Ashoka University
Economics Discussion Paper 133

Regularized Opinion Pools for Density Forecasts under Bayesian Inspired Framework

November 2024

Parush Arora, Ashoka University

https://ashoka.edu.in/economics-discussionpapers

Regularized Opinion Pools for Density Forecasts under Bayesian Inspired Framework

Parush Arora*†

Abstract

The paper considers the efficient estimation of opinion pools with regularization in the Bayesian paradigm and extends their application to cases where the number of competing models exceeds the number of observations. A Bayesian-inspired formulation and estimation algorithm is proposed whose 1) conditional density accommodates any proper scoring rule and 2) different priors allow weight shrinkage towards equality, extreme weights or any combinations under the Lasso, Ridge and Entropy penalty. Specifically, the Dirichlet prior allows shrinkage towards extreme weights which is useful for model selection applications. The simulation study explores and identifies situations where average log score is highest for opinion pools under shrinkage towards equality or extreme weights. An application involving the Survey of Professional Forecasters demonstrates that the Bayesian opinion pool's inflation forecast competes well with the equal-weight aggregated inflation forecast post 2013.

Keywords: *Inflation Expectation, Model Averaging, Predictive Density, Scoring Rule.*
JEL: C11, C15, C53, E17, E37

1 Introduction and Motivation

Forecasters' outlook toward any predictive exercise is reflected in how they formulate, specify, and estimate their model. The model dynamics depend on how the forecaster perceives and incorporates uncertainty (Steel (2020)). As a result, several competing forecasts emerge for a given random variable. For a researcher, a forecast combination is an intuitive way to utilize all this information (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 regularized forecast combinations for density forecasts aggregated under the linear opinion pool (Stone (1961), Bacharach (1974)).

Let y_t be a random variable and f_{kt} be the forecast density for y_t by forecaster k at time $t = 1, \ldots, T$. The combined forecast, f_t, under the linear opinion pool framework is obtained as

$$f_t = \sum_{k=1}^{K} w_k f_{kt}, \tag{1.1}$$

*I am grateful to Ivan Jeliazkov, Fabio Milani and Yingying Lee for their invaluable guidance and encouragement. All errors are my own.

†Department of Economics, Ashoka University.

where w_k is the weight allotted to f_{kt}. For simplicity, the paper submerged w notation from f_t. The weights are updated recursively once y_t is realized. The weights are estimated with respect to the unit simplex constraint: $\sum_{k=1}^{K} w_k = 1$ and $w_k \geq 0 \ \forall \ k$, ensuring that Eq. 1.1 is an appropriate probability density function.

Eq. 1.1 has been estimated in the past using proper scoring rules (Gneiting and Raftery (2007)). Key contributions include Bates and Granger (1969), Degroot and Mortera (1991) Geweke and Amisano (2011), Opschoor et al. (2017) and Garratt et al. (2023). Estimation becomes challenging when the number of forecasters is large relative to the number of past observations of y_t (micronumerosity). This becomes a binding constraint, especially in time-series forecasting, where the frequency of observations limits the data length. Researchers have used penalized forecast combinations to deal with the small sample problem (Capistrán and Timmermann (2009), Elliott (2011), Conflitti et al. (2015), and Samuels and Sekkel (2017) among many others). Diebold et al. (2023) extends regularization to the mixture of density forecasts allowing for shrinkage toward equal weights.

This paper proposes to estimate linear opinion pools using the Bayesian inspired formulation and hence calls it the Bayesian Opinion Pool (BOP). The framework 1) is general to accommodate all scoring rules, 2) different penalties and 3) allows the opinion pool to be estimated when the number of forecasting densities exceeds the number of observations. The Normal prior truncated over simplex introduce ridge penalty, the Laplace prior truncated over simplex introduce Lasso penalty and the Dirichlet prior introduce entropy penalty. Though Normal and Laplace only allow shrinkage towards equality (or any pre-given combination), the Dirichlet prior also allows classical shrinkage where coefficients are pushed towards 0 making it attractive for model selection applications. The researcher under Dirichlet prior can choose to shrink the weights on the spectrum with one extreme of allotting equal weights to all the models to another where all the weights are allotted to the best model. 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. This makes BOP useful for applications related to model averaging and model selection.

Unlike the usual regularization of selecting few variables out of many, the weight's shrinkage towards equality is more useful for density aggregation given the simplex constraint. The idea behind regularization is to avoid overfitting and improve out of sample prediction which translates into equal weights for density combination. Equal weights provide insurance against bad forecasts and their performance is been found competitive with optimized weights (Hendry and Clements (2004) and Wallis (2005)). The aim of BOP in such settings is to find the right balance between exploiting past information through optimization and regularizing forecasts through shrinkage towards equality.

One strand of literature deviated from opinion pools when combining forecast densities. 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.

The paper uses BOP in an application involving the survey of professional forecasters (SPF) to improve inflation density forecast. The aggregated predictive density for the inflation rate is estimated and compared with the equal weights strategy (simple opinion pools or SOP) published by the Federal Bank of Philadelphia. The issue of ignoring past predictive accuracy information could have been tackled through scoring rules optimization, but due to low data frequency, it is infeasible to use optimized-based methods for any length of the training window. The inflation forecast obtained through the BOP at various levels of shrinkage competes well with the Federal Reserve Bank of Philadelphia's published SOP.

Section 2 gives a brief overview of traditional opinion pools. Section 3 covers the explanation, derivation and estimation algorithm of the Bayesian opinion pool. Section 4 presents the simulation study where the choice related to penalty through Dirichlet prior is explored along with BOP's predictive performance. Section 5 covers an empirical exercise involving the Survey of Professional Forecasters (SPF) where inflation forecast densities are combined using BOP and compared with the equal weights combination as published by the Federal Bank of Philadelphia. Section 6 concludes the paper.

2 Proper Scoring Rules for Opinion Pools

This section summarizes the asymptotic properties of opinion pools optimized using proper scoring rules. Let θ_0 be the true vector of parameters. The researcher aims to fit a parametric model p_θ based on a sample $y_1, \ldots, y_T$. Let any proper scoring rule be presented as $S(\cdot, \cdot)$. Gneiting and Raftery (2007) showed that asymptotically

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

If the constraints on weights in Eq. 1.1 are satisfied the opinion pool satisfies the conditions of an appropriate probability distribution. Then asymptotically,

$$arg \ \max_{w} \frac{1}{T} \sum_{t=1}^{T} S(f_t, y_t) \longrightarrow w_0 \quad as \quad T \longrightarrow \infty. \tag{2.2}$$

where $w = \{w_1, \ldots, w_K\}$ is the parameter of interest. Bernardo and Smith (2000) considered three possible scenarios in the context of model averaging. First is the M-closed case where M_0, the true model, is identified and available in the model list. In this case, opinion pools will converge to M_0 asymptotically and $w_0 = \{1, 0, \ldots, 0\}'$ 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 (M-complete case). The third one is the M-open case, the most applicable and is considered in this paper, is when M_0 is not part of the model list. In this case, w will converge to some weight vector $w_0 = w^*$, which is related to the properties of the metric implied by the scoring function. For example, the log score minimizes the Kullback–Leibler divergence from M_0 to the opinion pool (Gneiting and Raftery (2007)).

