Combining Density Forecasts using Bayesian Opinion Pools

Parush Arora[*] [†]

JOB MARKET PAPER

Abstract

The paper considers the efficient estimation of opinion pools in the Bayesian paradigm and extends their application to cases where the number of competing models exceeds the number of observations. An appropriate Bayesian formulation and estimation algorithm is proposed, allowing for shrinkage of weights towards any possible combination and thus applicable to problems related to model averaging and model selection. Results from a simulation study reveal that the proposed Bayesian opinion pool methodology improves prediction accuracy and stability in weights compared to opinion pools estimated using scoring rules. An application involving the Survey of Professional Forecasters demonstrates that the Bayesian opinion pool's inflation forecast competes well with the inflation forecast obtained from the simple opinion pool (published by the Federal Bank of Philadelphia). The application showcases the usefulness of the Bayesian solution in situations where optimization-based opinion pools fail.

Keywords: *Inflation Expectation, Model Averaging, Predictive Density, Scoring Rule.*

JEL: C11, C15, C53, E17, E37

[*]I am grateful to Ivan Jeliazkov, Fabio Milani and Yingying Lee for their invaluable guidance and encouragement. All errors are my own. The latest version of the paper can be found on my website

[†]Department of Economics, University of California, Irvine, CA 92697. email: parusha@uci.edu, website: https://www.parush-arora.com/

1 Introduction and Motivation

Every forecaster's take on a problem is reflected in how they formulate, specify and estimate a predictive model. All these decisions are predicated on how the forecaster perceives and incorporates uncertainty (Steel (2020)). As a result, several competing predictive models can emerge for a given random variable. For a researcher, an intuitive way to utilize all this information is by aggregating all the predictive densities (See Hoeting et al. (1999) for Bayesian Model averaging, Wang et al. (2009) for frequentist model averaging, Moral-Benito (2015) for model averaging in economics, Gneiting and Ranjan (2013) for predictive model aggregation and Clyde and George (2004) for model uncertainty). This paper focuses on linear opinion pool (Stone (1961), Bacharach (1974)), a simple and widely used method for model aggregation, and explores its utility for time series forecasting applications under the Bayesian foundation.

To set up the framework, let y_t be a random variable and $Y_T = \{y_1, y_2, \ldots, y_T\}$ be a sequence of ordered random variables up to time T. Let M_k be the model estimated by forecaster k and $p(y_{T+1}|Y_T, M_k)$ be the predictive density of y_{T+1} associated with M_k, where $k = 1, 2, \ldots, K$. The aggregate predictive density, $p(y_{T+1}|Y_T)$, under the linear opinion pool framework is obtained as

$$p(y_{T+1}|Y_T) = \sum_{k=1}^{K} w_{k,T} p(y_{T+1}|Y_T, M_k), \tag{1.1}$$

where $w_{k,T}$ is the weight allotted to M_k with the time subscript implying the use of information up to time T. It also means that the weights are updated recursively once y_{t+1} is realized. The weights are estimated with respect to the constraints $\sum_{k=1}^{K} w_k = 1$ and $w_k \geq 0 \ \forall \ k = 1, 2, \ldots, K$, which ensures that Eq. 1.1 is an appropriate probability density.

Researchers have estimated Eq. 1.1 by optimising different objective functions (further referred to as traditional opinion pools or TOP). There are two issues with the approach. First, estimating unique weights is infeasible if the number of predictive models exceeds the number of observations (micronumerosity). A non-negative degree of freedom is a necessary condition for any optimization problem. Often, micronumerosity (or near micronumerosity) becomes a binding constraint,

especially in time series forecasting, where the frequency of observations limits the data length. This restricts the researcher from considering fewer models in the final analysis or using a shorter rolling window for training the opinion pools. Second, the weights under TOP have a high variance when the sample size is small. Since weights based on past information are used to obtain the predictive density concerning the future variable, instability in these weights over time means that the opinion pools are responding to noise in the data. In these cases, weights associated with the expert's prediction react to their immediate past forecasting accuracy and do not consider their consistency as the data is limited. This could be one of the reasons why equal weights perform competitively with optimized weights (Hendry and Clements (2004) and Wallis (2005)), as equal weights provide insurance against bad forecasts and optimized weights can overfit the data.

The paper proposes to estimate the opinion pools using the proper Bayesian formulation and hence calls it the Bayesian Opinion Pool (BOP). This approach resolves the two issues discussed. First, the Bayesian framework allows the opinion pool to be estimated when the number of forecasting densities exceeds the number of observations. The proposed algorithm is effective even when dealing with a high number of forecasters since the whole vector of weights is sampled in a single block, leading to computational efficiency. Second, the BOP's weights are less volatile under the small sample setting. The BOP utilizes the Dirichlet prior which allows the opinion pool's weight to shrink towards any possible combination including one extreme of allotting equal weights to all the models to another where all the weights are allotted to the best model. This makes BOP useful for applications related to both model averaging and model selection. Due to the flexibility, the prior can introduce stability in the weights under the small sample settings, by shrinking them towards equal weights and allowing deviations only if enough evidence is available in the data. The stability in the weights over time leads to improvement in prediction accuracy since the shrinkage avoids overfitting. The simulation study (Section 4) found evidence that the BOP is stable under a small sample setting and is highly competitive with the TOP (the five proper scoring rules considered are log, quadratic, spherical, continuous ranked probability score (CRPS), and the first two moments score (FTMS)). Finally, the paper uses BOP in an application involving

the survey of professional forecasters (SPF) where traditional optimization-based opinion pools fail due to micronumerosity. The aggregated predictive density for the inflation rate is estimated and compared with the equal weights strategy published by the Federal Bank of Philadelphia. The paper finds evidence for lower mean square prediction error associated with inflation estimated using the BOP.

Even though the current framework uses Bayesian formulation, that does not restrict the experts from using only Bayesian models. In the Bayesian setting, the predictive density $p(y_{T+1}|Y_T, M_k)$ can be written as

$$p(y_{T+1}|Y_T, M_k) = \int p(y_{T+1}|\theta_k, Y_T, M_k)p(\theta_k|Y_T, M_k)d\theta_k, \tag{1.2}$$

where θ_k be the set of parameters used to specify M_k, $p(\theta_k|Y_T, M_k)$ is the posterior distribution of θ_k and $p(y_{T+1}|\theta_k, Y_T, M_k)$ is the likelihood function associated with M_k evaluated at the value y_{T+1}. The parameter θ_k has been integrated out, so it does not appear in the Eq. 1.2. In the likelihood-based perspective, the predictive density can take the form

$$p(y_{T+1}|Y_T, M_k) = p(y_{T+1}|\theta_k^{MLE}, Y_T, M_k). \tag{1.3}$$

The density is conditioned on $\theta_k = \theta_k^{MLE}$, the maximum likelihood estimator, on the right side of Eq. 1.3, and thus, θ_k got absorbed into M_k on the left side of Eq. 1.3.

Extensive research has been done involving aggregation of the predictive densities. Mitchell and Hall (2005) combined density forecasts using Kullback–Leibler information criterion. Billio et al. (2013) used state space modelling to aggregate predictive densities and used Bayesian formulation to estimate time-varying weights. Busetti (2017) discussed quantile aggregation of predictive densities. Bassetti et al. (2018) used the Bayesian method to estimate the beta transformation of the opinion pool. McAlinn and West (2019) develop a novel class of dynamic latent factor models for time series forecast synthesis called Bayesian predictive synthesis which encompasses several existing forecast pooling methods.

For forecasting applications, an appropriate class of objective functions to estimate opinion pools are scoring rules. A scoring rule is a function that assigns a score to a probabilistic distribution based on how well it performs in predicting the realized event. Scoring rules can be judged based on *ex-ante* and *ex-post* properties (Winkler et al. (1996)). A scoring rule is called proper (an *ex-ante* property) when it disincentivizes the forecaster from revealing the probability distribution different from their true belief. *Ex-post* properties are concerned with how the scoring rule evaluates the performance of a probabilistic distribution. Gneiting and Raftery (2007) covered a thorough discussion on proper scoring rules and their theoretical properties. Bates and Granger (1969) optimized weights in Eq. 1.1 by minimizing the variance. Geweke and Amisano (2011, 2012) optimized weights using the log score and showed its usefulness in predicting stock index data. Degroot and Mortera (1991) estimated optimal weights by minimizing the expected quadratic score under the Bayesian framework. Opschoor et al. (2017) compared opinion pools optimized from censored likelihood score (CLS), CRPS, and log score on stock market indices data and found that CLS performed the best, whereas the log score performed the worst. The properties of opinion pools vary based on the scoring rule used to estimate them.

This paper contributes to the renewed interest in survey-based measures of inflation expectations. Coibion et al. (2018) referred to SPF extensively and argued for improved models that rely on variables with expectations. Inflation forecasts are integral to many macroeconomic models as they are used as an estimator for inflation expectations. For example, the augmented Phillips curve under aggregate price formation captures the relation where the expectations of future inflation partly drive the current inflation (Phelps (1967), Friedman (1968)). In business cycle analysis, the efficacy of a real shock depends on how much future inflation is anticipated (Kydland and Prescott (1982), Long Jr and Plosser (1983)). Under the rational expectations hypothesis, only unexpected changes in inflation lead to a change in real macro variables (Muth (1961)). The new Keynesian theory of price dynamics is based on inflation driven by its own expectations (Ball et al. (1988)).

The Federal Reserve Bank of Philadelphia collects inflation predictive densities from several forecasters and publishes both at the individual and aggregate levels. They weigh all the predictive

densities equally to calculate the aggregate level density, thus obtaining the simple opinion pool (SOP). The issue with that approach is that the equal weights do not extract sufficient information from the density of the forecasters who have been more accurate than others in the past. Providing unoptimized weights loses valuable information and leads to an inefficient estimator of predictive density.

The inflation forecast obtained through the BOP at various levels of shrinkage competes well with the Federal Reserve Bank of Philadelphia's published SOP. The average density allotted to realized inflation is higher, and the mean square prediction error associated with the estimated expected value is lower for the BOP than the SOP. These issues could have been tackled by using TOP, but due to the low data frequency, it is infeasible to estimate weights by optimizing an objective function for any training window possible.

