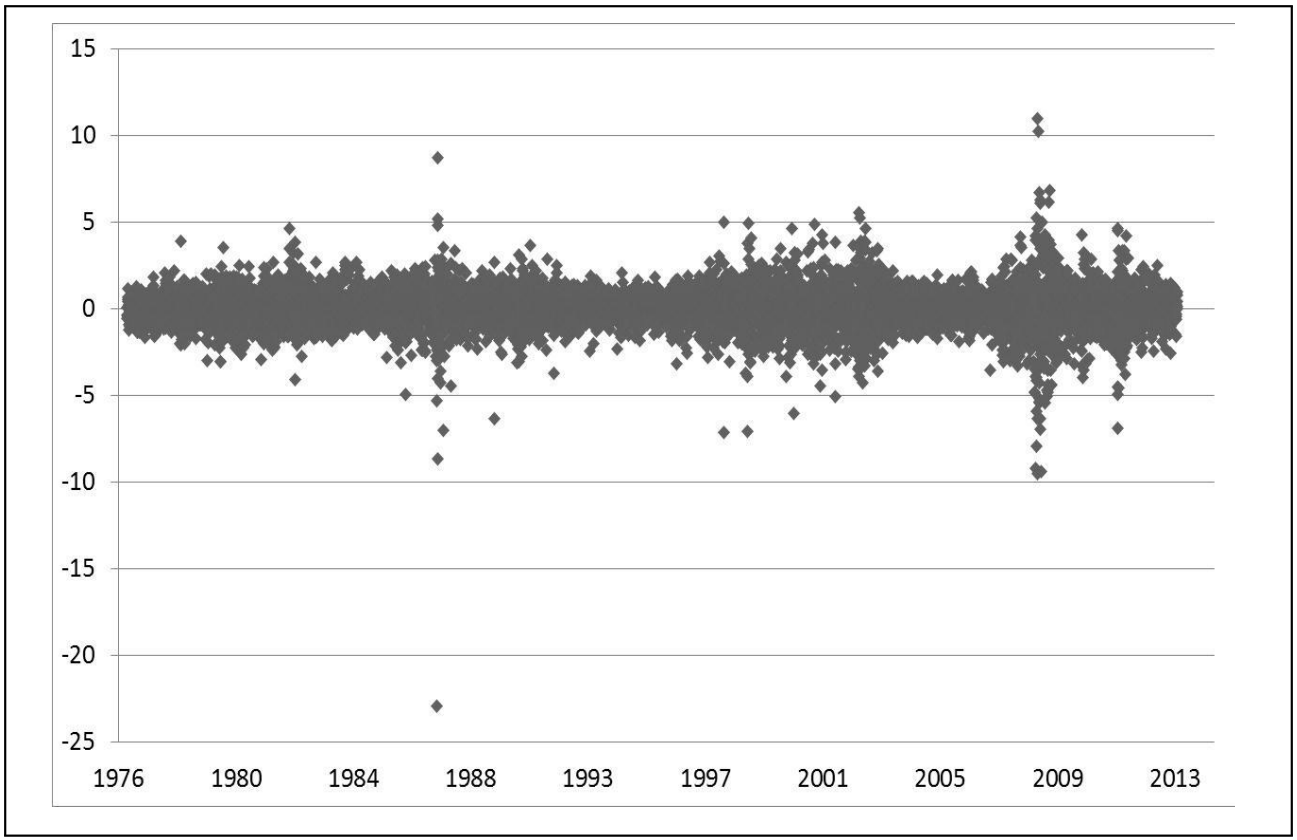
Rejection probabilities in favour of the Generalised (G) and Linear (L) pools using the Giacomini-White test for equal density forecast performance

Outline

Data and set up

- S&P 500 daily percent log returns data 3 January 1972 to 9 September 2013 extending Geweke & Amisano (2010, 2011) in their analysis of optimal linear pools.
- Following them we estimate GARCH(1,1), Student t-GARCH(1,1), Gaussian exponential GARCH(1,1) models *via* maximum likelihood (ML) and an SV model by Kim *et al* (1998) by Bayesian sampling methods using rolling samples of 1250 trading days (about five years).
- One-day-ahead recursive density forecasts for returns 15 December 1976 through to 9 September 2013 (out-of-sample $T = 9268$).
- Predictive densities formed by substituting the estimates for the unknown parameters.
- The two component densities are then combined using either a linear or generalised combination scheme.

S&P 500 daily %age log returns 15 Dec 1976 - 9 Sept 2013



Evaluation

- Fit generalised and linear combinations whole dataset, in-sample ($t = 1, ., T$), extending part of Geweke & Amisano (2011).
- G&A found LP outperformed any component.
- Can replicate G&A to their sample end in 2005, but inference sensitive to sample.
- Over longer sample including crisis LP at best matches best component.
- GP estimating p in range 2-10, thresholds chosen with a grid search width 0.5 in interval -2.5% to 2.5%.
- In all cases large and highly significant benefits to using GP *versus* LP.
- Pools using T-GARCH perform best, and also much better than T-GARCH alone.