Elliott et al. (2016) argued that there is no natural choice for choosing the scoring rule under the M-open case. The current paper considers log (l), quadratic (q), spherical (s), CRPS (c), and FTMS (m) rules (Brier (1950); Good (1952); Roby (1965); Shuford et al. (1966); Winkler and Murphy (1968); Epstein (1969); Selten (1998); Dawid and Sebastiani (1999)). The scores to opinion pool are provided as follows

$$\begin{aligned}
l(f_t, y_t) &= log(f_t) \\
q(f_t, y_t) &= 2p(f_t) - \int_{-\infty}^{\infty} f_t^2 dy_t \\
s(f_t, y_t) &= \frac{f_t}{(\int_{-\infty}^{\infty} f_t^2 dy_t)^{0.5}} \\
c(f_t, y_t) &= -\int_{-\infty}^{y_t^r} F_t^2 dy_t - \int_{y_t^r}^{\infty} (F_t - 1)^2 dy_t \\
m(f_t, y_t) &= -\Big(\frac{y_t - \mu}{\sigma}\Big)^2 - log(\sigma^2),
\end{aligned} \tag{2.3}$$

where μ is the mean, σ is the standard deviation and F_t is the cumulative predictive density of opinion pool. The scoring rules are altered to have higher scores implying improved forecast performance. Gneiting and Raftery (2007) discusses in detail the divergence or distance, different scoring rules are minimizing. For example, The log and quadratic (or Brier) scores minimize the Kullback–Leibler divergence and squared Euclidean distance between true DGP and predictive model, respectively. Dawid and Sebastiani (1999) suggested four proper scoring rules based on the first two moments of the predictive distribution, and $m(f_t, y_t)$ is among the popular ones. Given the decision maker has access to $\{f_{1t}, \ldots, f_{Kt}\}$, and $\{y_1, \ldots, y_T\}$, the opinion pool is estimated as

$$w^* = arg \ \max_{w} \sum_{t=1}^{T} S\Big(\sum_{k=1}^{K} w_k \, f_{kt}\Big), \tag{2.4}$$

where $w^* = \{w_1^*, \ldots, w_K^*\}$. The opinion pool for the prediction of y_{T+1} will take the form

$$f_{T+1} \quad = \quad \sum_{k=1}^{K} w_k^* \, f_{k,T+1}. \tag{2.5}$$

3 Bayesian Opinion Pool

This section lays out the estimation procedure for opinion pools under the Bayesian-inspired formulation. To obtain the posterior density of weights given data, the Bayes theorem is utilized and is given as

$$\pi(w|y) \quad \propto \quad \pi(y|w)\pi(w)$$

where $\pi(w|y)$ is the posterior distribution of weights, $\pi(y|w)$ is the likelihood function and $\pi(w)$ is the prior distribution of weights. The paper considers the following representation of the likelihood function

$$\pi(y|w) = \frac{1}{C_1} \prod_{t=1}^{T} e^{S(f_t, y_t)}$$

where C_1 is the normalizing constant for $\pi(y|w)$. The paper will call $\pi(y|w)$ as conditional density since it does not represent the researcher's imposed data-generating process, which is the intuition for the likelihood function. This Bayesian-inspired methodology treats weights as a K-dimensional, simplex bound, random variable and thus, the paper considers truncated Laplace (l), truncated Normal (n) and Dirichlet (d) priors to complete the formulation. These priors induce different regularizations on weights leading to the posterior modes overlapping with the optimized weights derived by Diebold et al. (2023).

3.1 Laplace Prior truncated on a Simplex

The Laplace prior truncated on a simplex is given as

$$\pi_l(w) \quad = \quad \mathbb{1}(w \in B) \frac{1}{C_2} \prod_{k=1}^{K} \frac{\lambda_l}{2} e^{-\lambda_l |w_k - w_{lk0}|}$$

where C_2 is normalizing constant, w_{lk0} is the location hyperparameter, $\lambda_l > 0$ is the scale hyperparameter and

$$B = \Big\{ w | \sum_{k=1}^{K} w_k = 1, w_k \geq 0 \, \forall \, k \Big\}.$$

The posterior distribution will take the form

$$\pi_d(w|y) \propto \frac{1}{C_1} \prod_{t=1}^{T} e^{S(f_t, y_t)} \mathbb{1}(w \in B) \frac{1}{C_2} \prod_{k=1}^{K} \frac{\lambda_l}{2} e^{-\lambda_l |w_k - w_{lk0}|}.$$

Dropping terms which do not depend on w, the posterior mode is

$$w_l \quad = \quad \arg\max_{w} \Bigg(\sum_{t=1}^{T} S(f_t, y_t) - \underbrace{\lambda_l \sum_{k=1}^{K} |w_k - w_{lk0}|}_{L^1 \text{ lasso penalty}} \Bigg) \qquad \text{s.t } w \in B$$

Park and Casella (2008) used Laplace prior to estimate Lasso regression under the Bayesian framework. Thus, this paper uses Laplace prior truncated over simplex for opinion pools which imposes L^1 lasso penalty on weights while satisfying the simplex constraints. The magnitude of λ_l will detrmine the strength of penalty. If $w_{lk0} = \frac{1}{K} \ \forall \ k$, the weights will be shrunk towards equality.

3.2 Normal Prior truncated on a Simplex

The Normal prior truncated on a simplex is given as

$$\pi_n(w) \quad = \quad \mathbb{1}(w \in B)\frac{1}{C_3} f_N(w|w_{n0}, I_K \frac{\lambda_n^{-1}}{2})$$

where f_N is multivariate normal distribution, C_3 is normalizing constant and $w_{n0} = \{w_{n10}, \ldots, w_{nK0}\}$ and $\frac{\lambda_n^{-1}}{2}$ are mean and variance of f_N respectively. The posterior distribution will take the form

$$\pi_d(w|y) \propto \frac{1}{C_1} \prod_{t=1}^{T} e^{S(f_t, y_t)} \mathbb{1}(w \in B)\frac{1}{C_3} f_N(w|w_{n0}, I_K \frac{\lambda_n^{-1}}{2})$$

Dropping terms which do not depend on w, the posterior mode is

$$w_n \quad = \quad \arg\max_w \left(\sum_{t=1}^{T} S(f_t, y_t) - \underbrace{\lambda_n \sum_{k=1}^{K} (w_k - w_{nk0})^2}_{L^2 \text{ ridge penalty}} \right) \qquad \text{s.t } w \in B$$

The well-known Bayesian ridge regression uses Normal prior over coefficients to induce L^2 penalty. A Normal prior truncated on a simplex for weights induces the same penalty for opinion pools. As was the case with Laplace prior, the magnitude of λ_n determines the strength of penalty and keeping $w_{nk0} = \frac{1}{K} \ \forall \ k$ leads to the weights being shrunk towards equality.

3.3 Dirichlet Prior

The Dirichlet prior is given as

$$\pi_d(w) \quad = \quad \frac{1}{B(\alpha)} \prod_{k=1}^{K} w_k^{\alpha_k - 1}$$

where α_k is a hyperparameter of Dirichlet prior. The paper assumes no prior information about forecasters and thus $\alpha_k = \alpha \ \forall \ k$. The posterior distribution will take the form

$$\pi_d(w|y) \propto \frac{1}{C_1} \prod_{t=1}^{T} e^{S(f_t, y_t)} \frac{1}{B(\alpha)} \prod_{k=1}^{K} w_k^{\alpha - 1}$$

Dropping terms which do not depend on w, the posterior mode is

$$w_d \quad = \quad \arg\max_w \left(\sum_{t=1}^{T} S(f_t, y_t) - \underbrace{\lambda_d \sum_{k=1}^{K} \log(w)}_{\text{entropy penalty}} \right) \qquad \text{s.t } w \in B$$

where, $\lambda_d = 1 - \alpha$ and $-\infty < \lambda_d \leq 1$ since $0 \leq \alpha < \infty$. Diebold et al. (2023) showed that their simplex+entropy regularized estimator coincides with posterior mode under Bayesian analysis when Dirichlet prior and log score

conditional density is considered. If $\lambda_d < 0$ (or $\alpha > 1$), the prior shrinks the weights towards equality. As $\lambda_d \to -\infty$, the tendency of weights towards equality grows stronger. If $\lambda_d = 0$ (or $\alpha = 1$), the prior is uniform but still imposes mild shrinkage towards equality since the prior's mean is $\frac{1}{K}$. The unique property about Dirichlet prior is the allowance for standard shrinkage of weights towards 0 which the L^1 or L^2 penalty does not allow. This adaptability is not been explored in the density combination literature. If $0 < \lambda_d \leq 1$ (or $0 \leq \alpha < 1$), the prior incentivizes extreme weights for some models. As $\lambda_d \to 1$, the tendency of weights towards choosing the best model increases. This is useful in case the application requires model selection.