Section 2 gives a brief overview of traditional opinion pools. Section 3 introduces the Bayesian opinion pool. Section 4 presents the simulation study where the performance of TOP and BOP are investigated in several settings. Section 5 covers the macroeconomic application involving the SPF data. Section 6 concludes the paper.

2 Proper Scoring Rules for Traditional Opinion Pools

This section summarizes the asymptotic properties of TOP and how it is used to estimate opinion pools. For a parametric probability distribution $p(Y_T, \theta)$, let θ_0 be the true vector of parameters. Let any proper scoring rule be presented as $S(\cdot)$. Gneiting and Raftery (2007) showed that asymptotically

$$arg \ \max_{\theta} \frac{1}{T} \sum_{t=1}^{T} S(p(Y_t, \theta)) \longrightarrow \theta_0 \qquad as \qquad T \longrightarrow \infty. \tag{2.1}$$

Suppose the constraints on weights in Eq. 1.1 are satisfied. It implies that the opinion pool satisfies the conditions of an appropriate probability distribution and can be presented as $p(Y_T, w_T)$, where

$w_T = \{w_{1T}, w_{2T}, \ldots, w_{KT}\}$ be the parameter of interest. Then asymptotically,

$$arg \ \max_{w_T} \frac{1}{T} \sum_{t=1}^{T} S(p(Y_t, w_T)) \longrightarrow w_0 \qquad as \qquad T \longrightarrow \infty. \tag{2.2}$$

Let M_0 be the true model or DGP. Bernardo and Smith (2000) defined three scenarios possible. First is when M_0 is identified and available in the model list (M-close case). In this scenario, opinion pools will converge to M_0. The weight vector $w_0 = \{1, 0, \ldots, 0\}'$ where the weightage of 1 is allotted to M_0 and 0 to other models. The second case is when M_0 is available, but the researcher decides to intentionally leave it out of the model set taken into consideration (M-complete case). The third case, which is the most applicable and is considered in this paper, is when M_0 is not part of the model list (M-open case). In this case, w_T will converge to some weight vector $w_0 = w^*$, which is related to the properties of the metric implied by the scoring function.

Elliott et al. (2016) argued that there is no natural choice for choosing the scoring rule under the M-open case. Thus, the paper considers the log (L), quadratic (Q), spherical (S), CRPS (C), and FTMS (F) for estimating weights in Section 4. For a given predictive density $p(y_{T+1}|Y_T, M_k)$ and realization of $y_{T+1} = y^*$, these scoring rules will allot a score as

$$\begin{aligned}
L(y^*) &= log(p(y^*|Y_T, M_k)) \\
Q(y^*) &= 2p(y^*|Y_T, M_k) - \int_{-\infty}^{\infty} p(y_{T+1}|Y_T, M_k)^2 dy_{T+1} \\
S(y^*) &= \frac{p(y^*|Y_T, M_k)}{(\int_{-\infty}^{\infty} p(y_{T+1}|Y_T, M_k)^2 dy_{T+1})^{0.5}} \\
C(y^*) &= \int_{-\infty}^{y^*} F(y_{T+1}|Y_T, M_k)^2 dy_{T+1} + \int_{y^*}^{\infty} (F(y_{T+1}|Y_T, M_k) - 1)^2 dy_{T+1} \\
F(y^*) &= -\Big(\frac{y^* - \mu_i}{\sigma_i}\Big)^2 - log(\sigma_i^2),
\end{aligned} \tag{2.3}$$

where μ_i is the mean, σ_i is the standard deviation and $F(y_{T+1}|Y_T, M_k)$ is the cumulative predictive density for M_k. An issue with the CRPS rule is that the optimization becomes computationally heavier as the number of predictive densities increases due to the presence of integration (Gneiting

and Raftery (2007)). Dawid and Sebastiani (1999) suggested four proper scoring rules based on the first two moments of the predictive distribution, and $F(y^*)$ is chosen to be the most popular one in this paper. Given Y_T is realized, the weights for the opinion pools are estimated as

$$w_T^* \quad = \quad arg \ \max_{w_T} \sum_{t=1}^{T} S\Big(\sum_{k=1}^{K} w_{k,T} p(y_t|Y_{t-1}, M_k)\Big), \tag{2.4}$$

where $w_T^* = \{w_{1T}^*, w_{2T}^*, \ldots, w_{KT}^*\}$. The opinion pool for y_{T+1} will take the form

$$p(y_{T+1}|Y_T) \quad = \quad \sum_{k=1}^{K} w_{k,T}^* \, p(y_{T+1}|Y_T, M_k). \tag{2.5}$$

3 Bayesian Opinion Pool

The paper attempts to estimate opinion pools using the proper Bayesian formulation and obtain the posterior distribution of weights. The weights are treated as a K-dimensional, simplex bound, random variable endowed with a Dirichlet prior.

$$p(w_t) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_K), \tag{3.1}$$

where the hyperparameter α_k determines the relative weight given to $M_k \ \forall \ k = 1, 2, \ldots, K$, thus allowing the incorporation of prior information for any forecaster. One of the important properties of the Dirichlet Prior is its adaptability to applications related to model selection and model averaging. If the value of α_k is kept above 1 for all k, the prior penalizes allotting extreme weights to some models; thus, the posterior mean of weights tend to be closer to each other. As α_k tends towards infinity for all k, the posterior mean of weights tends towards equal weights. If the value of α_k is kept below 1 for all k, the prior incentivizes extreme weights for some models. As α_k tends towards 0, the posterior weights tend towards choosing the best model among the set. This is useful in case the application requires model selection. If $\alpha_k = 1$ for all k, then the prior becomes uniform and lets the data steer the posterior mean of weights towards optimized values.

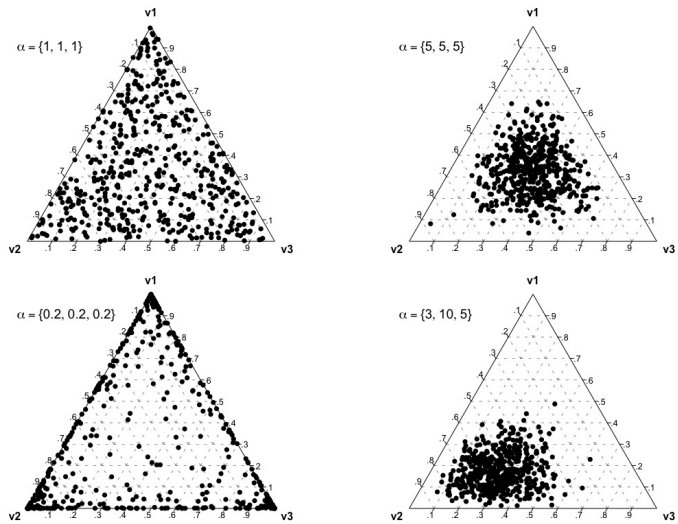


Figure 1: Draws from 3-dimensional Dirichlet with different α

To illustrate, Figure 1 shows 4 different cases of draws from 3-dimensional Dirichlet distribution for different values of $\alpha = \{\alpha_1, \ldots, \alpha_k\}$. The top-left figure represents the uniform prior with $\alpha = \{1, 1, 1\}$ which also shrinks weights towards equality since the mean of weights is $1/3$. The top-right figure represents a stronger shrinkage of weights towards 1/3 since $\alpha = \{5, 5, 5\}$ which can be useful for model averaging. The bottom-left figure represents the shrinkage of weights towards boundary cases since $\alpha = \{0.2, 0.2, 0.2\}$ which can be useful for model selection. The bottom right figure represents the case when non-sample information is available about the experts and all experts are not preferred equally.

Given that the opinion pool itself is an appropriate density function, it makes sense to treat it as the joint conditional distribution (equivalent to the joint likelihood function), a sequence of one-step-ahead conditional distributions, each of them a mixture generated by these weights, which is given as

$$\begin{aligned} p(Y_T|w_T) &= \prod_{t=1}^{T} \Big(p(y_t|Y_{t-1)} \Big) \\ &= \prod_{t=1}^{T} \Big(\sum_{k=1}^{K} w_{k,T}\, p(y_t|Y_{t-1}, M_k) \Big). \end{aligned} \tag{3.2}$$

Given the prior and the conditional distribution, the posterior distribution of the weights will look like

$$\begin{aligned} p(w_T|Y_T) &\propto p(Y_T|w_T)p(w_T) \\ &\propto \prod_{t=1}^{T}\Big(\sum_{k=1}^{K} w_{k,T}\, p(y_t|Y_{t-1}, M_k)\Big)\prod_{k=1}^{K} w_{k,T}^{\alpha_k-1}. \end{aligned} \tag{3.3}$$

It is easy to see that the log score optimization function (optimal prediction pools by Geweke and Amisano (2011)) is a monotonic transformation of the conditional distribution.

$$log\Big(\prod_{t=1}^{T}\big(\sum_{k=1}^{K} w_{k,T}\, p(\pi_t|\pi_{1:t-1})\big)\Big) = \sum_{t=1}^{T} log\Big(\sum_{k=1}^{K} w_{k,T}\, p(\pi_t|\pi_{1:t-1})\Big)$$

Therefore, the mean of the posterior distribution of weights will coincide with the optimal prediction pool asymptotically. Gneiting and Raftery (2007) showed that the log score minimizes the Kullback–Leibler divergence from DGP to the prediction model. This means the weights under the BOP minimize the Kullback–Leibler divergence distance from DGP to the model since the prior disappears in a large sample (Proof in the Appendix). In small sample settings, the estimates of BOP will differ from the optimal prediction pool as the BOP weights will shrink towards the prior. Since, under micronumerosity, it is not feasible to optimize the function, the BOP with a uniform prior can be seen as an extension of the optimal prediction pool, broadening its applicability.

