Quantile aggregation of density forecasts

Fabio Busetti*

Bank of Italy

Revised Version. September 2015

Abstract

Quantile aggregation (or 'Vincentization') is a simple and intuitive way of combining probability distributions, originally proposed in Vincent (1912). In certain cases, such as under Gaussianity, the Vincentized distribution belongs to the same family as that of the individual distributions and it can be obtained by averaging the individual parameters. This paper compares the properties of quantile aggregation with those of the forecast combination schemes normally adopted in the econometric forecasting literature, based on linear or logarithmic averages of the individual densities. Analytical results and Monte Carlo experiments indicate that the properties of quantile aggregation are in between those of the linear and the logarithmic pool. Larger differences among the combination schemes occur when there are biases in the individual forecasts. In this case quantile aggregation seems overall preferable, in terms of both logarithmic score and density calibration. The practical usefulness of Vincentization is illustrated empirically in the context of linear forecasting models for Italian GDP and quantile predictions of euro area inflation.

JEL Classification: C53, E17.

Keywords: fan charts, macroeconomic forecasts, model combination.

*I thank Malte Knuppel, Juri Marcucci, Andrea Silvestrini and Ken Wallis for useful comments on a previous version of this paper. All errors are mine. The views expressed here are those of the author and do not necessarily reflect those of the Bank of Italy. Email address: fabio.busetti@bancaditalia.it

1 Introduction

Economic forecasts are increasingly reported as point estimates supplemented by confidence bands or selected quantiles of the predictive distributions, in order to provide measures of uncertainty and risks around the central outcome. Indeed a common practice for central banks is to present forecasts of inflation and output in the form of 'fan charts' that describe in probabilistic terms the evolution of these variables along the forecast horizon. The (subjective) assessment of the likelihood of alternative macroeconomic scenarios is accounted using skewed density forecasts, that reflect higher probability for events in either tail of the distribution. For example, the Bank of England publishes fan charts for inflation since 1996; see Britton et al. (1998).

The econometric literature on forecast evaluation has been progressively extended to density forecasts in order to gauge the relative performance of different prediction models in terms of their distributions; see, inter alia, Diebold et al. (1998), Corradi and Swanson (2003, 2006), Mitchell and Hall (2005), Amisano and Giacomini (2007). In parallel, the idea of forecast combination, initiated by the classical paper of Bates and Granger (1969), has been applied to predictive distributions. The basic tools have been borrowed from the statistics literature on aggregation of subjective distribution functions, where the the task is to form an 'opinion pool'; cf. Genest and Zidek (1996) for a review. Econometric studies have focussed on linear or logarithmic weighting of the individual densities, where the weights may be data-driven reflecting the past performance of different models. Some examples are Wallis (2005, 2011), Hall and Mitchell (2007), Mitchell and Wallis (2010), Geweke and Amisano (2011), Fawcett et al. (2013). Kascha and Ravazzolo (2010) provide an empirical comparison of the linear versus the logarithmic opinion pool of several forecasting models of inflation. A comprehensive review of recent developments in density forecasting is Hall and Mitchell (2009).

This paper considers combining forecast distributions by quantile aggregation (or 'Vincentization'). This simple and intuitive approach, that consists in averaging the quantiles of the individual distributions, was originally proposed in Vincent (1912). Ratcliff (1979) and Thomas and Ross (1980) show that in certain cases, such as under Gaussianity, the Vincentized distribution belongs to the same family as that of the individual distributions and it can be obtained by averaging the individual parameters. In a recent work Liechtendahl et al. (2013) provide analytical properties of quantile aggregation and the linear opinion pool in terms of calibration, sharpness and shape. In econometrics a related approach has been proposed in Granger et al. (1989), where individual quantiles are modelled separately and, for each of them, a

certain linear combination of the forecasts is taken.

In this paper the properties of quantile aggregation are compared with those of the linear and the logarithmic opinion pool from an econometric forecasting perspective. Analytical results and Monte Carlo experiments indicate that the properties of quantile aggregation are in between those of the linear and the logarithmic pool. It is found that larger differences among the combination schemes occur when there are biases in the individual forecasts. In this case quantile aggregation seems overall preferable, in terms of both logarithmic score and density calibration. The practical usefulness of Vincentization is illustrated empirically in the context of linear forecasting models for Italian GDP and quantile predictions of euro area inflation.

The paper proceeds as follows. Section 2 defines quantile aggregation vis a vis the linear and the logarithmic opinion pools and it presents some general features of the various aggregation methods, focussing on the Gaussian case. Section 3 sets several Monte Carlo experiments to evaluate the relative properties of the different forecast density combinations in the context of standard econometric models. The empirical illustrations regarding linear forecasting models for Italian GDP and quantile models for euro area inflation are given in section 4. Section 5 provides concluding remarks.

2 Quantile aggregation and other density forecast combinations

Let $f_{it}(y_t)$ be forecast densities for a scalar variable y_t and denote by $F_{it}(y_t)$ the corresponding cumulative distribution functions, for $i = 1, 2, ..., n$. For a set of non-negative weights ω_i such that $\sum_{i=1}^{n} \omega_i = 1$, the combined distribution defined by quantile aggregation (or 'Vincentization') is given by

$$\overline{F}^{-1}_{\text{vin},t}(\alpha) = \sum_{i=1}^{n} \omega_i F^{-1}_{it}(\alpha), \quad 0 < \alpha \leq 1, \tag{1}$$

where $F^{-1}_{it}(\alpha) = \inf\{y : \ F_{it}(y) \geq \alpha\}$ are the quantile functions of the individual forecast distributions. Quantile averaging was originally proposed in Vincent (1912), hence it is sometimes called 'Vincentization'.