3.4 Estimation

Since the final form of the posterior is non-standard, the paper uses the Metropolis-Hasting (MH) algorithm to draw from the posterior density. When K is small, a uniform proposal density like the Dirichlet distribution with $\lambda_d = 0$ will be able to cover the whole parameter space. Although, When K is large, the acceptance rate associated with any uniform proposal may be too low due to high dimensionality. Alternatively, the paper explores a tailored proposal density where the vector of weights are transformed to be defined on an unbounded domain using a multivariate logit transformation. Given $\theta = \{\theta_1, \ldots, \theta_{K-1}\}$, the transformation will look like

$$\theta_k = \ln(\frac{w_k}{w_K}) \tag{3.1}$$

for all $k = 1, \ldots, K-1$. The draws are sampled from a tailored proposal normal density as $\theta \sim N(\bar{\theta}, \bar{\Omega})$. The mean of the Gaussian proposal, $\bar{\theta}$ is the mode of the $p(y|w)$. In case of micronumerosity, where numerical optimization fails, the mode is calculated using a back-fitting MCMC algorithm (details can be found in the Appendix). The covariance matrix, $\bar{\Omega}$ can be kept equal to either σI_{K-1} where σ is decided based on the rejection rate or $\bar{\Omega}$ is proportional to the inverse Hessian of the conditional density at $\bar{\theta}$. Let $\theta^{(g)}$ be θ drawn in the g^{th} iteration. The MCMC estimation of the BOP for the transformed proposal is summarized in the following steps. Let $w^{(g)}$ be w drawn in the g^{th} iteration.

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

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

STEP 3. Transform $\theta^{(g)}$ to obtain $w^{(g)}$.

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

STEP 5. If

$$u \leq \min\left(\frac{\pi(w^{(g)}|y)q(w^{(g-1)})}{\pi(w^{(g-1)}|y)q(w^{(g)})}, 1\right),$$

return $w^{(g)}$, else return $w^{(g-1)}$. 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 framework is not restrictive to one-step-ahead forecast and can be extended for long horizons forecasting densities like $f_{k,t+h}$.

4 Monte Carlo

The current simulation study tests logarithmic, quadratic and CRPS score conditional density with Dirichlet, Normal and Laplace priors.

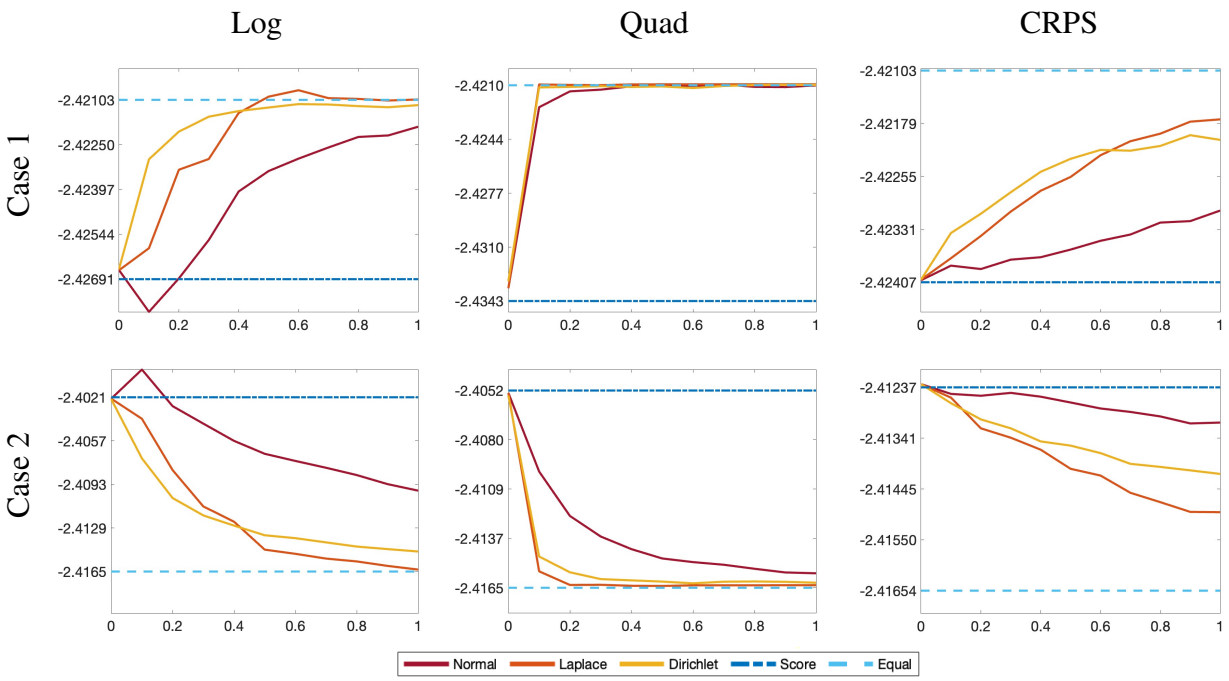

Figure 1: Average Log Score as a function of Penalty for Bayesian Opinion Pools for $T = 3$

4.1 Data-generating process and forecasts

The DGP and individual models are considered under the linear setting to preserve useful insights that might get lost in a complicated analysis. Let the variable of interest be z_t whose DGP is given as

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

Three individual forecasters submit their predictive densities for z_t as $N(z_{kt}, 4)$ where

Case 1: $z_{kt} \sim N(z_t, 4) \;\forall\; k = 1, 2$ and 3.
Case 2: $z_{1t} \sim N(z_t, 2), z_{2t} \sim N(z_t, 4)$ and $z_{3t} \sim N(z_t, 6)$.

The forecasters under the case 1 predict with equal accuracy whereas Case 2 has a clear ordering where forecaster 1 is the most accurate and forecaster 3 is the least. The opinion pool is trained using the sample size (T) of $3, 5$ and 10 which tests the cases of near-micronumerosity and small sample. The forecasts are calculated one step ahead and the exercise is repeated 500 times. The predictive exercise uses the rolling window approach.

4.2 Shrinkage Towards Equality

Figure 1 shows the sensitivity of the average log score to penalty for $T = 3$ for different scoring rule likelihood functions and priors (check Fig. 9 for $T = 10$ in the appendix). Sub-figures in the first, second and third column represent logarithmic, quadratic, and CRPS scoring rules, respectively. Sub-figures in the first and second row represent Case 1 and Case 2 of DGP, respectively. Each figure presents the average log score for Bayesian opinion pools under Dirichlet, Normal, and Laplace priors along with equal weights and respective score optimized weights. Uniform or no penalty is imposed when $\lambda_d = 1$, $\lambda_n = 0$ and $\lambda_l = 0$. The uniform penalty for Dirichlet is rescaled to start at 0 rather than 1 to align with Normal and Laplace.

For Case 1, the log score for equal weights is higher since all forecasters are equally competitive. On the contrary, the log score for score optimized weights is higher for Case 2 since forecasters vary in terms of their