To illustrate how the prior distribution interacts with the conditional distribution, Figure 3 showcases how uniform prior and boundary prior ($\alpha < 1$) shrink weights. The uniform prior expands the set of feasible vectors of weights to cases where all models are relatively equally considered. Increasing the value of α will increase the strength of shrinkage and thus be useful in applications requiring all models to participate in the final analysis (model averaging). The boundary prior shrinks weights towards the best model which can be useful for applications related to model selection.

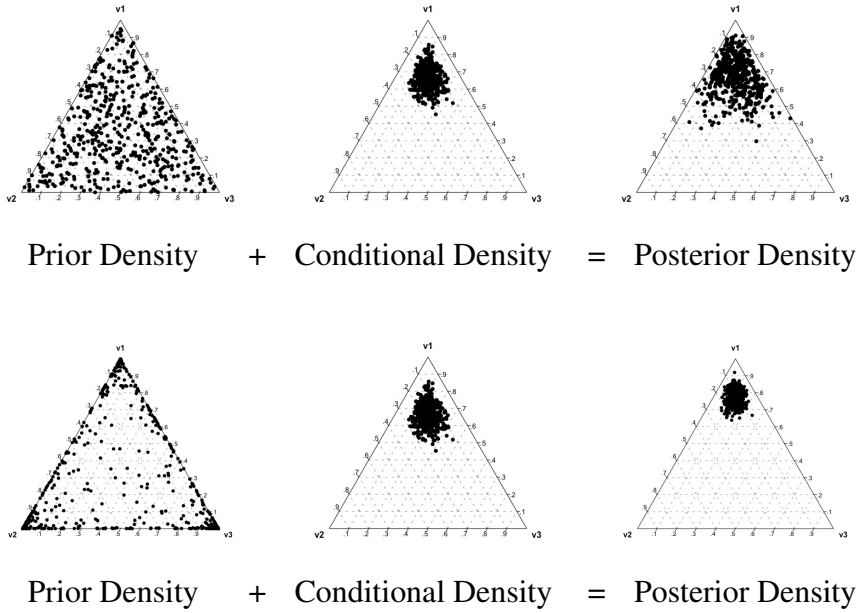


Figure 3: Illustration of weight's shrinkage under Bayesian opinion pool

Since the final form of the posterior is non-standard, the paper uses the Metropolis-Hasting (MH) algorithm to draw from the posterior density. For the proposal density, one potential candidate can be the Dirichlet distribution centred at the previous draw. Let $w_T^{(g)}$ be w_T drawn in the gth iteration. The Markov Chain Monte Carlo (MCMC) estimation of the BOP for the Dirichlet proposal is summarized in the following steps.

STEP 1. Choose a value of $w_T = w_T^{(0)}$.

STEP 2. At the gth iteration, sample $w_T^{(g)} \sim Dir(\alpha^{(g-1)})$ where $\alpha^{(g-1)}$ is chosen to center the distribution at $w_T^{(g-1)}$.

STEP 3. Generate $u \sim U(0,1)$.

STEP 4. If $u \leq min\left(\frac{p(w_T^{(g)}|Y_T)Dir(w_T^{(g-1)}|w_T^{(g)})}{p(w_T^{(g-1)}|Y_T)Dir(w_T^{(g)}|w_T^{(g-1)})}\right)$, return $w_T^{(g)}$, else return $w_T^{(g-1)}$. Go to step 2 and continue until the desired number of iterations is obtained.

Since there is no obvious choice for $\alpha^{(g-1)}$ in step 2, the researcher can choose $\alpha^{(g-1)} = cw_T^{(g-1)}$, where c is chosen to make the mean of $\alpha^{(g-1)}$ equal to 1. For the Dirichlet proposal, the acceptance

rate may be too low if the posterior density is narrow or the dimension is high. Alternatively, the vector of weights can be transformed to be defined on an unbounded domain using a multivariate logit transformation. Given $\theta_T = \{\theta_{1,T}, \ldots, \theta_{K-1,T}\}$, the transformation will look like

$$\theta_{k,T} = \ln(\frac{w_{k,T}}{w_{K,T}}) \tag{3.4}$$

for all $k = 1, \ldots, K-1$, where $\theta_T \sim N(\bar{\theta}_T, \bar{\Omega}_T)$. The mean of the Gaussian proposal, $\bar{\theta}_T$ is kept as some optimized value, which is calculated using an imperfect back-fitting MCMC algorithm (details can be found in the Appendix) since the numerical optimization of the conditional distribution fails in the case of micronumerosity. The covariance matrix, $\bar{\Omega}_T$ can be kept equal to either σI_{K-1} where σ is decided based on the rejection rate or proportional to the inverse Hessian of the conditional distribution at $\bar{\theta}_T$. Let $\theta_T^{(g)}$ be θ_T drawn in the gth iteration. The MCMC estimation of the BOP for the transformed proposal is summarized in the following steps.

STEP 1. Choose a value of $\theta_T = \theta_T^{(0)}$

STEP 2. At the gth iteration, sample $\theta_T^{(g)} \sim N(\bar{\theta}_T, \bar{\Omega}_T)$.

STEP 3. Transform $\theta_T^{(g)}$ to obtain $w_T^{(g)}$.

STEP 4. Generate $u \sim U(0, 1)$.

STEP 5. If $u \leq min\left(\frac{p(w_T^{(g)}|Y_T)q(w_T^{(g-1)})}{p(w_T^{(g-1)}|Y_T)q(w_T^{(g)})}\right)$, return $w_T^{(g)}$, else return $w_T^{(g-1)}$. Go to step 1 and continue until the desired number of iterations is obtained.

The density $q(\cdot)$ is the transformed density for w_T obtained after incorporating the Jacobian of the transformation. The mean of the Gaussian proposal, $\bar{\theta}_T$, can also be kept as the previous draw (just like in the case of the Dirichlet proposal). In that case, a starting value of $\theta_T = \theta_T^{(0)}$ would be required.

The framework is not restrictive to one-step-ahead densities and can be extended for long horizons. For a given predictive density $p(y_{t+h}|Y_t, M_k)$ representing h step ahead forecast, the posterior

distribution of the weights will look like

$$p(w_T^h|Y_T) \propto \prod_{t=1}^{T-h} \Big(\sum_{k=1}^{K} w_{k,T}^h \, p(y_{t+h}|Y_t, M_k) \Big) \prod_{k=1}^{K} (w_{k,T}^h)^{\alpha_k - 1}. \tag{3.5}$$

where w_T^h represents weights optimized using information upto time T and used for predictions in period $T + h$.

4 Simulation Study

This section explores the predictive performance of BOP and TOP on simulated data. The DGP and individual models are considered under the linear setting to preserve useful insights that might get lost in a complicated analysis. The data is artificially generated for a dependent variable as

$$DGP : y_t = 0.5 + 0.5y_{t-1} + \epsilon_t \ , \ \text{ where } \ \epsilon_t \overset{\text{iid}}{\sim} N(0, 5) \tag{4.1}$$

Three experts submit their forecasts for y_t as $N(\hat{y}_{kt}, 4)$, where $E(\hat{y}_{kt}) = y_t$ for $k = 1, 2$ and 3. Therefore, the expert's predictions are unbiased and only differ in the variance as follows

	Case 1	Case 2
Expert 1	var$(\hat{\pi}_{1t}) = 4$	var$(\hat{\pi}_{1t}) =$ AR(1)
Expert 2	var$(\hat{\pi}_{2t}) = 8$	var$(\hat{\pi}_{2t}) =$ AR(1)
Expert 3	var$(\hat{\pi}_{3t}) = 16$	var$(\hat{\pi}_{3t}) =$ AR(1)

Case 1 tests the situation where there exists a clear ranking in accuracy for the expert's predictions. Expert 1 is the most accurate whereas expert 3 is the least. This scenario tests the ability of BOP and TOP to identify the best model as the training sample increases. Case 2 models the variance of the predictions to follow an autoregressive process with varying degrees of persistence. This scenario introduces persistence to the accuracy of an expert's prediction while allowing the

ranking of experts to change over time. Thus, one expert can predict accurately for some periods and inaccurately for others. This tests the ability of BOP and TOP to identify the combination of weights which optimizes forecasting accuracy while ensuring against bad predictions due to the reliance on any one of the experts.

The opinion pools are trained using the following sample sizes: $T \in \{5, 10, 20, 30, 50, 100\}$ for case 1 and $T = \{30, 100, 200\}$ for case 2. Case 2 is tested with a larger T allowing for rankings to change over time. The testing sample is kept at 30 observations and tested for short-term (one step ahead), medium-term (three steps ahead), and long-term (six steps ahead) forecasting horizons (h). The persistence levels of 0.2, 0.5 and 0.8 are used to test the sensitivity of results in case 2. The predictive exercise uses the rolling window approach with the window length equivalent to the size of T. The BOP is estimated using the Dirichlet proposal density with varying values of α_k.

4.1 Volatility of Weights under Small Sample Setting

The paper considers case 1 to study the behaviour of weights since a clear ranking of models is defined. This exercise finds that the weights under TOP have high volatility when T is small (near micronumerosity). This can be seen in Figure 4 which shows the weight evolution for TOP and BOP for $T = 5, 10$ and 30. Each row represents different types of opinion pools whereas each column represents the sample size. For $T = 5$ and 10, the weights for TOP have a high variance, especially for log, quadratic and spherical opinion pools where they oscillate as extremely as between 0 and 1. This indicates that TOP is relying on the immediate predictive accuracy of the expert since there is a lack of data regarding their consistency. On the other hand, BOP is much more stable, stays close to the prior and is able to identify the best model while not overly relying on it for prediction. This is due to the Dirichlet prior with $\alpha_k = 1$ which imposes sufficient shrinkage (since it is the case of near micronumerosity) on weights. With the data flowing in, the weights deviate from equal weights as enough evidence is present about the accuracy of the concerned model. For $T = 30$, almost all the opinion pools are able to identify the best model and allot weights according to the ranking of the models.