Table 8 - in-sample

	Comps: 1: GARCH, 2: EGARCH, 3: SV, 4: TGARCH					
	1	2	3	4		
Comp.	-1.169	-1.187	-1.580	-1.183		
	1,2	1,3	1,4	2,3	2,4	3,4
GP	-0.288	-0.313	2.762	-0.284	2.762	2.762
LP	-1.169	-1.169	-1.169	-1.187	-1.183	-1.183
$\widehat{p}$	9	9	9	9	9	9
p-val	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,3	1,2,4	2,3,4	1,2,3,4		
GP	-0.288	2.762	2.762	2.572		
LP	-1.169	-1.169	-1.183	-1.169		
$\widehat{p}$	9	9	9	3		
p-val	0.000	0.000	0.000	0.000		

In-sample (1976-2013) mean log scores for GP, LP and component densities: $\widehat{p}$ CV estimator for p. H_0 p-value null LP best.

Table 8 - in-sample - weights on T-GARCH

- Interesting that 5 of the 6 GPs including T-GARCH yield identical scores.
- This is despite having different weights.
- In each case T-GARCH invariably obtains a high weight somewhere in the density, often approaching unity in the centre of the distribution.
- Despite the prominence of T-GARCH the generalised pools offer substantial improvements, unlike the linear pool.
- The improvement is in the tails, despite advantage T-GARCH should have with fat tails.
- The number of regions $p = 9$ in most cases, but where all four models are combined, $p = 3$ (a more parsimonious weighting system).

Table 8 - in-sample - weights on T-GARCH

	1: GARCH, 2: EGARCH, 3: SV, 4:TGARCH					
	1,4	2,4	3,4	1,2,4	2,3,4	1,2,3,4
Weight 1	0.397	0.500	0.524	0.333	0.339	0.704
Weight 2	1.000	0.993	1.000	1.000	1.000	1.000
Weight 3	0.993	0.839	0.998	0.987	0.945	
Weight 4	0.838	0.980	0.998	0.805	0.886	
Weight 5	1.000	1.000	1.000	1.000	1.000	
Weight 6	0.993	0.993	1.000	0.992	0.998	
Weight 7	0.999	0.981	1.000	0.997	1.000	
Weight 8	0.500	0.500	0.525	0.333	0.339	

Table 9 - out-of-sample performance

- Clear risk flexible piecewise functions fit well in-sample but forecast poorly due to parameter error (overfitting).
- Consider samples 2004-2009, 2004-2007 and 2007-2013.
- Only past data used for optimisation.
- For some sub-samples, eg 3 September 2004 - 9 September 2013, optimal linear pool eg containing T-GARCH cannot beat T-GARCH alone.
- But for all samples and cases considered generalised pool does beat components.
- Moreover for all combinations and samples considered GP considerably and (highly) significantly outperforms LP.

Table 9 - out-of-sample mean log scores and p-values

	1: GARCH, 2: EGARCH, 3: SV, 4:TGARCH					
	2,4	3,4	1,2,3	1,2,4	2,3,4	1,2,3,4
	3 Sept 2004: 9 Sept 2013					
G	0.690	0.724	0.808	0.723	0.735	-0.296
L	-0.810	-0.610	-0.548	-0.793	-0.558	-0.554
G&W	1.000	1.000	1.000	1.000	1.000	1.000
	3 Sept 2004: 31 Aug 2007					
G	0.861	0.861	0.910	0.862	0.861	0.406
L	-0.146	-0.193	-0.131	-0.143	-0.137	-0.135
G&W	1.000	1.000	1.000	1.000	1.000	1.000
	4 Sept 2007: 9 Sept 2013					
G	0.603	0.654	0.757	0.652	0.670	-0.655
L	-1.150	-0.823	-0.761	-1.126	-0.773	-0.769
G&W	1.000	1.000	1.000	1.000	1.000	1.000

Outline

Conclusions

- Well established that density combinations useful.
- Extends existing literature by letting the combination weights follow general schemes.
- Specifically, combination weights depend on the variable being forecast.
- Specifically, piecewise linear weight functions varying by region of the density.
- Examined theoretically with sieve estimation used to optimise the score of the generalised density combination.
- Monte Carlo experiments suggest
  - Powerful method, likely to deliver large improvements for modest numbers of threshholds.
  - Works best with large T but improves results in an exercise calibrated to macro inflation data as well.
- It works. Delivers very large and significant improvements relative to linear pool in a stock returns exercise.