Ratcliff (1979) and Thomas and Ross (1980) prove the following theorem on the properties of Vincentization for 'location-scale' families of distributions; see also Genest (1992).

Theorem. *Let n individual distributions $F_i(y)$ be of the form $F_i(y) = H((y - \lambda_i)/\gamma_i)$, where λ_i is a centering parameter, γ_i is the scale and H is*

some distribution function, $i = 1, ..., n$. *Then the Vincentized distribution is given by* $\overline{F}_{\text{vin}}(y) = H\left((y - \overline{\lambda})/\overline{\gamma}\right)$ *with* $\overline{\lambda} = \sum_{i=1}^{n} \omega_i \lambda_i$ *and* $\overline{\gamma} = \sum_{i=1}^{n} \omega_i \gamma_i$.

The Gaussian, Cauchy, exponential and logistic random variables are all 'location-scale' distributions. For these cases quantile averaging; (i) gives rise to a distribution belonging to the same family; (ii) the parameters of the combined distribution are obtained simply by averaging those of the individual ones. When not available in closed form, the Vincentized density can be obtained by numerical approximation as the derivative of the inverse of the quantile function.

The density combination methods typically adopted in the econometric forecasting literature are the 'linear opinion pool' and the 'logarithimic opinion pool'. The former, proposed by Stone (1961), is defined as

$$\overline{f}_{\text{lin},t}(y_t) = \sum_{i=1}^{n} \omega_i f_{it}(y_t);$$

the latter is given by

$$\overline{f}_{\text{log},t}(y_t) = \frac{\prod_{i=1}^{n} f_{it}(y_t)^{\omega_i}}{\int \prod_{i=1}^{n} f_{it}(y_t)^{\omega_i} dy}.$$

The linear opinion pool is thus a linear mixture distribution, which in general can be multi-modal even under Gaussianity of the individual forecasts. As in the case of Vincentization, the logarithmic opinion pool is closed under Gaussianity (more generally when the individual components belong to the regular exponential family; Faria and Mubwandarikwa, 2008).[1]

As an example, figure 1 shows the result of the three combination methods where the individual density functions are a N(0,1) and N(2,0.5). The linear pool gives rise to a bimodal distribution. The logarithmic pool is less dispersed and its location is closer to that of the individual distribution with lower variance; cf. section 3. As another example figure 2 shows the combinations of an exponential and a Weibull distribution, both with unit mean. Again the linear pool shows higher dispersion.

The advantages of forecast combinations are well understood for the case of point forecasts, where combinations are showed to work well in several empirical studies; cf. Timmermann (2006). Less studies have investigated

[1] One clear drawback of the logarithmic opinion pool is that it gives probability of zero to events that have zero probability under any of the individual distributions.

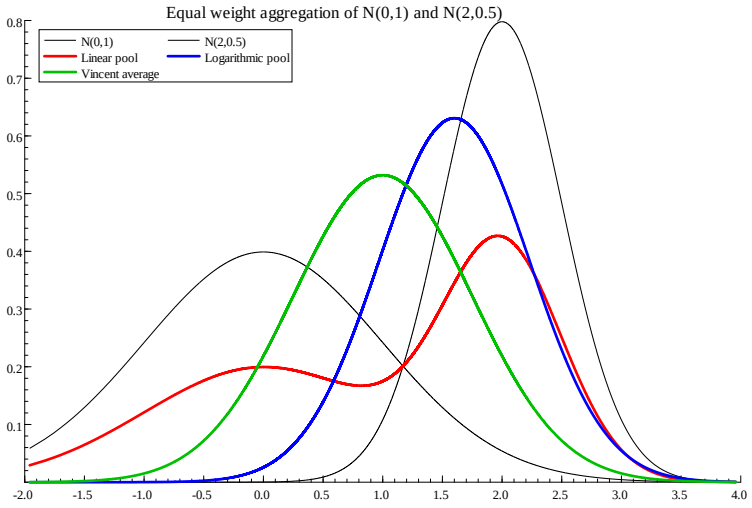

Figure 1: Aggregation of Gaussian densities

the advantages of combining predictive distributions. Importantly, Kascha and Ravazzolo (2010) show that a density forecast combination is at least as good as the worst model in terms of distributional accuracy, where the metrics is the average probability of observing the realized values (the 'log-score', defined in the subsection below).

2.1 The aggregation weights

The properties of forecast combinations to some extent depend on the weighting scheme adopted. Ideally, the weights should reflect the past performance of the different models and be time-varying, i.e. computed recursively at each point in time using all observations available. In practice equal weights forecasts are often adopted; empirically these are found difficult to beat. In the context of density forecasts, the empirical analysis in Kascha and Ravazzolo (2010) does not indicate any advantage of using time-varying over equal weights.

A simple metrics for setting weights is to use the prediction mean square error of different models ($PMSE_i$),

$$\omega_i = \frac{1/PMSE_i}{\sum_{i=1}^{n} 1/PMSE_i}.$$

These *inverse MSE weights* are widely used for point forecasts; they would be optimal if the forecasts were independent; see Bates and Granger (1969).

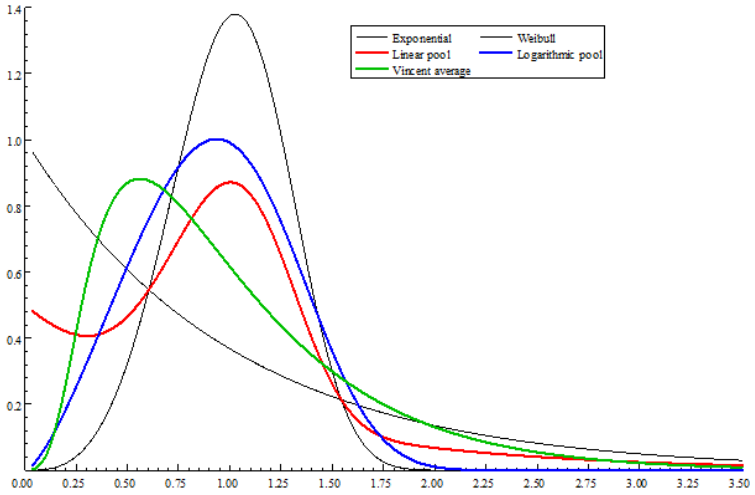

Figure 2: Equal weight aggregation of an exponential and a Weibull distribution.

