On the Aggregation of Probability Assessments: Regularized Mixtures of Predictive Densities for Eurozone Inflation and Real Interest Rates

Francis X. Diebold
University of Pennsylvania

Minchul Shin
Federal Reserve Bank of Philadelphia

Boyuan Zhang
University of Pennsylvania

January 2, 2022

Abstract: We propose methods for constructing regularized mixtures of density forecasts. We explore a variety of objectives and regularization penalties, and we use them in a substantive exploration of Eurozone inflation and real interest rate density forecasts. All individual inflation forecasters (even the ex post best forecaster) are outperformed by our regularized mixtures. From the Great Recession onward, the optimal regularization tends to move density forecasts' probability mass from the centers to the tails, correcting for overconfidence.

Acknowledgments: For guidance we are grateful to the editor and two referees. For helpful comments and/or assistance we are grateful to Umut Akovali, Brendan Beare, Graham Elliott, Rob Engle, Domenico Giannone, Christian Hansen, Nour Meddahi, Mike McCracken, Marcelo Medeiros, James Mitchell, Joon Park, Hashem Pesaran, Youngki Shin, Mike West, and Ken Wolpin. We are also grateful to conference participants at the 2020 EC^2 Meeting, the 2021 SoFiE Conference on Machine Learning in Finance, the 2021 SoFiE Annual meeting, and the 2021 NBER/NSF Time Series Conference, and to seminar participants at AMLEDS, KAEA, and the University of Oklahoma. The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

Key words: Density forecasts, forecast combination, survey forecasts, shrinkage, model selection, regularization, partially egalitarian LASSO, model averaging, subset averaging

JEL codes: C2, C5, C8

Contact: fdiebold@sas.upenn.edu, minchul.shin@phil.frb.org, boyuanz@sas.upenn.edu

1 Introduction

Forecast combination for a series y involves transforming a set of forecasts of y, $f = (f_1, ..., f_K)'$, into a "combined", and hopefully superior, forecast $c(f)$. Most of the huge literature focuses on linear combinations of univariate point forecasts, in which case we can write the combined forecast as $c(f;\omega) = \omega' f$, for combining weight vector $\omega = (\omega_1, ..., \omega_K)'$.[1] We typically proceed under quadratic loss, choosing the weights to minimize the sum of squared combined forecast errors (SSE),

$$SSE(c(f;\omega), y) = \sum_{t=1}^{T} (y_t - \omega' f_t)^2,$$

where the sample of forecasts and realizations covers $t = 1, ..., T$. That is, we simply run the least-squares regression $y \to f_1, ..., f_K$, so that[2]

$$\hat{\omega} = \arg\min_{\omega} \Big(SSE(c(f;\omega), y) \Big).$$

This is the classic Bates and Granger (1969) and Granger and Ramanathan (1984) solution.

Recent point forecast combination literature such as Diebold and Shin (2019), however, focuses instead on weights that solve a *penalized* estimation problem,

$$\hat{\omega} = \arg\min_{\omega} \Big(\ Objective(c(f;\omega), y) \ + \ \lambda \cdot Penalty(\omega) \ \Big), \tag{1}$$

where the Lagrange multiplier λ governs the strength of the penalty. Maintaining quadratic loss we have

$$\hat{\omega} = \arg\min_{\omega} \Big(\ SSE(c(f;\omega), y) \ + \ \lambda \cdot Penalty(\omega) \ \Big).$$

If λ=0 we obviously obtain the Bates-Granger-Ramanathan solution, but the recent literature focuses on λ>0. This produces regularization, which can be highly valuable in the finite samples often of practical relevance, particularly for economic survey forecasts where the sample size T is often very small relative to the number of forecasters K. The precise form of the penalty determines the precise form of regularization, but in general it involves selection and/or shrinkage in directions guided by the penalty. For example, the famous LASSO penalty of Tibshirani (1996), $Penalty(\omega) = \sum_{k=1}^{K} |\omega_k|$, induces both selection to 0

[1] Broad and insightful surveys include Timmermann (2006), Elliott and Timmermann (2016), and Aastveit et al. (2019).

[2] We assume unbiased forecasts, so there is no need for an intercept.

and shrinkage toward 0.