Figure 4: Weights evolution for opinion pools

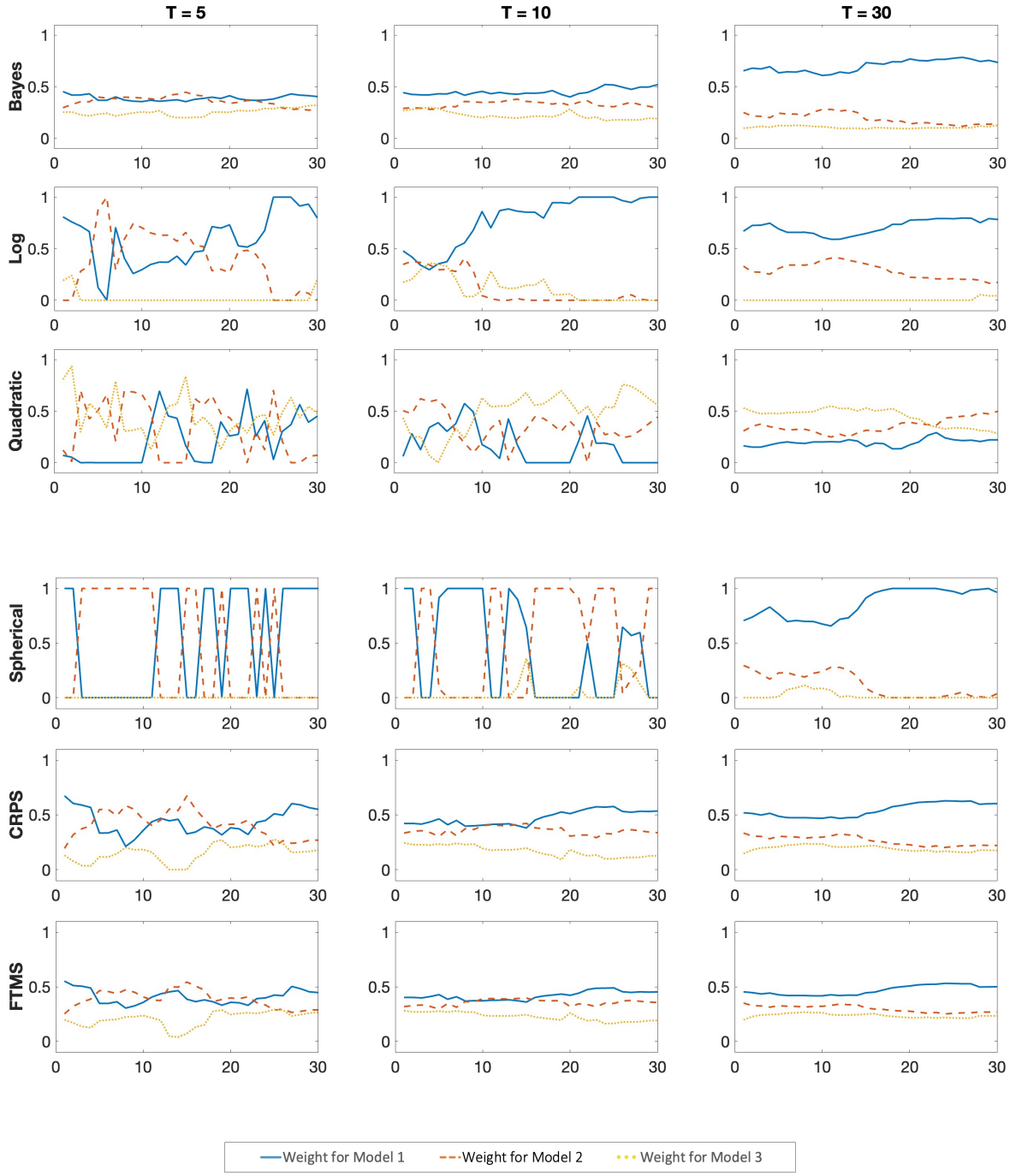


Table 1 contains the summary of the standard deviation for weights for TOP and BOP. The standard deviation is estimated by first calculating the standard deviations of the weight corresponding to the individual models over the testing period and then taking the mean of those standard devi-

ations (three models). The BOP has the lowest standard deviation for small samples ($T = 5$ and 10) because the prior stabilizes the weights around equality. As the sample size increases ($T = 20$ and 30), the weights for TOP become stable as well and start to converge towards the best model. This is intuitive since the opinion pools are able to identify the true DGP when T is high.

Table 1: Standard deviation of weights

Sample	Log	Quad	Sphere	CRPS	FTMS	Bayes
5	0.241	0.306	0.412	0.125	0.075	**0.040**
10	0.179	0.170	0.331	0.093	0.059	**0.045**
20	0.088	0.069	0.172	0.038	0.024	0.038
30	0.054	0.048	0.086	0.027	0.018	0.029

The Dirichlet prior stabilizes the evolution of weights over time; leading to BOP having the lowest volatility under the small sample setting. The stability over time allows the opinion pool to avoid overfitting, which is one of the crucial features of a good predictive model. This positive spillover affects the predictive performance which is discussed in Subsection 4.3.

4.2 Shrinkage of Weights

The paper considers case 1 to study weight's behaviour under shrinkage since a clear ranking of models is defined. Since the Dirichlet prior allows the researcher to choose the intensity of shrinkage, it is important to see how the weight's behaviour changes as α_k changes. Figure 5 shows the evolution of weights when $\alpha_k = 1, 3$ and 5. for $T = 10, 30$ and 50. It is observed that the weights converge towards equality as α_k is increased which is an expected result. This property allows BOP to be used in applications related to model averaging. As α_k tends to infinity, BOP tends towards the simple opinion pool (opinion pool with equal weights).

Figure 6 shows the evolution of weights when $\alpha_k = 1, 0.6$ and 0.3. for $T = 10, 30$ and 50. It is observed that the weights diverge away from equal weights and the best model is been preferred

Figure 5: Bayesian Opinion Pool with $\alpha > 1$ for Model Averaging

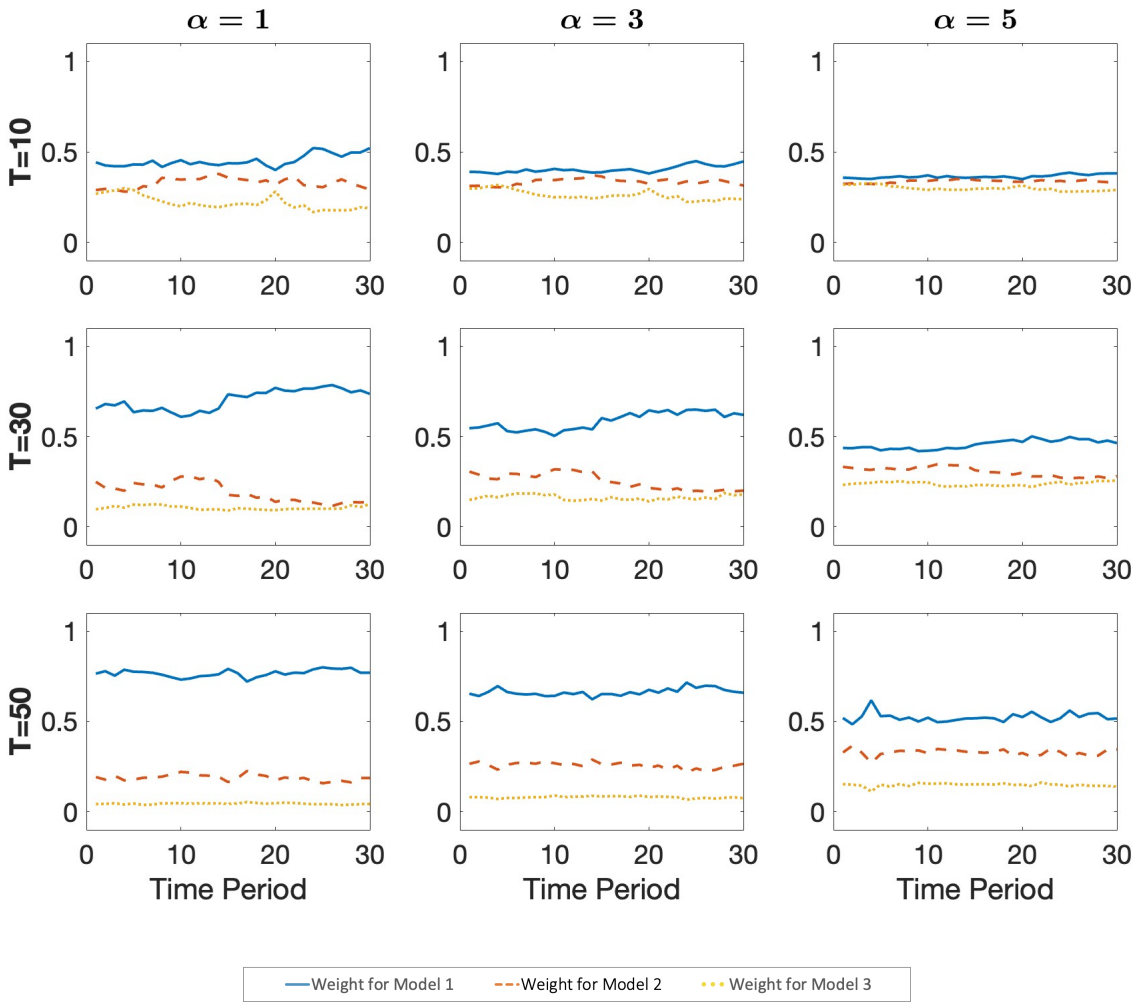
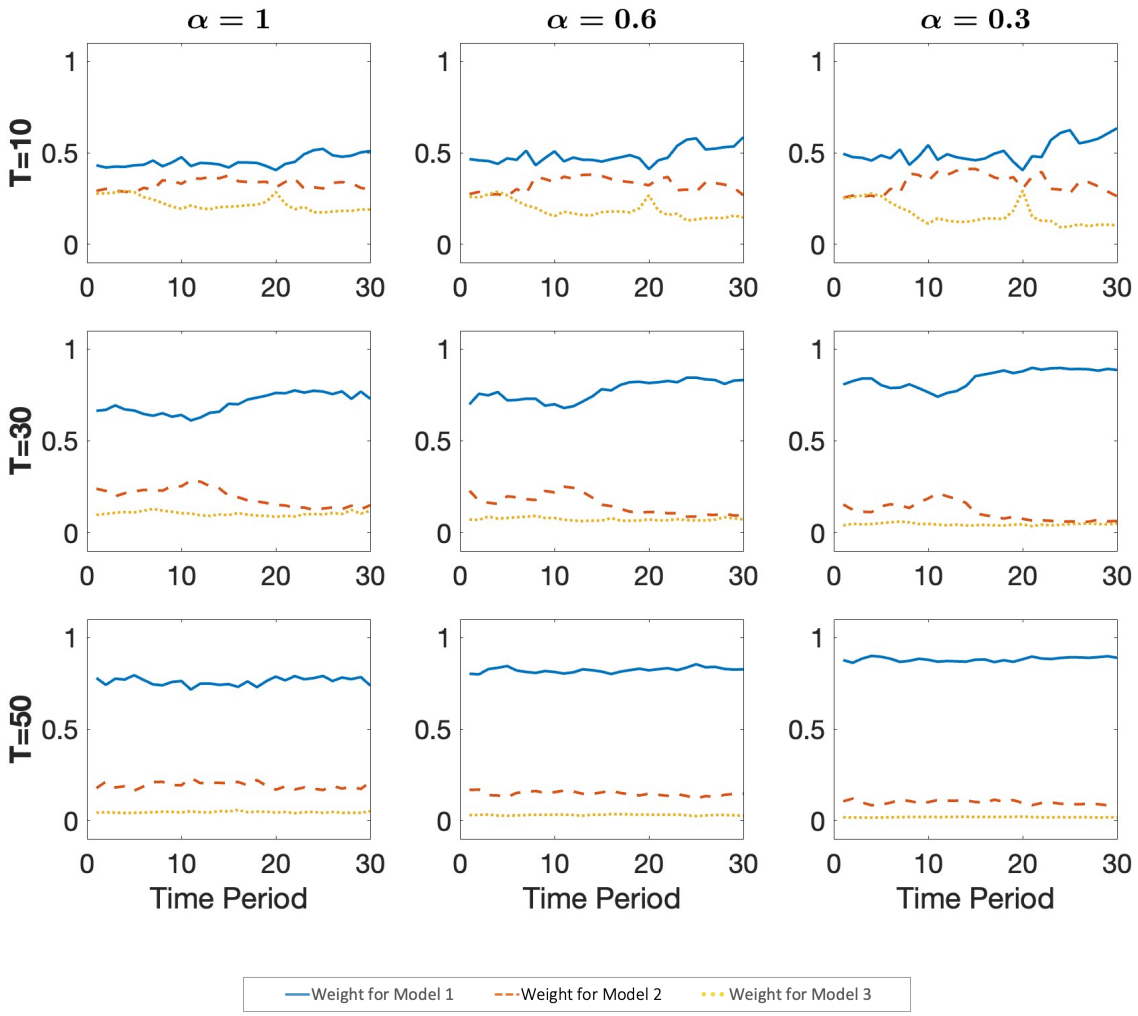