In the context of density forecasts, models are often compared in terms of their average (log) predictive density, the so-called log-score, $S_i = \frac{1}{T} \log \sum_{t \in \Upsilon} f_{it}(y_t)$, where the average is over some sample Υ of length T. In particular, the density forecasts of model i is seen as a better approximation of the true distribution than those of model j if the (out-of sample) log score is higher, $S_i > S_j$; that is, on average, model i gives higher probability to the events that really occurred; see e.g. Mitchell and Hall (2005).[2] The *log-score weights* are defined as

$$\omega_i = \frac{\exp(S_i)}{\sum_{i=1}^{n} \exp(S_i)}.$$

In a Bayesian framework these weights are related to the models' posterior probabilities[3].

2.2 Properties of the combined distributions under Gaussianity

Without imposing any distributional assumption, Lichtendahl et al. (2013) show that the linear opinion pool and the Vincent average have the same

[2] Amisano and Giacomini (2007) gives a formal test of equal forecast performance based on the difference in the log score of the two models.

[3] Hall-Mitchell (2007) suggest using 'optimal log-score weights', defined as those that maximize the log-score of the combined distribution under the linear opinion pool.

mean and that the $\overline{F}_{\text{vin}}$'s even central moments are less than or equal to those of $\overline{F}_{\text{lin}}$. Hence the Vincentized distribution is less dispersed and has thinner tails than the linear opinion pool.

Under Gaussianity of the individual distributions, $N(\mu_i, \sigma_i^2)$, $i = 1, 2, ..., n$, simple formulas apply. In particular: (i) the Vincentized distribution is Gaussian with mean $\mu_{\text{vin}} = \mu_{\text{lin}} = \sum_i \omega_i \mu_i$ and variance $\sigma^2_{\text{vin}} = \left(\sum_i \omega_i \sigma_i\right)^2$; (ii) the logarithmic pool is Gaussian with mean $\mu_{\log} = \left(\sum_i \omega_i/\sigma_i^2\right)^{-1} \sum_i \mu_i \omega_i/\sigma_i^2$ and variance $\sigma^2_{\log} = \left(\sum_i \omega_i/\sigma_i^2\right)^{-1}$; (iii) $\sigma^2_{\log} \leq \sigma^2_{\text{vin}} \leq \sigma^2_{\text{lin}}$.

Thus the linear pool, that in general is non-Gaussian, is the most dispersed distribution, while the logarithmic pool is the less dispersed[4]. In terms of location, the logarithmic pool rescales the weights such that $\mu_{\log}$ is 'closer' to the mean of the individual distributions with smaller variance. These features can be clearly seen in the example of figure 1.

A further message implied in figure 1 is that larger differences among the combination schemes are more likely to arise if the individual distributions are not centered around the same mean. This issue is investigated in the following simple experiment where we combine two Gaussian distributions with different locations but same mean square. Let f_1 be a $N(0, 1)$ and f_2 a $N(b, 1 - b^2)$, where b is a 'bias'; f_1 is the true model. The goal is to compare the various equal weight combinations of the two distributions; as the true model is assumed to be f_1 the equal weights are also inverse MSE weights. The metrics of comparison is the distance of f_{lin}, $f_{\log}$, f_{vin} from the true distribution f_1, as a function of the magnitude of the bias term. The results are showed in figure 3 for the Kullback-Leibler divergence and the L_1-distance. The distances are graphed as ratios to the one of the f_2 distribution: clearly these ratios are lower than 1, as the combinations are closer to f_1 than f_2 is. As the bias rises the logarithmic pool becomes increasingly worse than the linear pool and the Vincentized distribution, both in terms of the Kullback-Leibler divergence and the L_1-distance. $\overline{F}_{\text{lin}}$ and $\overline{F}_{\text{vin}}$ tend to behave more similarly.

[4]The proof that $\sigma^2_{\log} \leq \sigma^2_{\text{vin}}$ follows from applying twice the weighted arithmetic and geometric mean inequality (AM-GM), $\sum_i \omega_i x_i \geq \prod_i x_i^{\omega_i}$, where ω_i are the weights that sum to 1. First, squaring both sides of AM-GM, with $x_i = \sigma_i$, yields $\sigma^2_{\text{vin}} = \left(\sum_i \omega_i \sigma_i\right)^2 \geq \prod_i x_i^{2\omega_i}$. Second, by applying AM-GM with $x_i = 1/\sigma_i^2$ one gets $\left(\sigma^2_{\log}\right)^{-1} = \sum_i \omega_i/\sigma_i^2 \geq \prod_i \left(1/\sigma_i^2\right)^{\omega_i} = \left(\prod_i \sigma_i^{2\omega_i}\right)^{-1}$. Multiplying out, the terms at the right hand side cancel out leaving $\sigma^2_{\text{vin}}/\sigma^2_{\log} \geq 1$. The proof that $\sigma^2_{\text{vin}} \leq \sigma^2_{\text{lin}}$ follows directly by the Jensen inequality; it is a special case of the more general result proved in Lichtendal et al. (2013).