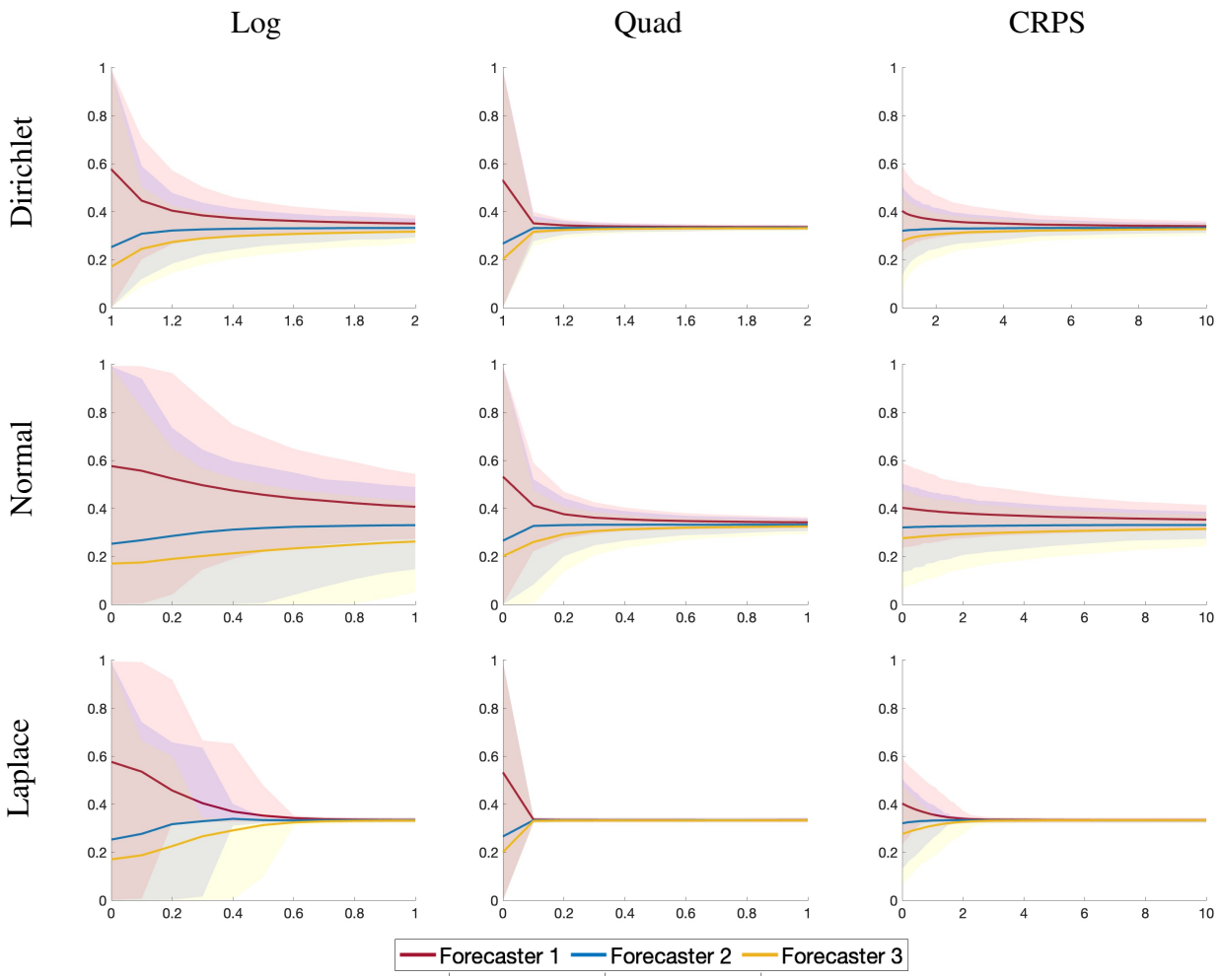

Figure 2: Weights as a function of Penalty for Bayesian Opinion Pools under Case 2 and $T = 3$

predictive accuracy. The opinion pools start from the score optimized value as we normalized 0 to represent uniform prior. As the penalty increases, the average log score converge towards equal weights due to stronger shrinkage toward equality. As the sample size increase, one can expect the log scores for equal and score optimized opinion pools to converge.

The convergence from score optimized to equal weights is fastest in quadratic score, followed by log and CRPS. Within each figure, Laplace shows the faster convergence followed by Dirichlet and Laplace. To see how sensitive the weights are to penalty, it will be interesting to test them under case 2 since they are expected to be unequal. Figure 2 shows the mean of the posterior modes of weights in 500 samples for Case 2 and $T = 3$. Sub-figures in the first, second and third column represent logarithmic, quadratic, and CRPS scoring rules, respectively. Sub-figures in the first, second and third column represent Dirichlet, Normal and Laplace prior, respectively. The opinion pools are able to correctly rank the forecasters, but the distance between weights for uniform prior depends on the combination of score function and prior. The weights for logarithmic and quadratic scores converge within 1 unit penalty increment irrespective of the prior. The convergence is relatively slow for CRPS, which was also observed in Fig. 1. Also, the Bayesian opinion pool does not weigh forecaster 1 heavily, since the sample does not have enough information due to its small size. As the sample size increases, it can be expected that the likelihood function will dominate, leading to forecaster 1 being assigned a higher weight (Figure 10 for $T = 10$ in the appendix shows a greater weight allotted to forecaster 1 at a uniform prior).

The shrinkage is strong when T is small as the prior dominates due to the insufficiency of information in the

conditional density. As T increases, the conditional density starts to dominate and the weights deviate from the equal weights. This property allows BOP to be used in applications related to model averaging. As α_k tends to infinity, BOP tends to the simple opinion pool (opinion pool with equal weights).

4.3 Standard Shrinkage towards 0

Forecaster selection in density combination is a challenging task, as the simplex constraint introduces natural shrinkage toward equal weights. Diebold et al. (2023) discussed the difficulty of introducing standard regularization in the density aggregation problem, since equal weights are as close to 0 as one can get while maintaining the sum-to-one restriction. The Dirichlet prior enables shrinkage towards extreme weights, and thus can be useful in application related to model selection. Figure 3 shows the mean of the posterior modes of weight in the 500 samples for Case 2 with $0 < \lambda_d < 1$. The subfigures in the first, second and third columns represent logarithmic, quadratic, and CRPS scoring rules, respectively. The subfigures in the first, second and third columns represent $T = 3, 5$ and 10, respectively. The weights under logarithmic and quadratic likelihood are sensitive to penalty value and immediately shrink towards choosing forecaster 1 whereas CRPS shows slower convergence. A similar pattern is observed for all $T = 3, 5$ and 10.

Figure 4 presents the average log score as a function of the penalty for Dirichlet prior for case 2. The subfigures in the first, second and third columns represent logarithmic, quadratic, and CRPS scoring rules, respectively. The subfigures in the first, second and third columns represent $T = 3, 5$ and 10, respectively. As the forecasters have a clear ordering based on predictive accuracy, standard shrinkage leads to increase in average log score. The opinion pool degenerates into the predictive density of forecaster 1. This setup highlights the usefulness of density selection when the gap between the predictive accuracy of the forecasters is significant.

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" (P.3).

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 it possible to track them. Anonymity is maintained to prevent strategic misreporting. The details of the data set and its significance can be found in Croushore et al. (2019), Clements et al. (2023) or on the Federal Reserve Bank of Philadelphia 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.

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 densities are standardized and take the form of a histogram. Engelberg et al. (2009) fit continuous densities to the individual surveyed histograms to undo discretization. However, this interprets the survey replies as some subjective continuous distribution that is present in the minds of individual forecasters (Kenny et al. (2015)). Moreover, it imposes distributional assumptions which may pose practical challenges given that individual histograms are often restricted to very few bins. Thus, this paper uses linear interpolation and assumes uniform probability mass within the bins which is implicit in the assumption of a histogram (Clements (2002)).

SPF is used practically for two purposes. First, it is used to estimate inflation expectations. Inflation forecasts are integral to many macroeconomic models as they are used to estimate 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)). Keane and Runkle (1990) argue that a model with rational agents can be better represented using the predictive data from SPF. Coibion et al.