Figure 6: Bayesian Opinion Pool with $\alpha < 1$ for Model Selection



among the available ones. As α_k tends to 0, BOP degenerates into the best model thus preferred for applications related to model selection.

4.3 Forecasting Performance

The paper uses MSPE to evaluate the predictive performance of various opinion pools. Table 2 contains MSPE values for case 1 for different sample sizes and forecasting horizons. The columns with the lowest MSPE are highlighted. BOP has the lowest MSPE for most of the scenarios. One can observe that as the sample size increases, the optimal α_k also increases. Since the conditional density dominates the prior when T is high, a stronger prior leads to optimal shrinkage. Among

Table 2: Mean Square Prediction Error for Case 1

Sample (T)	Horizon (h)	Log	Quad	Sphere	CRPS	FTMS	Bayes $\alpha_k = 1$	Bayes $\alpha_k = 2$	Bayes $\alpha_k = 3$
5	1	144	212	200	81.5	75.7	**74**	76.1	77.4
5	3	152	176	168	78.8	72.4	**70.5**	71.8	72.8
5	6	153	147	160	68.6	**62.5**	62.9	64.6	65.4
10	1	133	157	157	88.7	88.6	**86.9**	89.5	91.7
10	3	129	152	160	79.8	80.5	**79.1**	82.7	84.6
10	6	127	132	158	79.9	77.1	**73.5**	76.0	77.5
20	1	124	130	126	70.5	72.6	**68.1**	70.0	72.4
20	3	111	124	119	64.5	67.1	**62.5**	64.7	66.9
20	6	93.2	114	99.3	56.9	59.2	**55.2**	57.3	59.4
30	1	112	131	88.2	70.7	74.8	**67.4**	69.1	72.0
30	3	106	127	83.0	65.5	70	**62.8**	64.2	67.6
30	6	90.0	115.4	75.3	59.2	63.3	**56.3**	58.4	61.4
50	1	124	144.6	105	76.9	83.8	82.5	**76.1**	76.9
50	3	114	135	97.6	71.3	78.2	76.4	**71.0**	71.6
50	6	109	126	92.1	67.3	73.2	72.4	**67.1**	67.8
100	1	115	124	77.6	68.3	73.6	78.3	68.4	**65.4**
100	3	106	114	70.1	61	66.2	70.7	61.0	**58.1**
100	6	98.0	106	63.7	56.2	61.2	64.3	56.1	**54.1**

the TOP, CRPS and FTMS perform well and their MSPE is significantly lower than log, quadratic and spherical. For FTMS, since the predictions of experts vary only through mean and variance in the DGP (normal distribution is imposed), the first two moments incorporate sufficient information for opinion pools. For CRPS, it captures the idea of proximity better than other scoring rules and thus performs well in this simulation exercise.

Table 3 contains MSPE values for case 2 for different sample sizes, forecasting horizons and persistence. The results are similar to that of Table 2 where BOP dominates for the majority of cases. CRPS and FTMS perform significantly better than log, quadratic and spherical scoring rules. This shows that BOP is able to capture the dynamic behaviour of experts when the ranking based on predictive accuracy changes over time.

Table 3: Mean Square Prediction Error for Case 2

Persistence	Sample Size	Forecasting Horizon	Log	Quad	Sphere	CRPS	FTMS	Bayes $\alpha_k = 1$
0.2	30	1	14.30	6.84	10.24	5.26	5.25	**5.22**
0.2	30	3	13.22	6.14	8.88	**4.60**	4.61	4.62
0.2	30	6	11.29	5.83	8.31	4.18	4.18	**4.18**
0.2	100	1	12.64	4.75	6.45	**4.12**	4.13	4.16
0.2	100	3	10.84	4.36	5.64	3.74	3.75	**3.72**
0.2	100	6	9.32	4.05	4.95	3.47	3.48	**3.45**
0.2	200	1	10.82	5.98	6.42	5.63	5.64	**5.61**
0.2	200	3	10.28	5.13	5.55	4.91	4.92	**4.91**
0.2	200	6	7.99	4.61	4.78	4.34	4.35	**4.27**
0.5	30	1	15.41	7.87	10.81	5.54	5.55	**5.51**
0.5	30	3	12.97	7.061	9.39	**4.91**	4.92	**4.91**
0.5	30	6	12.84	6.70	8.54	4.45	4.45	**4.43**
0.5	100	1	12.52	5.17	7.03	**4.45**	4.47	4.46
0.5	100	3	11.22	4.74	6.25	4.00	4.03	**3.96**
0.5	100	6	9.61	4.37	5.48	3.72	3.74	**3.68**
0.5	200	1	12.41	6.90	7.25	6.33	6.35	**6.24**
0.5	200	3	11.28	5.97	6.37	5.59	5.60	**5.53**
0.5	200	6	8.88	5.41	5.55	5.00	5.02	**4.87**
0.8	30	1	18.27	10.62	13.77	7.03	7.10	**6.98**
0.8	30	3	18.97	9.70	14.27	6.47	6.50	**6.45**
0.8	30	6	19.40	8.82	13.28	6.00	5.99	**5.94**
0.8	100	1	15.57	6.59	9.17	5.76	5.80	**5.75**
0.8	100	3	14.51	6.17	9.02	5.38	5.42	**5.34**
0.8	100	6	14.00	5.69	8.01	4.97	5.00	**4.94**
0.8	200	1	18.22	9.79	10.42	8.77	8.81	**8.61**
0.8	200	3	16.74	8.45	9.21	7.74	7.77	**7.69**
0.8	200	6	13.73	7.69	8.18	7.01	7.05	**6.86**

5 Application: Inflation Prediction using the Survey of Professional Forecaster

The Survey of Professional Forecasters is a useful source of data for economists and policymakers. Croushore and Stark (2019) in ”The Fifty Years of the Survey of Professional Forecasters” stated, ”In 2018, the survey generated more than 45,000 unique hits to the Philadelphia Fed’s external webpages...The audience consists of academic researchers..., policymakers...and business people”. Figure 7 shows the increase in citations and publications of papers per year which contain ”Survey of Professional Forecasters” in their title, abstract or keywords.