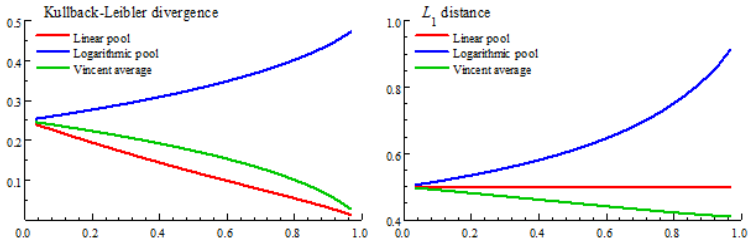

Figure 3: Distances from the true distribution plotted against the magnitude of the (square) bias.

3 Monte Carlo comparison of the forecast density combination methods

This section compares the density forecasts obtained with the different combination methods for simple time serie models and data generating processes. The first experiment is the same one considered in Mitchell and Wallis (2010), where the data are generated by an AR(2) process and the forecasting models to be combined contain only one lag of the dependent variable. In a second experiment the data process is an AR(1) process with GARCH disturbances, while the forecasts are generated ignoring either the autoregressing component or the time-varying conditional volatility.

The main metric over which forecast densities are compared is the KLIC divergence from the true distribution, $KLIC$=E[log $f_0(y)$ - logf(y)], $where f_0$ is the true density and f its forecast. Note that the relative performance based on KLIC divergence is the same as that obtained by computing the log-score. The calibration properties of the different forecast distributions are also evaluated. As in Mitchell and Wallis (2010), we report results based on 2000 Monte Carlo simulations and a sample size $T = 150$.

3.1 An AR(2) data generating process

We assume that the data are generated by the AR(2) process

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t, \qquad \varepsilon_t \sim NIID(0,\ \sigma_\varepsilon^2).$$

The 'ideal forecast'of y_t, distributed as $F_{0t} = N\left(a_1 y_{t-1} + a_2 y_{t-2},\ \sigma_\varepsilon^2\right)$, is first compared with two individual forecasts obtained from misspecified models where y_t is regressed on either y_{t-1} or y_{t-2} only. These are distributed as $F_{1t} = N\left(\rho_1 y_{t-1},\ \sigma_y^2(1-\rho_1^2)\right)$ and $F_{2t} = N\left(\rho_2 y_{t-2},\ \sigma_y^2(1-\rho_2^2)\right)$, where

$\sigma_y^2 = \sigma_\varepsilon^2/(1 - a_1\rho_1 - a_2\rho_2)$, $\rho_1 = a_1/(1 - a_2)$, $\rho_2 = a_1\rho_1 + a_2$.The individual forecasts are then aggregated with equal weights according to three schemes considered in this paper, yielding the combined distributions $F_{\text{lin},t}$, $F_{\text{log},t}$ and $F_{\text{vin},t}$ respectively.

The resulting density forecasts are compared against the ideal forecast in the table 1, where AR_1 and AR_2 denote the individual forecasts f_{1t} and f_{2t}, respectively. For each distribution we report the KLIC distance against the true density f_{0t} (the lower the better) and the percentage rejections of a test of correct calibration[5], run at the 5% significance level, denoted as $KS_{.05}$.

We consider three configurations of the autoregressive parameters of the data generating process: (1) $a_1 = 1.5$, $a_2 = -0.6$, (2) $a_1 = 0.15$, $a_2 = 0.2$, (3) $a_1 = -0.5$, $a_2 = 0.3$. The corresponding first and second order autocorrelation coefficients are: (1) $\rho_1 = 0.94$, $\rho_2 = 0.80$, (2) $\rho_1 = 0.19$, $\rho_2 = 0.23$, (3) $\rho_1 = -0.71$, $\rho_2 = 0.66$.

The first three rows of the table contain nearly the same numbers as those reported by Mitchell and Wallis (2010; table IV for KLIC and table II for $KS_{.05}$), to which we add the results for the logarithm opinion pool and the vincentization aggregation methods. In case (1) where data are very persistent ($\rho_1 = 0.94$), the AR_1 model achieves the lowest KLIC distance from the true distribution. Among the combined forecasts, the logarithmic opinion pool is the preferable aggregation scheme although it remains significantly worse than the AR_1. In case (2) and (3) the three aggregation methods delivers similar results (with a slightly inferior performance of the linear opinion pool), yielding a better outcome than the individual forecasts. In terms of calibration, the type of model misspecification considered in this experiment does not affect much the rejection frequencies of the KS test. Overall the linear pool appears less calibrated than the other two distributions; this may be partly related to its non-normality.

Table 1. Comparison of density forecasts for an AR(2) data generating process with unbiased forecasts.

[5]The Kolmogorov-Smirnov test of uniformity of the probability integral transform is used, with asymptotic critical values. Part of the (small) differences with respect to the figures reported in Mitchell and Wallis (2010) may be due to their use of finite sample critical values. However for the ideal forecast we obtain a rejection rate of the null hypothesis equal to 4.8%, hence very close to the nominal size of the test.

Forecast	Case (1)		Case (2)		Case (3)	
	$KLIC$	$KS_{.05}$	$KLIC$	$KS_{.05}$	$KLIC$	$KS_{.05}$
AR_1	22.5	1.7	2.1	9.3	4.8	11.8
AR_2	75.6	9.4	1.2	7.9	12.1	0.0
Lin. Pool	42.9	5.8	0.7	8.7	3.5	6.7
Log. Pool	36.3	4.9	0.7	8.9	1.9	3.9
Vincentization	49.5	6.2	0.6	9.0	2.2	3.7

In table 1 all forecasts are unbiased, while we have seen in section 2 that greater differences among the densities may occur when we allow for a bias. To this extent we also consider a forecast obtained from the true data generating process but evaluated under an asymmetric loss function of the 'linex' type, which delivers a (constant) forecast bias; see e.g. Christoffersen and Diebold (1997). The performance of the three aggregation schemes is evaluated when this biased forecast is combined with the AR_1 model. Table 2 reports the results for the parametrization labelled 'case (1)' of the AR(2) data generating process with $\sigma_\varepsilon^2 = 1$, for a forecast bias equal to 0.5, 1 and 2. We report the KLIC distances from the true distribution for both equal weight and inverse MSE weights combinations, denoted $KLIC_0$ e $KLIC_1$ respectively.