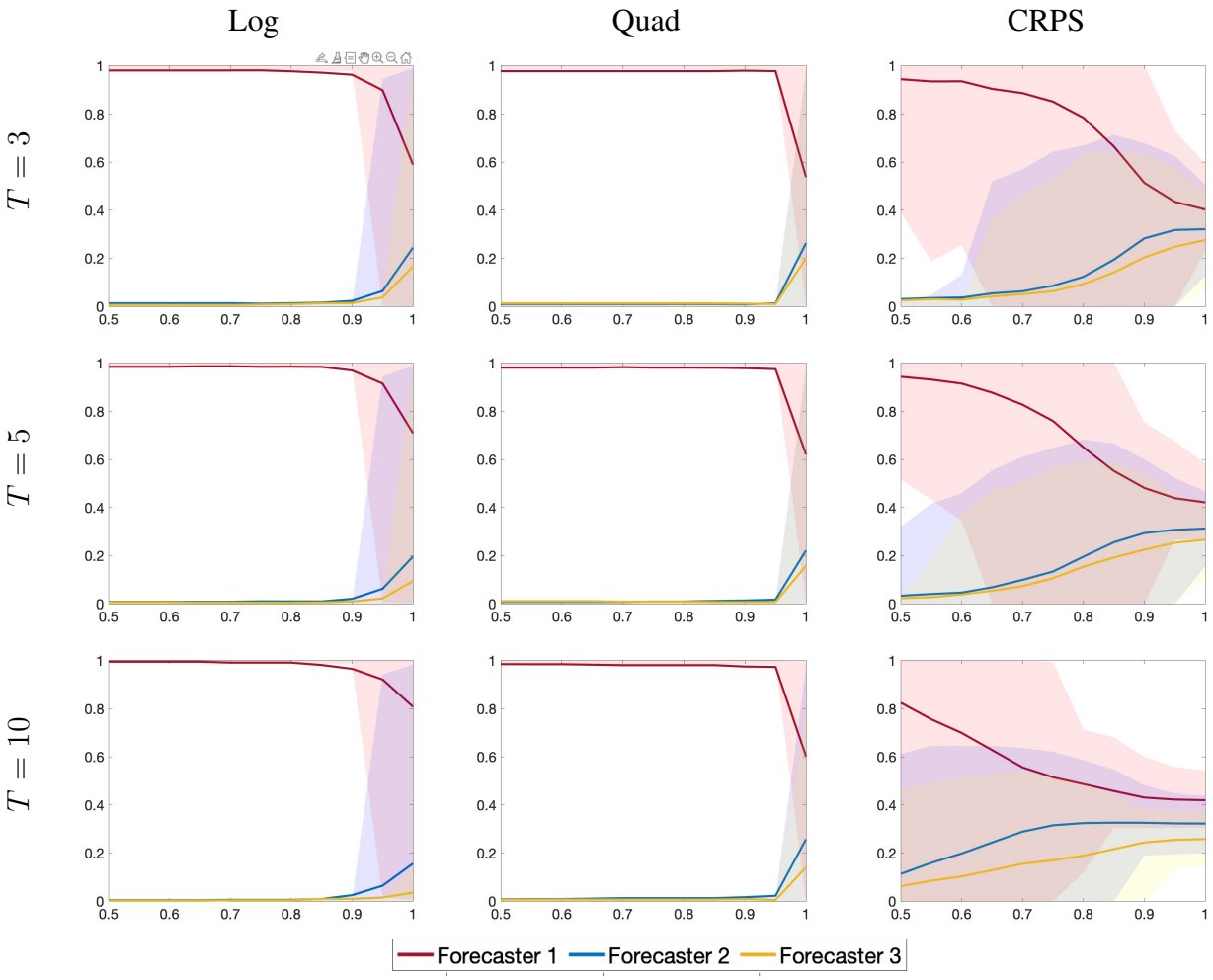

Figure 3: Weights under Dirichlet prior as a function of Penalty under Case 2

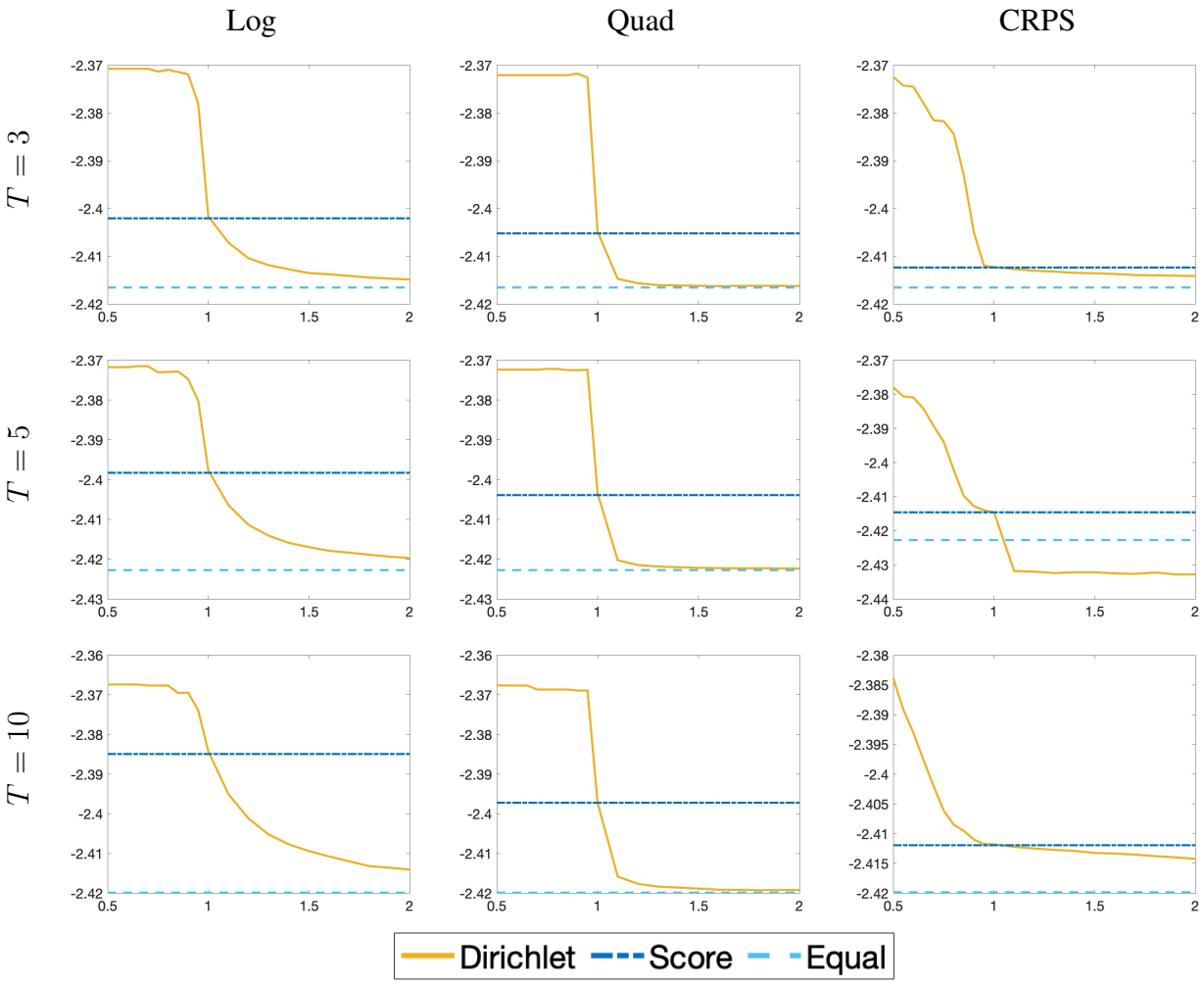

Figure 4: Average Log Score as a function of Penalty for Dirichlet Prior for Case 2

(2018) referred to SPF extensively and argued for improved models that rely on variables with expectations. 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)). Carroll (2003) evaluated the influence of SPF data on private-sector expectations.

Second, SPF is used to forecast inflation accurately or test forecasting models. This facilitates decisions requiring accurate inflation predictions (for example, setting wage contracts). Smets et al. (2014) incorporated SPF data to measure the forecasting accuracy of New Keynesian DSGE models. Forecasts based on the neural networks and several linear econometric models were compared to SPF data (Croushore (1993)). Swanson and White (1997) used model selection on multiple non-linear models and found that no one model was able to consistently beat SPF forecasts. Croushore et al. (2019) mentioned in their paper that "The SPF has become the gold standard for evaluating forecasts or comparing forecasting models" (P.5).

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 how the economy expects inflation to be. 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 window length.

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" (P.10). The issue with the simple opinion pools (SOP) approach is that it does not exploit the information about the past predictive performance of the experts. Figure 5 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 the forecast. The horizontal axis represents the unique ID of experts. The size of the points represents the number of quarters, an expert was active in the past. The figure depicts that some experts 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 for log score likelihood function with $\lambda_k = \{0.25, 0.5, 0.75, 1, 2, 5\}$ where $k = d, n$ and l for Dirichlet, Normal and Laplace priors respectively. The decision to choose penalty which shrinks weights towards equality 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 prior shrinks the BOP towards SOP but still allows deviations in case strong evidence for better relative predictive accuracy is present. The paper also explores shrinkage towards extreme weights by keeping $\lambda_d < 1$.

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 to deal with the unbalanced panel .