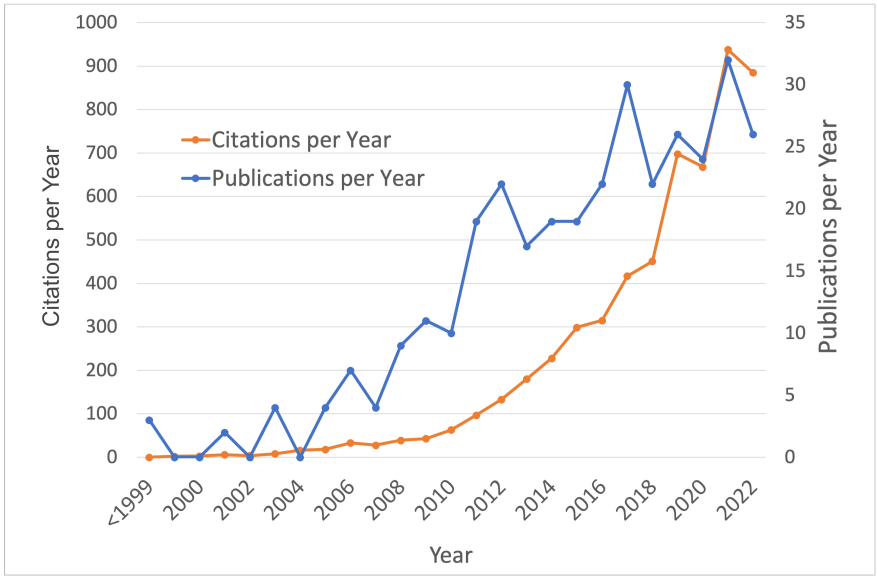


Figure 7

The Federal Reserve Bank of Philadelphia publishes individual and aggregate density projections (and point estimates) for macroeconomic variables every quarter. They survey individual professional forecasters immediately after the U.S. Bureau of Economic Analysis (BEA) releases data. A unique ID is assigned to each forecaster, making tracking them possible. Anonymity is maintained to prevent strategic misreporting. The experts submit their forecast densities by allotting probabilities to bins (range of inflation rates) which are predetermined by the Fed so that the final density takes the form of a histogram. The details of the data set and its significance can be

found in Croushore et al. (2019), Clements et al. (2023) or on their website. This paper focuses on inflation density forecasts. (Diebold et al. (1997)) argued that point forecasts from SPF are extensively used in macroeconomic literature, but density forecasts are relatively less explored.

SPF is used practically for two purposes. First, it is an estimator for inflation expectations and thus is used to track them. Keane and Runkle (1990) argue that a model with rational agents can be better represented using the predictive data from SPF. Professional forecasters should be better informed and thus justify the assumption of rationality in macro models. Carroll (2003) evaluated the influence of SPF data on private-sector expectations. Second, SPF is used to forecast inflation accurately or test forecasting models thus, facilitating decisions requiring accurate inflation predictions. Smets et al. (2014) incorporated SPF data to measure the forecasting accuracy of New Keynesian DSGE models.

The Federal Bank of Philadelphia publishes aggregated inflation forecasts density calculated by taking a simple average of density estimates submitted by individual experts. Equal weights are a reasonable choice if the objective is to track inflation expectations. Since, the aim is to capture how rational agents perceive future inflation, including everyone's opinion captures the idea of inflation expectations. Also, numerical optimization is infeasible as 160 forecasters participated during 120 quarters (Q1 1992 to Q4 2021), with an average of 35 active forecasters per quarter. The number of forecasters is always higher than the number of data points for any rolling window.

If the objective of SPF is inflation forecasting, then equal weights are a sub-optimal choice. Aastveit et al. (2018) mentioned that "Despite the long history of the SPF, little attention has historically been paid to how the weights on the competing forecast densities in the finite mixture should be determined". The issue with the simple opinion pools (SOP) approach is that it leads to the loss of past predictive performance information and produces an inefficient estimator. Figure 8 presents the predictive performance of experts who are active for at least 10 quarters in the period of Q1 1992 to Q4 2021. The vertical axis represents the probability allotted by an expert to the bin which contained the realized value of the inflation rate. Thus, higher the probability allotted by the expert, better is the forecast. The horizontal axis represents the unique ID of experts. The

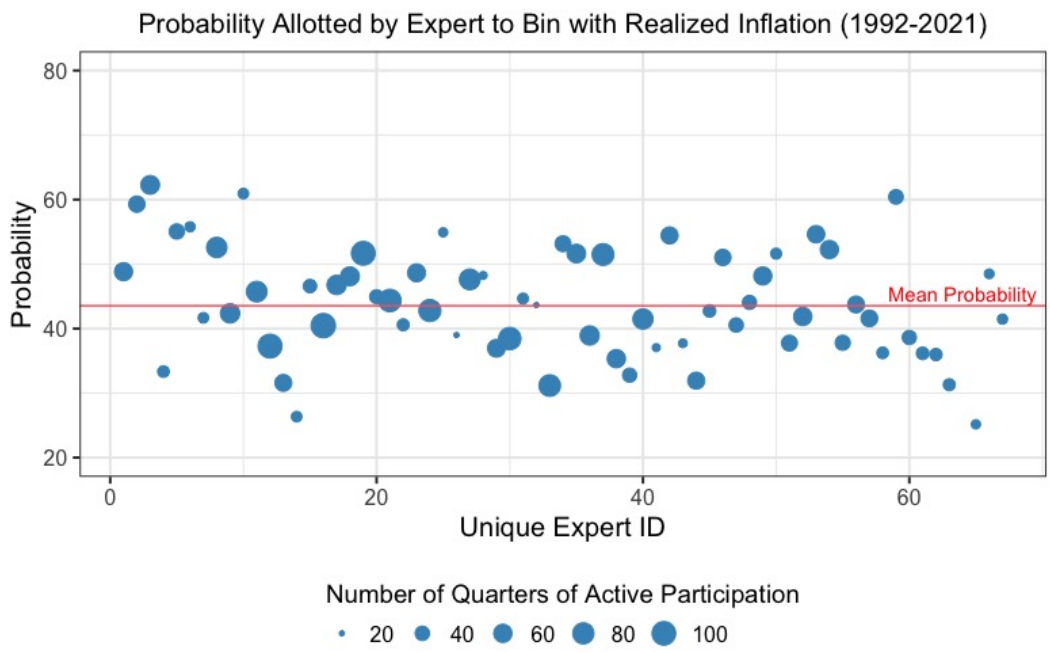


Figure 8

size of the points represents the number of quarters, an expert was active in the past. The figure clearly depicts that there are some experts who were consistently active and allotted much higher probability to the realized inflation rate than the average and vice versa. Using equal weights ignores this information and thus there is an opportunity to improve the predictive accuracy of aggregated inflation forecast density.

This paper aggregates inflation density forecasts using the BOP with a uniform prior ($\alpha_k = 1$). BOP not only accounts for important information related to the past accuracy of individual forecasters but also allows estimation under micronumerosity where optimization-based methods fail. The decision to choose a uniform prior is guided by the non-sampling information. Since the Fed uses equal weights to aggregate densities, it can be considered a good benchmark to start from. Also, researchers in the past have frequently found combining point forecasts with equal weights to be very competitive with the more complicated weighting techniques. Clemen (1989) shows in his review that equal weights are difficult to beat. Similar results were concluded by Stock and Watson (1999) and Fildes and Ord (2002). The paper also considers tight priors of $\alpha_k = 1.5$ and 3. The prior shrinks the BOP towards SOP but still allows some movement in case strong evidence

for better relative predictive accuracy is present.

Frequent entry and exit of forecasters make optimization of the opinion pool more involved. Capistrán and Timmermann (2009) elaborated on the problem of having an unbalanced panel and recommended filling in the missing values before aggregation. They also considered using the unbalanced panel by keeping only the frequent forecasters. However, they had to resort to the simple average when there were fewer remaining forecasters than parameters to be estimated.

This paper does not fill in for the missing forecasting density and follows the following method.

- Entry: Suppose a forecaster is unavailable in the training data (m quarters moving window) but submits the prediction for the $(m+1)^{th}$ quarter. Thus, there is no information on the past predictive performance. In that case, their density is allotted $1/A$ weight (equal weight), where A is the number of active forecasters in the $(m+1)^{th}$ quarter (Alternatively, the researcher can decide to include the expert only if they have participated for a certain number of quarters).

- Exit: Suppose a forecaster was available in the training data (m-quarter moving window) but not for the $(m+1)^{th}$ quarter. In that case, their density will be allotted 0 weight, and they will not be considered in the optimization process.

To explain it better, let us assume that 40 forecasters were active in the last 20 quarters (not necessarily for every quarter), which is the training period for this case. Only 10 forecasters submitted their predictions for the 21st quarter, including 2 new ones. Then, the weights allotted to these 2 new ones would be $1/10$ each, and the weights for the remaining 8, whose values were estimated based on the past data (excluding the twelve inactive forecasters), would be normalized so that the total sum of these active 10 weights is 1.

The application considers moving windows approach with varying lengths from 21 to 29 quarters of training data and the rest of the period until 2021 Q4 as testing data. The prior with $\alpha_k > 1$ tends to bring weights closer $\frac{1}{K}$ where K is the total number of active experts.

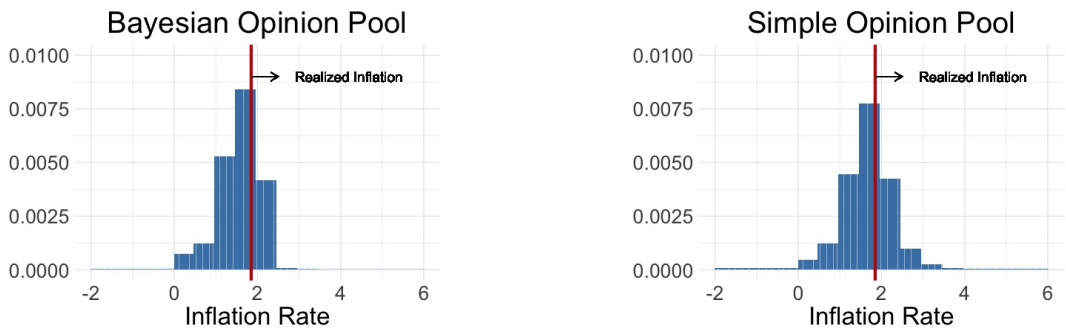


Figure 9: Simple and Bayesian opinion pool densities for Q1 2014

To visually aid the understanding of the information aggregation process under the BOP (uniform prior) and SOP, Figure 9 demonstrates BOP and SOP final predictive densities for inflation for Q1 2014. A total of 40 forecasters submitted their predictions. The final densities look different representing the fact that BOP allotted different weights to each expert than SOP. In this particular case, BOP allots a higher probability to the realized inflation than SOP