When the bias is relatively small, less or equal to 1, the performance of the unbiased AR_1 forecasting model is worse than that of any of the three combination schemes. Overall Vincentization appears to be the preferable aggregation scheme: it is significantly better when the bias is larger, while being not much different from the log opinion pool otherwise. Using inverse MSE weights (columns $KLIC_1$) may improve significantly the accuracy of the combined distributions, but it does not change the relative rankings of the three methods. In terms of calibration, the bias forecast strongly reject the KS test, affecting the rejection frequencies also of the combined densities (results are not reported but they are available upon request).

Table 2. Comparison of density forecasts for an AR(2) data generating process, with a bias individual forecast (asymmetric loss function).

Forecast	$bias = 0.5$		$bias = 1.0$		$bias = 2.0$	
	$KLIC_0$	$KLIC_1$	$KLIC_0$	$KLIC_1$	$KLIC_0$	$KLIC_1$
AR_1	22.5	22.5	22.5	22.5	22.5	22.5
Asymmetric loss	13.0	13.0	51.1	51.1	202.0	202.0
Lin. Pool	10.9	10.5	22.1	21.1	51.9	31.9
Log. Pool	8.6	8.4	20.3	18.4	66.6	28.1
Vincentization	9.6	9.1	17.1	16.3	46.8	22.9

3.2 Time-varying volatility

Here we assume that the data are generated by the AR(1)-GARCH(1,1) process

$$\begin{aligned} y_t &= \rho y_{t-1} + \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim NIID(0,1), \\ \sigma_t^2 &= \gamma + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2. \end{aligned}$$

We compare the properties of combining, with equal weights, the forecast densities of two misspecified models: an AR(1) with constant conditional variance and a GARCH(1,1) with constant conditional mean, denoted as f_{1t} and f_{2t} respectively. For each individual distribution and forecast combination Table 3 reports the KLIC distance from the true density f_{0t} and the percentage rejections of the KS test of calibration at the 5% significance level. We report results only for a typical parametrization of the GARCH ($\alpha = 0.04$, $\beta = 0.95$) and for two cases of high and low persistence, $\rho = 0.75$ and $\rho = 0.25$ respectively.

For the case of highly persistent data ($\rho = 0.75$) the AR(1) model with constant variance provides the relatively more accurate approximation of the density forecast to the true distribution. Among the combined distributions, the logarithmic pool has comparably better properties than the linear pool and the Vincent average, the latter two behaving similarly. On the other hand, when the data persistence is lower all combined distributions behave comparably better than each individual forecasts, with the Vincent average being only slightly superior.

In terms of calibration, when the data are persistent ($\rho = 0.75$) the GARCH model with constant conditional mean rejects the null hypothesis of the KS test with the highest frequency; the combined densities are also less calibrated than the misspecified AR(1) model. On the other hand, for $\rho = 0.25$ the AR(1) model is the worst calibrated, while the Vincent average is the model that works relatively better.

Table 3. Comparison of density forecasts for an AR(1)-GARCH(1,1) data generating process

Forecast	$\rho = 0.75$		$\rho = 0.25$	
	$KLIC$	$KS_{.05}$	$KLIC$	$KS_{.05}$
AR	2.6	19.1	2.6	18.5
GARCH	31.4	50.8	2.9	11.7
Lin. Pool	11.2	36.4	1.1	12.2
Log. Pool	5.8	27.9	1.2	12.9
Vincentization	12.9	37.1	0.9	1.7

4 Empirical illustrations

4.1 Linear forecasting models for Italian GDP

As a first example we consider combining forecasts of Italian GDP from two simple linear models: (1) an autoregression of order 4; (2) a three variables VARX model for GDP, inflation and long-term interest rate, with two lags of the endogenous variables and additional exogenous regressors for foreign demand, oil prices and the short term interest rate.[6] The models are estimated on quarterly data for the period 1986.1-2006.4. The out-of-sample evaluation period is 2007.1-2014.4: it includes the 'Great Recession' of 2008-09 and the later 'Sovereign Debt Crisis'.

Figure 4 provides in-sample and out-of-sample point predictions of the two models for the four-step ahead percentage growth rate of GDP, $y_{t+4|t} = 100(\log GDP_{t+4} - \log GDP_t)$. The troughs of the two (out-of-sample) crises are clear outliers with respect to the forecast distributions based on the in-sample fit.[7]

Turning to distributions, as an example Figure 5 shows the forecast densities for the first three years of the out-of-sample period. As expected, all combined distributions tend to look more or less the same when there are not large differences in the point predictions of the AR and VARX models; otherwise the linear prediction pool displays asymmetry and bimodality, whereas the logarithmic opinion pool and the Vincent average are more similar to each other.

Table 4 reports the out-of-sample fit of the 5 forecast distributions, measured through the log-score and the MSE of the point estimates. The p-values of the test of Amisano and Giacomini (2007) of the null hypothesis of equality of distributions are also reported. A one-sided version of the test is considered where the alternative hypothesis is that each of the combined distributions

[6] The VARX model can be viewed as a rough approximation of the macroeconometric models typically used for producing conditional forecasts, under specific assumptions on the future paths of foreign variables and of the monetary policy rate. If the assumptions turn out to be more or less correct, then the conditional forecasts can be much more accurate than the unconditional ones. In our estimates the in-sample variance of the four-step ahead prediction errors of percentage GDP growth is 1.9 for the AR(4) model and 0.8 for the VARX.