- 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

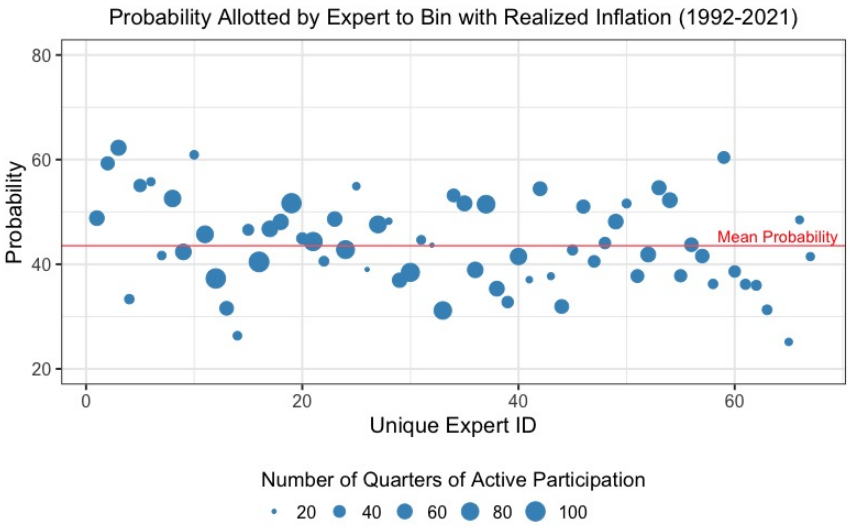

Figure 5

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 choose to include the expert only if they have participated for a certain number of quarters (Conflitti et al. (2015) used 5 quarters of data).

- 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.

- Partial availability: Suppose a forecaster submits the prediction for the $(m+1)^{th}$ quarter but was available in s periods out of the m training period where $s < m$. The weights associated with the forecaster will enter the joint conditional density (Eq. B.6) in the periods where they were available (total of s times). Thus, the methodology rewards consistency as the forecaster with active participation will have a greater influence on the opinion pool density than an inactive one.

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 21^{st} 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 the weights for 10 active experts is 1.

The paper considers the moving windows approach with 24 quarters (6 years) of training data and the rest of the period until 2020 Q1 (pre-Covid) as testing data (one step out of sample prediction). The paper identifies pre 2013 as the period where SOP performs better than BOP and post 2013 as the period when BOP performs better than SOP, especially for Dirichlet and Normal priors.

Figure 6 shows the log score difference between BOP and SOP for Normal and Laplace prior for pre and post 2013 period. The penalty term λ_k takes the values $0.25, 1$ and 5 for $k = n$ or l where 0.25 represents weak shrinkage and 5 represent strong shrinkage. The difference is normalized to 0 and thus the vertical line at the origin is represented by SOP (equal weights). The upper sub-figures shows the log score difference for Normal prior for different values of shrinkage, which is, on average, worse than SOP for pre 2013 and better than SOP in post 2013. This is confirmed in Tab. 1 and Tab. 2 where the average log score difference for Normal is negative for pre 2013 and positive for post 2013. The log score for normal is not sensitive to the penalty values chosen which hints towards increasing them. For Laplace prior, the performance ordering is not clear in the bottom sub-figures of Fig. 6. Compared to Normal, the log score of Laplace responds to change in penalty strength though there

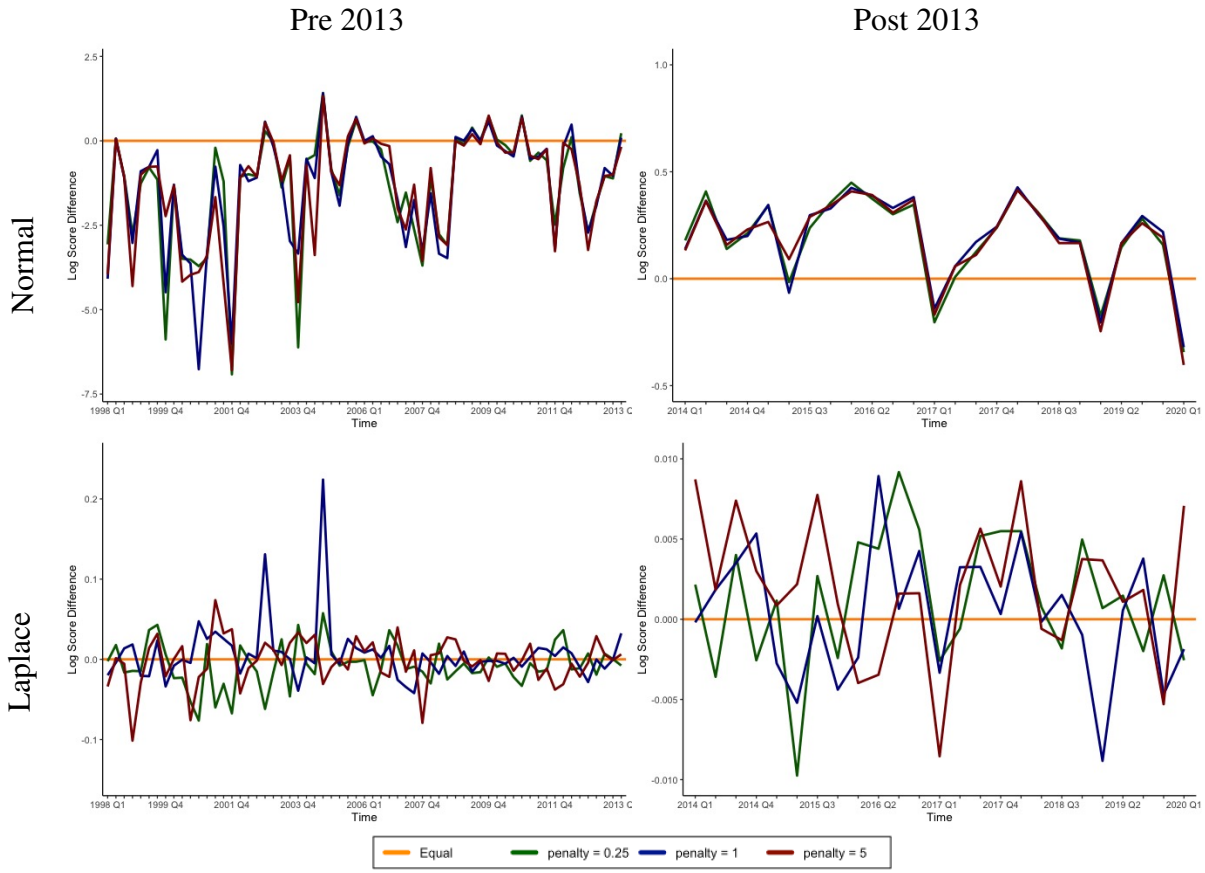

Figure 6: Log Score Difference between BOP (Log) and SOP for Normal and Laplace

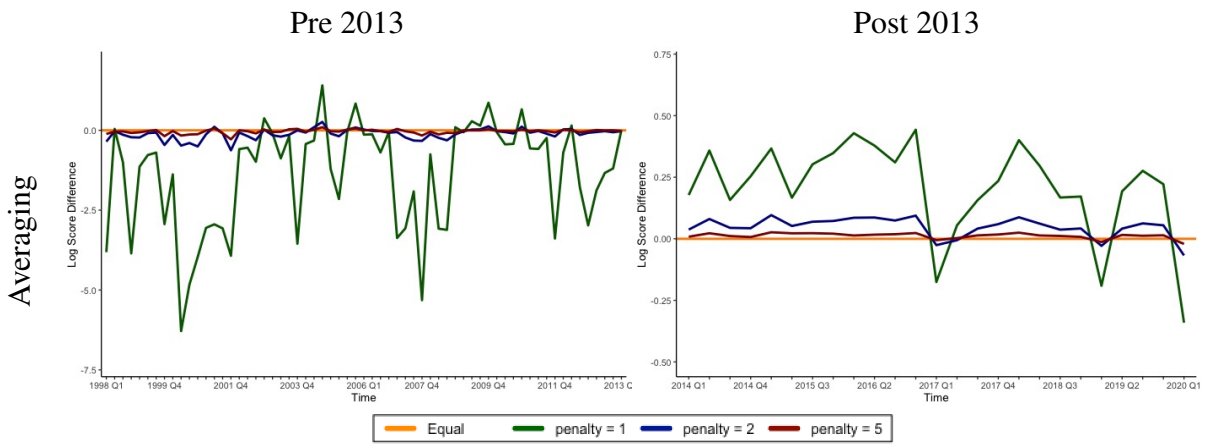

Figure 7: Log Score Difference between Dirichlet BOP (Log) and SOP when $\lambda_d \geq 1$