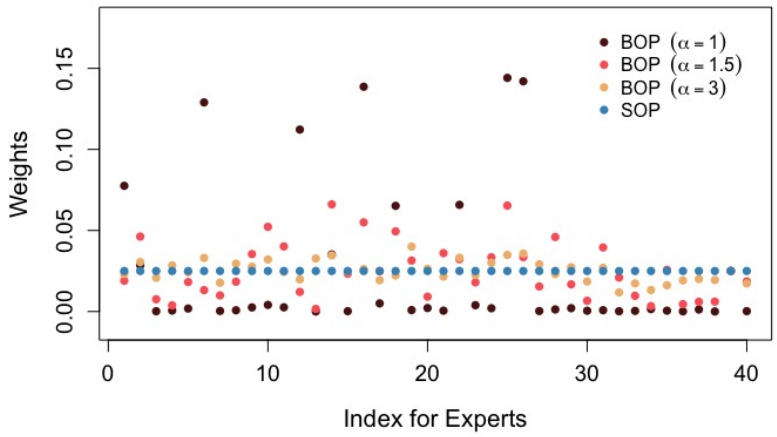
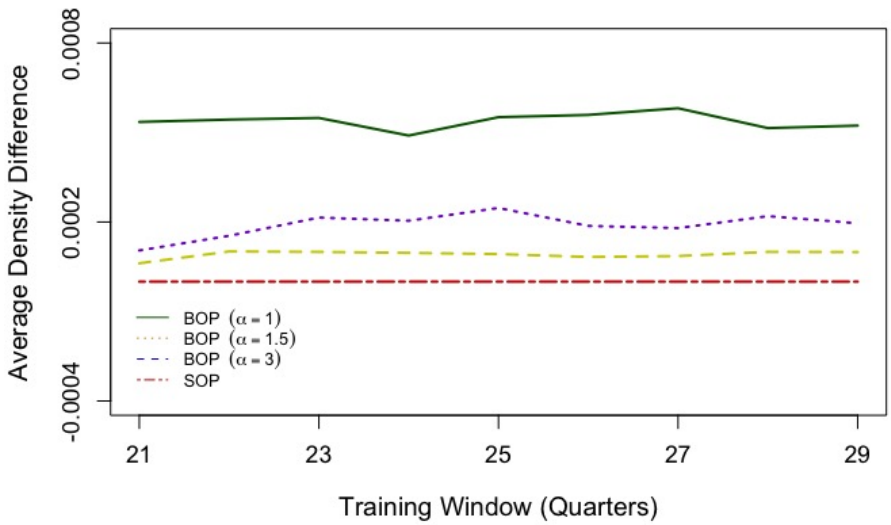


Figure 10: Weights allotted to experts under BOP and SOP for Q1 2014

Figure 10 shows estimated weights under BOP for different values of α_k for Q1 2014. It shows how the weights become concentrated around equal weights when α_k increases, showcasing the strong shrinkage implied.

The aggregated predictive densities, representing out-of-sample forecasts, are tested based on

Figure 11: Average Density Evaluated at the Realized Inflation Rate for Opinion Pools



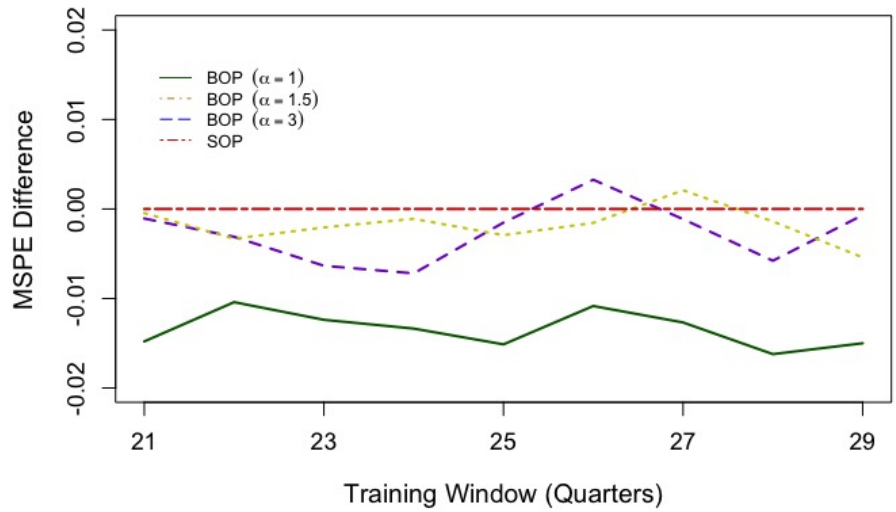
the average density allotted to the realized inflation rate. Figure 11 shows the average density difference between SOP and BOP evaluated at the realized inflation rate for different rolling windows. The difference is normalized to 0 and thus the vertical red line at the origin is represented by SOP. As the training window varies, the predictive accuracy of opinion pools is tested on the corresponding remaining quarters of data until Q4 2021, and the weights are updated each quarter (recursive). The average density allotted by the BOP is always higher than the SOP. As α_k increases, the MSPE difference between SOP and BOP becomes smaller representing that BOP is tending towards SOP. The difference is significant at 5% for the uniform prior for all training window lengths considered.

Table 4: MSPE associated with opinion pools for different training windows

	21	22	23	24	25	26	27	28	29
BOP	0.433	0.440	0.445	0.453	0.451	0.457	0.462	0.467	0.468
SOP	0.449	0.454	0.459	0.463	0.467	0.471	0.476	0.482	0.488
Diff	-0.016***	-0.014**	-0.014**	-0.01*	-0.016***	-0.014***	-0.014**	-0.015***	-0.02***

Significance: 0.1 (*), 0.05 (**) and 0.01 (***)

Figure 12: MSPE associated with opinion pools for different training windows



To test whether the improvement in the density forecast estimator spills over to the estimates of point forecasts, opinion pools are compared using MSPE. Table 4 presents MSPE for BOP with a uniform prior ($\alpha_k = 1$) and SOP. The MSPE for BOP is smaller than SOP for every length of training window considered. The difference in MSPE between BOP and SOP is significant. Figure 12 shows the MSPE difference between BOP and SOP for different training windows. The difference is normalized to 0. The figure also shows the MSPE difference for BOP with tight priors ($\alpha_k = 1.5$ and 3) which is smaller. This shows that the complete information incorporated in estimating the forecast density has a positive spillover effect on the estimate of the point forecast. It also suggests that all forecasters are not equal in their predicting abilities, and the BOP exploits this asymmetry to its advantage.

6 Conclusion

This paper identifies two limitations associated with traditional opinion pools. First, optimizing opinion pools using a scoring function is not feasible if the number of models exceeds the data length, an issue faced in macroeconomics applications. Second, high volatility in weight associated

with TOP affects the predictive accuracy due to overfitting. Thus, a lot of researchers resort to equal weights since the marginal gains from optimized weights do not justify the cost of involving oneself in a complicated procedure.

This paper proposes estimating opinion pools under the Bayesian framework to resolve these issues. The Bayesian formulation allows the weights to be estimated under micronumerosity. The MCMC algorithm enables sampling of the high dimensional weight vector from its joint posterior distribution leading to efficiency gains. The use of a Dirichlet prior makes the weights relatively more stable over time and allows the researcher to control the shrinkage level. Apart from solving the issues, the BOP is found to be highly competitive compared to TOP, when micronumerosity is not the case.

In the application, the paper uses SPF data to obtain better estimates of inflation forecasts using the BOP. The Federal Bank of Philadelphia estimates the aggregate inflation density by allotting equal weights to all the individual densities. The BOP uses optimized weights based on the past accuracy of the forecaster, thus utilizing richer information and discouraging forecasters from providing extreme predictions. The application showed that the BOP (with various levels of shrinkage) outperforms the SOP published by the Federal Bank of Philadelphia.

The applications of BOP extend to macroeconomics or finance, especially in settings which deal with aggregating predictive densities. Gneiting and Ranjan (2013) combined predictive cumulative distributions and tested the approach on forecasting S&P 500 returns. McAlinn et al. (2020) used the Bayesian predictive synthesis for applications related to macroeconomic forecasting. Del Negro et al. (2016) estimated time-varying weights in linear opinion pools (Dynamic Pools) and used them to investigate the relative forecasting performance of dynamic stochastic general equilibrium (DSGE) models with and without financial frictions for output growth and inflation. Baştürk et al. (2019) combined density forecasts to improve portfolio strategies.

Given that the BOP can be applied to settings involving micronumerosity, where TOP can not be estimated, and its predictive accuracy is highly competitive with TOP when micronumerosity is not an issue, a stronger case can be made for its exploration and adoption in future research.

While the discussion in the paper focused on macroeconomic time series data, the usefulness of the techniques can be extended to any application involving model averaging and prediction. The utility of the BOP in other simulation settings, improvements in the MCMC algorithm and estimation of optimal shrinkage can be explored in future research work.

Acknowledgement

The author is grateful to Prof. Ivan Jeliazkov for his guidance, to Prof. Yingying Lee and Prof. Fabio Milani for their helpful comments and to UCI students and faculties for their valuable feedback.

References

Aastveit, K. A., J. Mitchell, F. Ravazzolo, and H. K. Van Dijk (2018). The evolution of forecast density combinations in economics. Technical report, Tinbergen Institute Discussion Paper.

Bacharach, J. (1974). Bayesian dialogues. *Unpublished manuscript, Christ Church College, Oxford University*.

Ball, L., N. G. Mankiw, D. Romer, G. A. Akerlof, A. Rose, J. Yellen, and C. A. Sims (1988). The new keynesian economics and the output-inflation trade-off. *Brookings papers on economic activity 1988*(1), 1–82.

Bassetti, F., R. Casarin, and F. Ravazzolo (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association 113*(522), 675–685.

Baştürk, N., A. Borowska, S. Grassi, L. Hoogerheide, and H. K. van Dijk (2019). Forecast density combinations of dynamic models and data driven portfolio strategies. *Journal of Econometrics 210*(1), 170–186.