[7] Model misspecification can explain the overprediction of the VARX during the crises. In particular, the VARX does not take into account the effects of: (i) the decline of confidence and the increased uncertainty of households and firms, (ii) the fiscal consolidation measures enacted during the european sovereign debt crisis. Busetti and Cova (2013) estimate that these factors alone contributed for more than 3 percentage points to the Italian GDP fall in 2012-13.

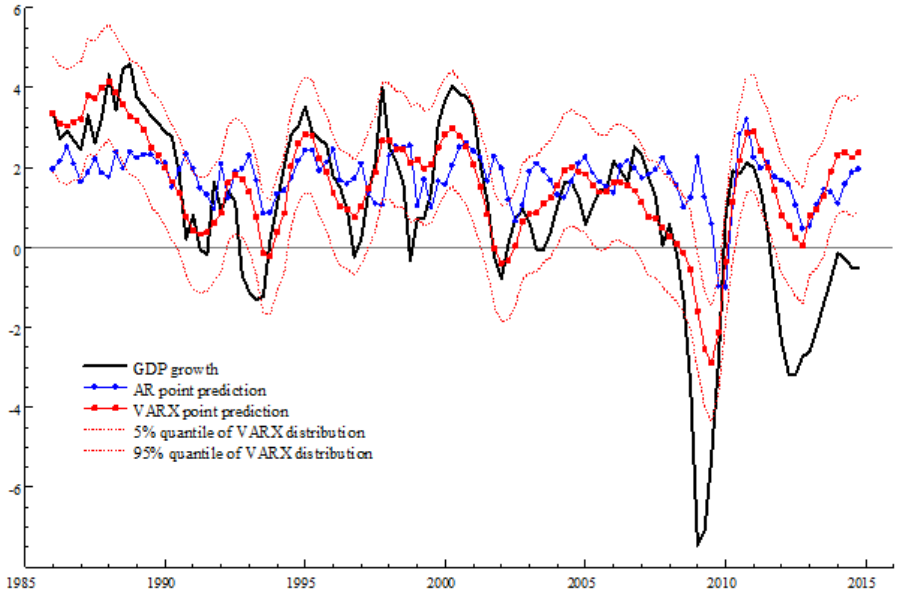

Figure 4: GDP growth and 4-step ahead predictions with the AR and VARX model; the estimation sample is 1986.1-2006.4

in the columns of the table fits the data better than the distribution in the rows.

In terms of point forecasts, the VARX displays the smallest MSE, even lower than that of forecast combinations. The VARX is however the worst in terms of density forecasts, as the variance of its forecast distribution is probably too small for the out-of-sample data. The Vincentized distribution achieves the highest log-score, followed by the linear pool; the logarithmic pool performs worst, despite a smaller MSE (which is computed comparing the realized value with the mean of the logarithmic pool).

The statistical tests of superior forecast distribution tell a similar story. All combined forecasts are significantly better than the VARX. Among the combination schemes, both the linear pool and Vincentization beat the logarithmic pool at the 10% significance; the test of a superior froecasting performance of Vincentization over the linear pool has a p-value of 0.17.

Table 4. Out-of-sample fit and pairwise test of superior performance for the forecasts of Italian GDP.

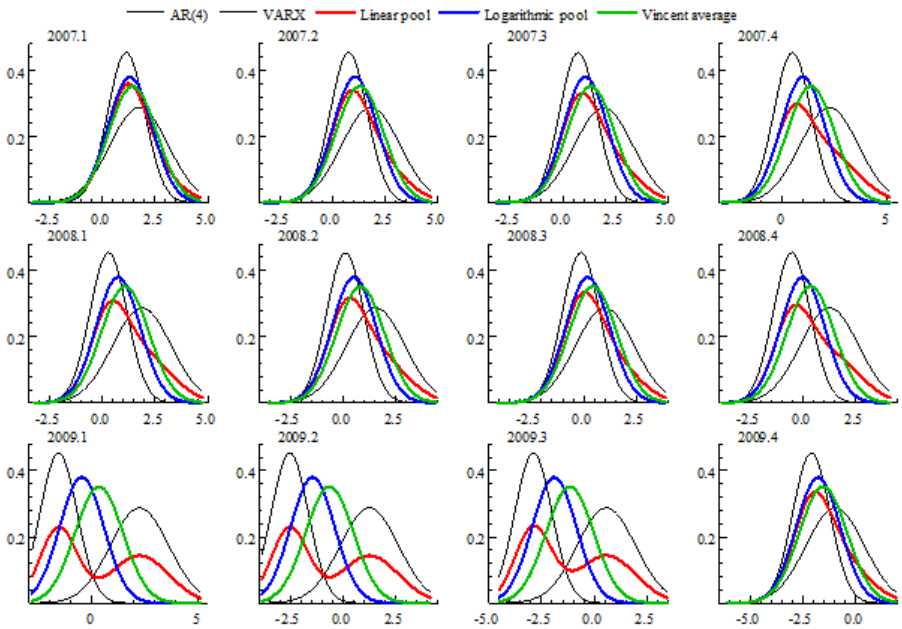

Figure 5: Forecast distributions for Italian GDP growth; 2007.1-2009.4

Forecasts	*MSE*	*Log-Score*	*Test of superior performance of*		
			Lin. Pool	Log. Pool	Vincentiz.
AR(4)	11.6	-4.28	.21	.49	.15
VARX	6.1	-4.74	.02	.03	.04
Lin. Pool	8.4	-4.03	-	.99	.17
Log. Pool	7.3	-4.27	.01	-	.09
Vincentization	8.4	-3.36	.83	.91	-

4.2 Quantile forecasts of euro area inflation

As a second, less standard, example we consider forecasting the distribution of euro area inflation by means of dynamic quantile regressions, following Busetti et al. (2015). The empirical distribution obtained by the quantile forecasts is then combined with the one of a times series model with changing volatility, which is generally viewed as a good benchmark for modelling inflation (Stock and Watson, 2007).