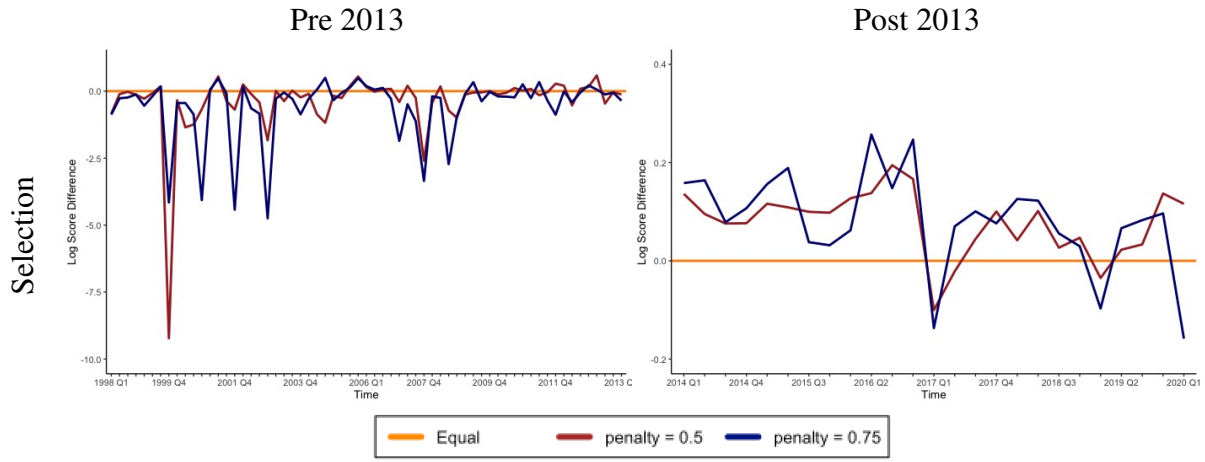

Figure 8: Log Score Difference between Dirichlet BOP (Log) and SOP when $\lambda_d < 1$

functional relation is not obvious. In Tab. 1 and Tab. 2, the average log score difference for Laplace is negative for pre 2013 and positive for post 2013.

Figure. 7 shows the log score difference between BOP and SOP for Dirichlet prior for pre and post 2013 period when $\lambda_d = 1, 2$ and 5. This setting explores the shrinkage of weights towards equality as analysed in Fig. 6. Similar to Normal, Dirichlet performs worst than SOP in pre 2013 and better than SOP in post 2013. The log score for Dirichlet converges towards log score of SOP as λ_d increases.

Figure. 8 shows the log score difference between BOP and SOP for Dirichlet prior for pre and post 2013 period when $\lambda_d = 0.5$ and 0.75. This setting explores the shrinkage towards extreme weights. The pattern is similar to what was observed for $\lambda_d = 1$. Dirichlet prior with $\lambda_d \leq 1$ shows significant improvement in predictive accuracy for post 2013 period compared to $\lambda_d > 1$. Keeping $\lambda_d \leq 1$ turns out to be high risk high reward strategy considering the two periods data is divided into. In Tab. 1 and Tab. 2, the average log score difference for Dirichlet for any value of λ_d is negative for pre 2013 and positive for post 2013.

Table 1: Pre 2013 Average Log Score Difference between BOP and SOP

λ	Dirichlet	Normal	Laplace
0.25	-0.358	-1.309	-0.008
0.5	-0.383	-1.241	-0.001
0.75	-0.578	-1.329	-0.008
1	-1.382	-1.298	-0.009
2	-0.116	-1.352	0.006
5	-0.035	-1.361	-0.002

Table 2: Post 2013 Average Log Score Difference between BOP and SOP

λ	Dirichlet	Normal	Laplace
0.25	0.071	0.186	0.001
0.5	0.078	0.188	0.001
0.75	0.083	0.194	0.001
1	0.206	0.200	0.002
2	0.048	0.195	0.001
5	0.012	0.184	0.002

There could be couple of potential reasons for improved performance of BOP over SOP post 2013. First, since the training window of 6 years is used, predicting inflation in 2013 excludes the Great Recession from the training data which could be seen as a structural break. Second, with advancement in machine learning and predictive models, forecasters on average has become better in predicting inflation in the last decade. Further analysis is need cement these potentila factors as causes.

6 Conclusion

This paper provides a Bayesian inspired framework which is general enough to accommodate any scoring rule and penalty for estimation of regularized opinion pools considering low frequency nature of time series data. 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. While the discussion in the paper focused on macroeconomic time series data, the usefulness of the techniques can be extended to other applications like gambling, stock market, election polls etc. 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.

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.

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

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. (2002). An evaluation of the survey of professional forecasters probability distributions of expected inflation and output growth. *Manuscript, Department of Economics, University of Warwick*.

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.

Conflitti, C., C. De Mol, and D. Giannone (2015). Optimal combination of survey forecasts. *International Journal of Forecasting 31*(4), 1096–1103.

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

Croushore, D. D. (1993). Introducing: the survey of professional forecasters. *Business Review-Federal Reserve Bank of Philadelphia 6*, 3.

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., M. Shin, and B. Zhang (2023). On the aggregation of probability assessments: Regularized mixtures of predictive densities for eurozone inflation and real interest rates. *Journal of Econometrics 237*(2), 105321.

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

Elliott, G. (2011). Averaging and the optimal combination of forecasts. *University of California, San Diego*.

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.

Engelberg, J., C. F. Manski, and J. Williams (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics 27*(1), 30–41.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982) 8*(6), 985–987.

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*.

Garratt, A., T. Henckel, and S. P. Vahey (2023). Empirically-transformed linear opinion pools. *International Journal of Forecasting 39*(2), 736–753.

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

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.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological) 14*(1), 107–114.

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.

Kenny, G., T. Kostka, and F. Masera (2015). Density characteristics and density forecast performance: a panel analysis. *Empirical Economics 48*, 1203–1231.

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.

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.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the american statistical association 103*(482), 681–686.

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

Roby, T. B. (1965). Belief states and the uses of evidence. *Behavioral science 10*(3), 255–270.

Samuels, J. D. and R. M. Sekkel (2017). Model confidence sets and forecast combination. *International Journal of Forecasting 33*(1), 48–60.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics 1*(1), 43–61.

Shuford, E. H., A. Albert, and H. E. Massengill (1966). Admissible probability measurement procedures. *Psychometrika 31*(2), 125–145.

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.

Swanson, N. R. and H. White (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International journal of Forecasting 13*(4), 439–461.

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. and A. H. Murphy (1968). Evaluation of subjective precipitation probability forecasts. In *Proceedings of the first national conference on statistical meteorology*, pp. 148–157. American Meteorological Society Boston Boston, MA, USA.

Appendix A: Back-fitting MCMC algorithm

This subsection presents the MCMC algorithm used in the estimation of BOP. 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 $u_2 \leq min\Big(\frac{p(Y_T|w_n)}{p(Y_T|w_o)}, 1\Big)$, return w_n, else return w_o and store the value of conditional density 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 density 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 density has converged and use w_0^* stored in the N^{th} iteration as $\bar{w_T}$.

The value N in iteration cycle 2 can be decided based on how much the maximum conditional density 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. There is a possibility that $\bar{w_T}$ is not a global maximum. The paper suggests using the algorithm multiple times from different initial conditions to verify.

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. Gneiting and Raftery (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.

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

where, $\frac{1}{T}\sum_{t=1}^{T} l(w_T|Y_t) = \bar{l}(w_T|Y_T) \xrightarrow{a.s.} l(w|Y)$. Using this result, the posterior density 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 density 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 density 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 density becomes relatively small in determining the posterior density when T is large. The posterior density converges to a degenerate density at w^0 as $T \longrightarrow \infty$ then $\frac{v}{T} \longrightarrow 0$ and $w_T^* \longrightarrow w^0$, and the posterior density is approximately normally distributed with mean w_T^*.