Bates, J. M. and C. W. Granger (1969). The combination of forecasts. *Journal of the Operational Research Society 20*(4), 451–468.

Bernardo, J. M. and A. F. Smith (2000). *Bayesian theory*, Volume 405. John Wiley & Sons.

Billio, M., R. Casarin, F. Ravazzolo, and H. K. Van Dijk (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics 177*(2), 213–232.

Busetti, F. (2017). Quantile aggregation of density forecasts. *Oxford Bulletin of Economics and Statistics 79*(4), 495–512.

Capistrán, C. and A. Timmermann (2009). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking 41*(2-3), 365–396.

Carroll, C. D. (2003). Macroeconomic expectations of households and professional forecasters. *the Quarterly Journal of economics 118*(1), 269–298.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting 5*(4), 559–583.

Clements, M. P., R. W. Rich, and J. S. Tracy (2023). Surveys of professionals. In *Handbook of Economic Expectations*, pp. 71–106. Elsevier.

Clyde, M. and E. I. George (2004). Model uncertainty. *Statistical science 19*(1), 81–94.

Coibion, O., Y. Gorodnichenko, and R. Kamdar (2018). The formation of expectations, inflation, and the phillips curve. *Journal of Economic Literature 56*(4), 1447–91.

Croushore, D., T. Stark, et al. (2019). Fifty years of the survey of professional forecasters. *Economic Insights 4*(4), 1–11.

Dawid, A. P. and P. Sebastiani (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 65–81.

Degroot, M. H. and J. Mortera (1991). Optimal linear opinion pools. *Management Science 37*(5), 546–558.

Del Negro, M., R. B. Hasegawa, and F. Schorfheide (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics 192*(2), 391–405.

Diebold, F. X., A. Tay, and K. Wallis (1997). Evaluating density forecasts of inflation: the survey of professional forecasters.

Elliott, G., D. Ghanem, and F. Krüger (2016). Forecasting conditional probabilities of binary outcomes under misspecification. *Review of Economics and Statistics 98*(4), 742–755.

Fildes, R. and K. Ord (2002). Forecasting competitions–their role in improving forecasting practice and research. *A companion to economic forecasting*, 322–353.

Friedman, M. (1968). The role of monetary policy the american economic review. *New york 58*.

Geweke, J. and G. Amisano (2011). Optimal prediction pools. *Journal of Econometrics 164*(1), 130–141.

Geweke, J. and G. Amisano (2012). Prediction with misspecified models. *American Economic Review 102*(3), 482–86.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*(477), 359–378.

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronic Journal of Statistics 7*, 1747–1782.

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

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

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science 14*(4), 382–417.

Keane, M. P. and D. E. Runkle (1990). Testing the rationality of price forecasts: New evidence from panel data. *The American Economic Review*, 714–735.

Kydland, F. E. and E. C. Prescott (1982). Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, 1345–1370.

Long Jr, J. B. and C. I. Plosser (1983). Real business cycles. *Journal of political Economy 91*(1), 39–69.

McAlinn, K., K. A. Aastveit, J. Nakajima, and M. West (2020). Multivariate bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association 115*(531), 1092–1110.

McAlinn, K. and M. West (2019). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of econometrics 210*(1), 155–169.

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 niesr 'fan'charts of inflation. *Oxford bulletin of economics and statistics 67*, 995–1033.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys 29*(1), 46–75.

Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica: Journal of the Econometric Society*, 315–335.

Opschoor, A., D. Van Dijk, and M. van der Wel (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics 32*(7), 1298–1313.

Phelps, E. S. (1967). Phillips curves, expectations of inflation and optimal unemployment over time. *Economica*, 254–281.

Smets, F., A. Warne, and R. Wouters (2014). Professional forecasters and real-time forecasting with a dsge model. *International Journal of Forecasting 30*(4), 981–995.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature 58*(3), 644–719.

Stock, J. H. and M. W. Watson (1999). Forecasting inflation. *Journal of monetary economics 44*(2), 293–335.

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

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

Wang, H., X. Zhang, and G. Zou (2009). Frequentist model averaging estimation: a review. *Journal of Systems Science and Complexity 22*(4), 732–748.

Winkler, R. L., J. Munoz, J. L. Cervera, J. M. Bernardo, G. Blattenberger, J. B. Kadane, D. V. Lindley, A. H. Murphy, R. M. Oliver, and D. Ríos-Insua (1996). Scoring rules and the evaluation of probabilities. *Test 5*(1), 1–60.

Appendix A: Back-fitting MCMC algorithm

This subsection presents the MCMC algorithm used in the MCMC estimation of BOP. There is a possibility that $\bar{w_T}$ obtained is not the global maximum. Let w_o be the weight vector drawn in

the previous iteration and w_n be the weight vector drawn in the current iteration. The steps are as follows.

STEP 1. Draw w_n from a proposal density, be it Dirichlet, Normal distribution with logistic transformation (discussed in Section 3) or truncated normal (defined on the interval $[0, 1]$), where the proposal is centred at w_o. Normalize w_n so that the sum is 1 in case needed, and choose the variance so that the whole space can be explored.

STEP 2. Generate $u_2 \sim uniform(0, 1)$

STEP 3. If $u_2 \leq min\left(\frac{L(Y_T|w_n)}{L(Y_T|w_o)}, 1\right)$, return w_n, else return w_o and store the value of conditional distribution evaluated at w_o. Since uniform Dirichlet distribution is considered as prior, it disappears from the formula.

STEP 4. Repeat the above three steps M times (call it iteration cycle 1) and name the weights as w_0^* with the highest conditional distribution value.

STEP 5. Repeat the above 4 steps N times (call it iteration cycle 2) with $w_0 = w_0^*$ in each iteration. Stop once the value of conditional distribution has converged and use w_0^* stored in the Nth iteration as $\bar{w_T}$.

The value N in iteration cycle 2 can be decided based on how much the maximum conditional distribution value changes after every M iteration in iteration cycle 1. Similarly, the number of iterations M in iteration cycle 1 is decided based on the trade-off between exploring the solution space and computational time.

Appendix B: Asymptotic Properties

Under the M-closed case, when the true model (let's say D) is part of the set of available models, the opinion pool degenerates to the true model since all the weight is allotted to it (Geweke and

Amisano (2011)). This situation rarely arrives in real life, and D is generally unknown to the forecaster and the decision maker. The weights become relevant under the M-Open case when D is not part of the set of available models. In that case, the true weights (let's say $w^0 = \{w_1^0, w_2^0, \ldots, w_K^0\}$) can be interpreted as the ones which give the minimum Kullback-Leibler divergence from D to the opinion pool. Hall and Mitchell (2007) showed that the opinion pool optimized based on log predictive score minimizes the Kullback–Leibler directed distance from the data generating process to the prediction model. For K prediction models, the log prediction score for an opinion pool for $w_T = \{w_{1,T}, w_{2,T}, \ldots, w_{K,T}\}$ where $w_{k,T} \geq 0 \ \ \forall \ \ k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} w_{k,T} = 1$ for a given period t will look like

$$\begin{aligned} l(w_T|Y_T) &= \sum_{t=1}^{T} log\Big(\sum_{k=1}^{K} w_{k,T}\, p(y_t|Y_{t-1}, M_k)\Big) \\ &= \sum_{t=1}^{T} l(w_T|Y_t) \end{aligned} \tag{B.1}$$

One of the advantages of the log prediction score is that it is closely related to the likelihood function, which can be seen in the relation $l(w_T|Y_T) = log(p(Y_T|w_T))$. Geweke and Amisano (2011) showed that the weights obtained from optimizing $l(w_T|Y_T)$ asymptotically minimizes the Kullback-Leibler distance from the true model D.

$$w_T^* = arg\ \ max_w\ \ l(w_T|Y_T) \overset{a.s.}{\rightarrow} arg\ \ max_w\ \ l(w|Y) = w^0 \tag{B.2}$$

where, $\frac{1}{T}\sum_{t=1}^{T} l(w_T|Y_t) = \bar{l}(w_T|Y_T) \overset{a.s.}{\rightarrow} l(w|Y)$. Using this result, the posterior distribution of weights can be rewritten as

$$\begin{aligned} p(w_T|Y_T) &\propto p(Y_T|w_T)p(w_T) \\ &\propto exp\{log(p(Y_T|w_T))\}p(w_T) \\ &\propto exp\{\sum_{t=1}^{T} l(w_T|Y_t)\}p(w_T) \\ &\propto exp\{T\bar{l}(w_T|Y_T)\}p(w_T) \end{aligned} \tag{B.3}$$

As T increases, the exponential term dominates, and the effect of the prior, which does not depend on T, becomes relatively smaller. To analyse the posterior distribution further, let's take a second-order Taylor series approximation of $l(w_T|Y_T)$ around w_T^*

$$\begin{aligned} l(w_T|Y_T) &\approx l(w_T^*|Y_T) - \frac{T}{2}(w_T - w_T^*)^2(-\bar{l}''(w_T^*|Y_T)) \\ &\approx l(w_T^*|Y_T) - \frac{T}{2v}(w_T - w_T^*)^2 \end{aligned} \tag{B.4}$$

where $\bar{l}''(w_T^*|Y_T) = \frac{1}{T}\sum_{t=1}^{T} l''(w_T^*|Y_t)$ and $v = [\bar{l}''(w_T^*|Y_T)]^{-1}$. The term with first-order derivative disappears as $l(w_T|Y_T)$ is maximized at $w_T = w_T^*$. The posterior distribution can be approximated as

$$p(w_T|Y_T) \propto exp\{-\frac{T}{2v}(w_T - w_T^*)^2\}p(w_T) \tag{B.5}$$

The first term is in the form of a normal distribution with mean w_T^* and variance $\frac{v}{T}$. In summary, the role of the prior distribution becomes relatively small in determining the posterior distribution when T is large. The posterior distribution converges to a degenerate distribution at w^0 as $T \longrightarrow \infty$ then $\frac{v}{T} \longrightarrow 0$ and $w_T^* \longrightarrow w^0$, and the posterior distribution is approximately normally distributed with mean w_T^*.