Let $Q_\alpha(z)$ denote the quantile of order α of a variable z, for $0 < \alpha < 1$. Motivated by Phillips curve arguments, we estimate the following prediction

model of conditional quantiles,

$$Q_\alpha(\pi_{t+1}) = \beta_0(\alpha) + \beta_1(\alpha)\pi_t + \beta_2(\alpha)y_t + \beta_3(4)oil_t + \beta_4 ex_t, \qquad (2)$$

where π_t is inflation (the year-on-year change of the logarithm of the Harmonized Index of Consumer Prices), y_t the output gap, oil_t the change in oil prices in euro, ex_t the change of the nominal effective exchange rate of the euro. The model is estimated on quarterly data, in pseudo-real time with starting period 1990Q1,[8] for $\alpha = .05, .10, .15, ..., .85, .90, .95$. The covariates y_t, oil_t and ex_t are meant to capture the effects of, respectively, demand pressure, commodity prices and the exchange rate. The model is dynamic in the sense that the inflation quantiles depend on the past level of inflation, in addition to the covariates.

The out-of-sample forecast performance of the dynamic quantile regressions is evaluated over the 10 year period 2005Q1-2014Q4. The benchmark model for comparison is an AR(2)-GARCH(1,1). A Vincent average of the approaches is then considered. The main metric for evaluation is the out-of-sample 'check' loss function $L(\alpha)$,

$$L(\alpha) = \sum_{\pi_t \geq \widehat{Q}_{\alpha,t}} \alpha \left| \pi_t - \widehat{Q}_{\alpha,t} \right| + \sum_{\pi_t < \widehat{Q}_{\alpha,t}} (1-\alpha) \left| \pi_t - \widehat{Q}_{\alpha,t} \right|, \qquad \alpha = .05, ..., .95,$$

where $\widehat{Q}_{\alpha,t}$ is the (real time) forecast of the α-order conditional quantile of inflation. A summary measure of distributional accuracy is the weighted quantile scoring function (WQS), as in Gneiting and Ranjan (2011), defined as

$$WQS = \int_0^1 L(\alpha)\omega(\alpha)d\alpha,$$

where $\omega(a)$ are the weights. The *lower* WQS the better is the distributional forecast.

Table 5 reports the WQS of the forecasts for different weighting scheme, that give different importance to various parts of the distribution; the case of uniform weights corresponds to the simple average of $L(\alpha)$ across all quantiles.

According to this metric, the quantile regression is generally better than the AR(2)-GARCH(1,1) but both forecasts are less accurate than that resulting from quantile aggregation. Averaging seems to be especially useful for the

[8] The data prior to the inception of the euro in 1999 have been reconstructed based on the initial composition of the euro area.

tails of the distribution, whereas no advantage over the quantile regression approach is found in the center.

Table 5. Distributional Accuracy based on the WQS criterion.

Weighted Quantile Score	$\omega(\alpha)$	Q-REG	GARCH	Vincent
Uniform	1	5.23	5.46	5.12
Center	$\alpha(1-\alpha)$	1.00	1.07	1.00
Left tail	$(1-\alpha)^2$	1.64	1.64	1.59
Right tail	α^2	1.58	1.69	1.53
Both tails	$(1-2\alpha)^2$	0.58	0.61	0.54

An example of the types of distributions implied by the different models is provided in Figure 6, that contains forecast densities for the 12 quarter period 2008.1-2010.4. For quantile regression and Vincentization the densities have been approximated by fitting a generalized log-logistic distribution to the quantiles, so to allow for asymmetry. The density of the AR(2)-GARCH(1,1) forecast is Gaussian.

In general, the quantile regression forecast is less dispersed as it can account better of the large movements in the oil prices and the exchange rate occuring during that period. It can also exhibit a high degree of asymmetry (possibly magnified by the smooth approximation presented in the figure). The Vincent average seems a good compromise between the asymmetry of quantile regression and the larger dispersion of the conditionally Gaussian time series model.

5 Concluding remarks

This paper has compared quantile aggregation against the linear and the logarithmic opinion pool as methods for combining density forecasts. Overall the properties of quantile aggregation are in between those of the linear and the logarithmic pool. It maintains the location properties of the linear pool but it is less dispersed and more likely to yield a unimodal density. According to our results, quantile averaging appears particularly useful for combining forecast distributions with significant differences in location. Empirical illustrations on linear and non-linear models for macroeconomic data have been provided showing the practical usefulness of quantile aggregation.