Appendix C: Natural interpretation of Log Score under Bayesian framework

Given that the opinion pool itself is an appropriate probability distribution function, it makes sense to treat it as the joint conditional density (equivalent to the joint likelihood function) given as

$$\begin{aligned} p(Y_T|w_T) &= \prod_{t=1}^{T} p(y_t|Y_{t-1}) \\ &= \prod_{t=1}^{T} \Big(\sum_{k=1}^{K} w_{k,T}\, p(y_t|Y_{t-1}, M_k) \Big). \end{aligned} \tag{B.6}$$

The conditional density incorporates the past predictive performance of experts as it is defined as a sequence of one-step-ahead conditional densities from time 1 to T. Since each conditional density is a mixture generated by the weights which are not varying with respect to time, the weights are tied with past conditional densities.

Given the prior and the conditional densities, the posterior density of the weights takes the form

$$\begin{aligned} p(w_T|Y_T) &\propto p(Y_T|w_T)p(w_T) \\ &\propto \prod_{t=1}^{T} \Big(\sum_{k=1}^{K} w_{k,T}\, p(y_t|Y_{t-1}, M_k) \Big) \prod_{k=1}^{K} w_{k,T}^{\alpha_k - 1}. \end{aligned} \tag{B.7}$$

It is easy to see that the log score rule (optimal prediction pools by Geweke and Amisano (2011)) is a monotonic

transformation of the conditional density.

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

Therefore, the mode of the posterior density of weights will coincide with the weights under the optimal prediction pool asymptotically. The weights minimize the Kullback–Leibler divergence from DGP to the opinion pool since the prior disappears in a large sample. 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 it is not feasible to optimize the function under micronumerosity, the BOP with a uniform prior can be seen as an extension of the optimal prediction pool, broadening its applicability.

Appendix D: Extra Figures from Simulation Study

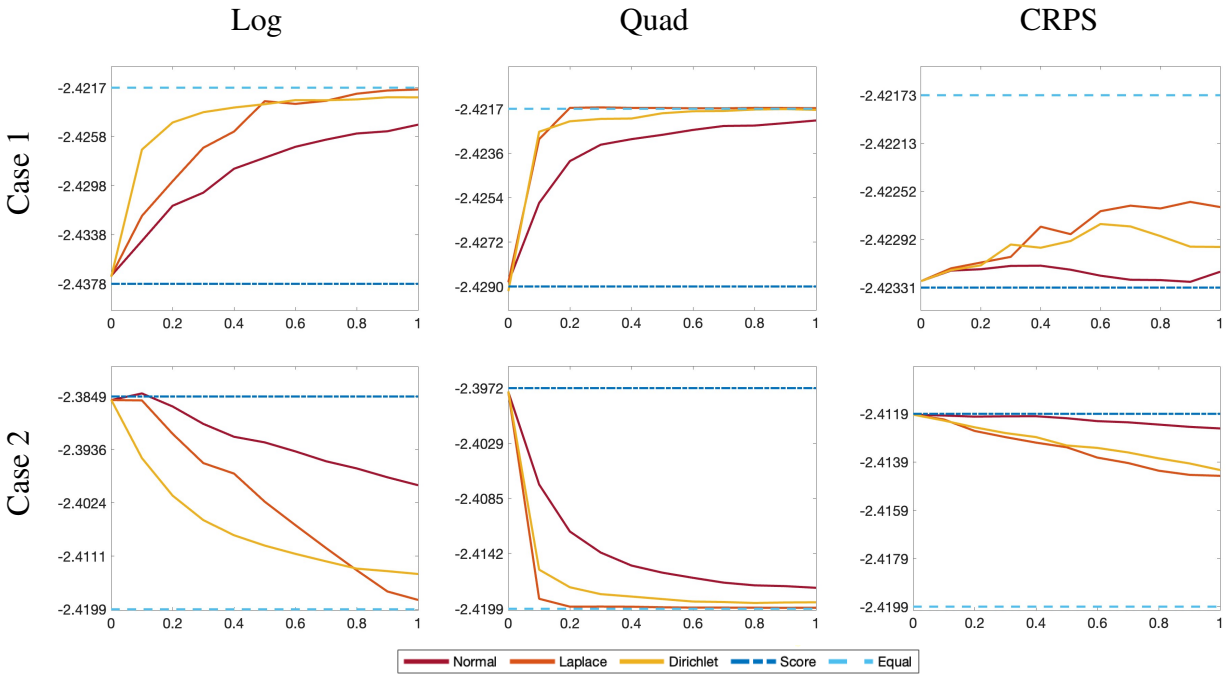

Figure 9: Average Log Score as a function of Penalty for Bayesian Opinion Pools for $T = 10$

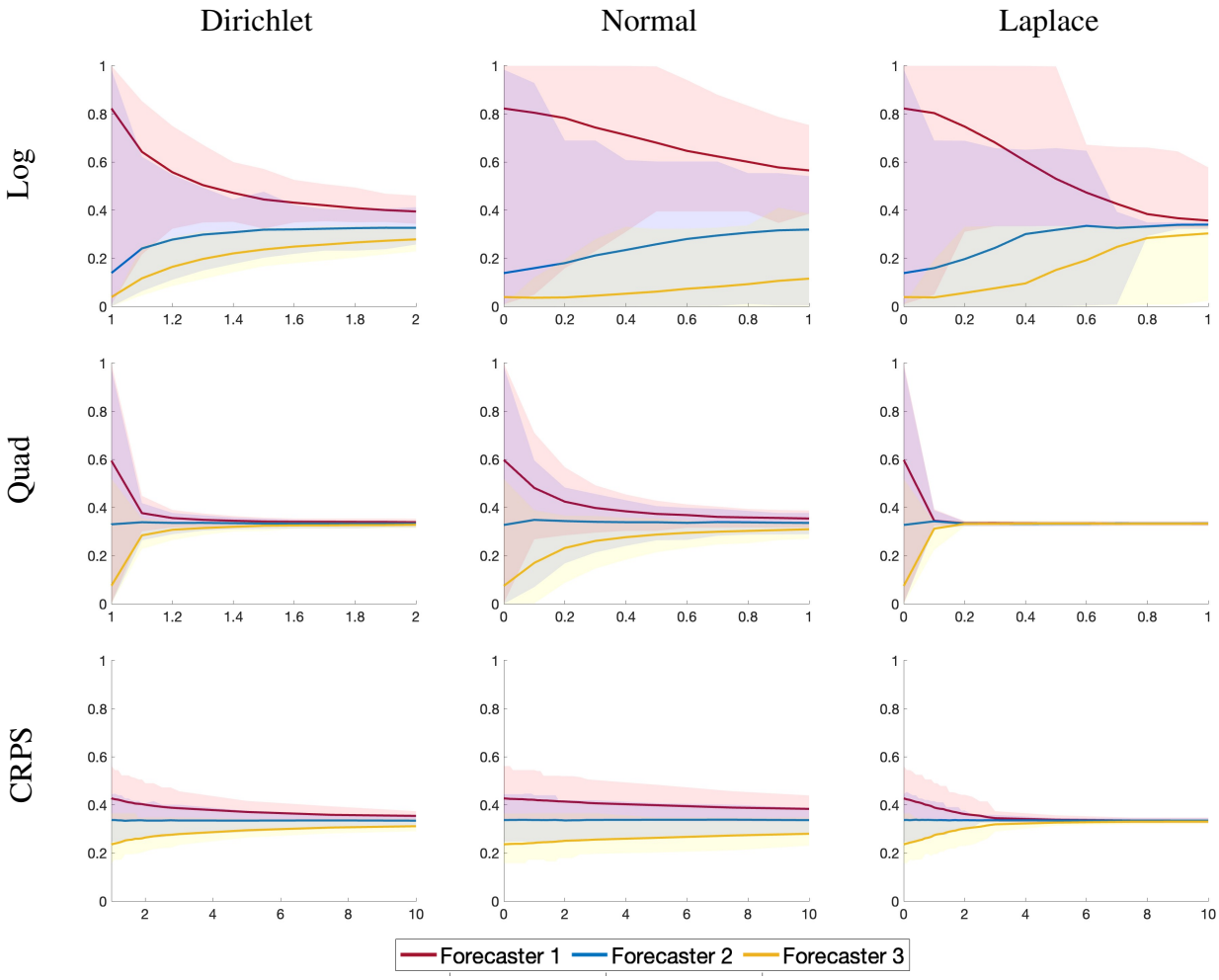

Figure 10: Weights as a function of Penalty for Bayesian Opinion Pools under Case 2 and $T = 3$