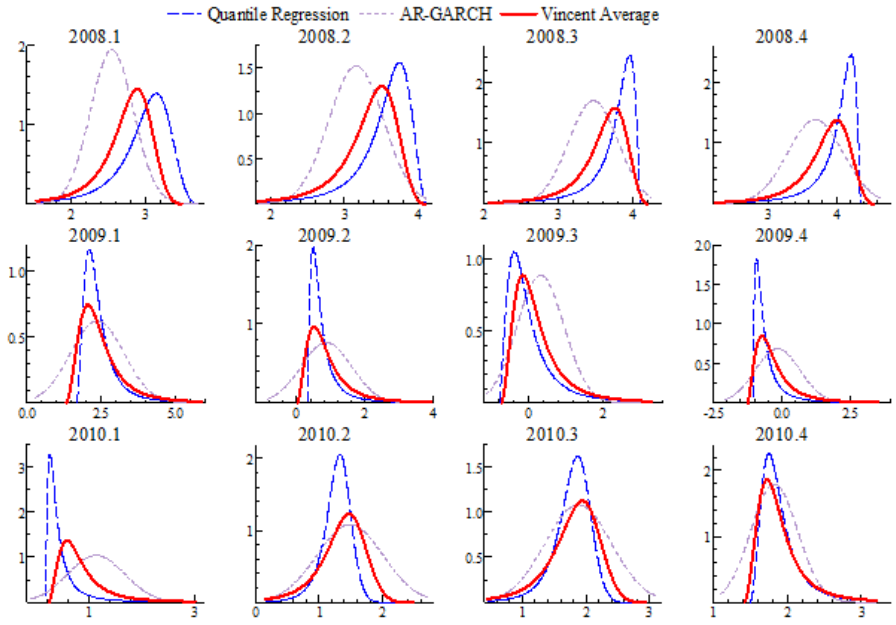

Figure 6: Forecast distributions for euro area inflation; 2008.1-2010.4

References

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

[2] Bates, J.M. and C.W.J. Granger (1969), Combination of forecasts, *Operational Research Quarterly*, 20, 451–468.

[3] Britton, E., Fisher, P. and J. Whitley (1998), "The Inflation Report Projections: Understanding the Fan Chart", *Bank of England Quarterly Bulletin*, 38, 30-37.

[4] Busetti, F. and J. Marcucci (2013), Comparing forecast accuracy: a Monte Carlo investigation, *International Journal of Forecasting*, 29, 13-27.

[5] Busetti, F., Caivano, M. and L. Rodano (2015), On the conditional distribution of euro area inflation forecast, Bank of Italy Working Papers, n. 1027.

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

[7] Corradi V. and N.R. Swanson (2003), Bootstrap conditional distribution tests in the presence of dynamic misspecification, *Journal of Econometrics*, 133, 779–806.

[8] Corradi V. and N.R. Swanson (2006), Predictive density evaluation, In *Handbook of Economic Forecasting*, Elliot G, Granger CWJ, Timmermann A (eds). Elsevier: Amsterdam; 197–284.

[9] Diebold, F.X, Gunther T. and A.S. Tay (1998), Evaluating density forecasts with applications to finance and management, *International Economic Review*, 39, 863–883.

[10] Faria, A. E. and E. Mubwandarikwa (2008), The geometric combination of Bayesian forecasting models, *Journal of Forecasting*, 27, 519–35.

[11] Fawcett, N., Kapetanios, G., Mitchell, J. and S. Price (2013), Generalised density forecast combinations, Bank of England Working Paper.

[12] Genest, C. (!992), Vincentization revisited, *The Annals of Statistics*, 20, 1137-1142.

[13] Genest, C. and J.V. Zidek (1986), Combining probability distributions: a critique and an annotated bibliography, *Statistical Science*, 1, 114-148.

[14] Geweke, J. and G. Amisano (2011), Optimal prediction pools, *Journal of Econometrics*, 164, 130-141.

[15] Gneiting, T. and R. Ranjan (2011), Comparing density forecasts using threshold- and quantile-weighted scoring rules, *Journal of Business & Economic Statistics*, 29, 411-422.

[16] Granger, C.W.J., H. White and M. Kamstra (1989), Interval forecasting: an analysis based upon ARCH-quantile estimators, *Journal of Econometrics*, 40, 87-96.

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

[18] Hall, S.G. and J. Mitchell (2009), Recent developments in density forecasting, In *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*, Mills TC, Patterson K (eds), MacMillan.

[19] Hendry, D.F. and M.P. Clements (2004), Pooling of forecasts, *Econometrics Journal*, 7, 1–31.

[20] Jore, A.S., Mitchell, J. and S.P. Vahey (2010), Combining forecast densities from VARs with uncertain instabilities, *Journal of Applied Econometrics*, 25, 621-634.

[21] Kascha, C. and F. Ravazzolo (2010), Combining inflation density forecasts, *Journal of Forecasting*, 29, 231-250.

[22] Lichtendahl, K.C., Grushka-Cockayne, Y. and R.L. Winkler (2013), Is It Better to Average Probabilities or Quantiles?, *Management Science*, 59, 1594-1611.

[23] Mitchell, J. and S.G. Hall (2005), Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESER 'fan' charts of infl ation, *Oxford Bulletin of Economics and Statistics*, 67, 995–1033.

[24] Mitchell, J. and K.F. Wallis (2010), Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness, *Journal of Applied Econometrics*.

[25] Ratcliff, R. (1979), Group reaction time distributions and an analysis of distribution statistics, *Psychological Bulletin*, 86, 446-461.

[26] Stock, J.H. and M.W. Watson M W (2007), Why has U.S. in‡ation become harder to forecast?, *Journal of Money, Credit and Banking*, 39, 3–33.

[27] Stone, M. (1961), The opinion pool, *Annals of Mathematical Statistics*, 32, 1339–1342.

[28] Thomas, E.A.C. and B.H. Ross (1980), On appropriate procedures for combining probability distributions within the same family, *Journal of Mathematical Psychology*, 21, 136-152.

[29] Timmermann A. (2006), Forecast combinations, In *Handbook of Economic Forecasting*, Elliot G, Granger CWJ, Timmermann A (eds). Elsevier: Amsterdam; 135–196.

[30] Vincent, S.B. (1912), The function of the viborissae in the behavior of the white rat. *Behavioral Monographs*, 1.

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

[32] Wallis, K.F. (2011), Combining forecasts – forty years later, *Applied Financial Economics*, 21, 33–41

[33] Winkler, R.L. (1968), The consensus of subjective probability distributions, *Management Science*, 15, B61–B75.