In this paper we extend the idea of regularized forecast combination to the density forecast case. Density forecasting is important because predictive densities are complete probabilistic statements, which are always desirable, sometimes invaluable, and increasingly available. Density forecasts provide much more information, for example, than interval forecasts, which in turn provide more information than point forecasts.[3]

We work with "linear opinion pools" (mixtures), as in the key contributions of Hall and Mitchell (2007), Geweke and Amisano (2011) and Amisano and Geweke (2017), but we consider a variety of estimation objectives, and most importantly, we introduce regularization constraints. Our regularized density forecast combinations are regularized mixtures, and important subtleties arise in constructing appropriate penalties for mixture regularization. In this paper we confront this situation and propose several solutions.

Our methods are related to earlier and current work in both the econometrics and statistics literatures. A basic insight underlying our work and much of the recent literature is that Bayesian model averaging (BMA) as traditionally implemented is unattractive for combining density forecasts from misspecified models, because it fails to acknowledge misspecification (Diebold, 1991). That is, it assumes implicitly or explicitly that one of the models is "true", in which case the posterior predictive density asymptotically puts all probability on that model, so that BMA actually *fails* to average. Instead, once we acknowledge that all models are misspecified, we want a method capable of delivering a defensible and *diversified* portfolio (weighted average) of models, even asymptotically.

In one strand of econometrics literature this led Hall and Mitchell (2007), Brodie et al. (2009), Geweke and Amisano (2011), and Amisano and Geweke (2017) *inter alia* to move away from BMA, working instead with linear opinion pools that optimize the log score. In a different strand of econometrics literature that also moved away from BMA, it led Billio et al. (2013) to treat density forecast combination as a nonlinear filtering problem, potentially with time-varying mixture weights. Parallel developments in the statistics literature now acknowledge misspecification, distinguishing between "M-closed" vs. "M-complete" situations, and achieve diversified density forecast mixtures by "stacking" predictive densities (Yao et al., 2018), or via "dynamic Bayesian predictive synthesis" (McAlinn and West, 2019).[4]

[3]The evaluation of interval forecasts, moreover, is fundamentally problematic, as detailed in recent work by Askanazi et al. (2018) and Brehmer and Gneiting (2021).

[4]"M-closed" refers to a situation where the true model is among those being combined (but of course the econometrician does not know which it is) and "M-complete" refers to a situation where a true model exists but is *not* among those being combined, thereby formalizing the situations described in Diebold (1991). For additional discussion and nuances, see Yao et al. (2018).

We pick up from there and proceed as follows. In section 2 we discuss objectives for mixture regularization, that is, various choices and issues associated with $Objective(c(f;\omega), y)$. Then in section 3 we treat choices and issues associated with $Penalty(\omega)$, starting with the key unit simplex penalty, which we maintain throughout, and then introducing hybrid penalties that blend the simplex penalty with others. In section 4 we present Monte Carlo evidence on the efficacy of our procedures. In section 5 we present empirical results for European Central Bank (ECB) survey density forecasts of Eurozone inflation and real interest rates. We conclude in section 6.

2 Objectives

Consider a discrete density (histogram) forecast for a scalar variable y, which takes values in $m = 1, ..., M$ bins, or categories.[5] Denote the forecast by $p = (p_1, ..., p_M)'$. We start with density forecast "scores" for a single forecaster in a single period in sections 2.1-2.3. We then extend the discussion to multiple forecasters and periods in section 2.4, and we provide additional discussion in section 2.5.

2.1 Log Score

The log score (Good, 1952; Winkler and Murphy, 1968) is

$$L(p, y) = -\log\left(\sum_{m=1}^{M} p_m \, 1(y \in b_m)\right), \tag{2}$$

where p_m is the probability assigned to bin b_m, and $1(y \in b_m) = 1$ if $y \in b_m$ and 0 otherwise.

Ranking density forecasts by L, where smaller is better, reflects a preference for "small surprises". In a frequentist interpretation, L is just the (negative of the) log predictive density evaluated at the realization; that is, it is the (negative of the) predictive log likelihood. In a Bayesian interpretation, L is, desirably, a strictly proper scoring rule.[6]

2.2 Brier Score

The Brier score (Brier, 1950) is:

[5] We focus largely on the discrete case, because it is the one of practical relevance for survey forecasts that we eventually analyze. Parallel developments of course exist for the continuous case.

[6] On scoring rules see Gneiting and Raftery (2007) and the references therein.

$$B(p,y) = \frac{1}{M}\sum_{m=1}^{M}\left(p_m - 1(y \in b_m)\right)^2.$$

The Brier score generalizes the idea of quadratic loss to density forecasts. Indeed B is effectively the same as the so-called "quadratic score",

$$Q(p,y) = -2\left(\sum_{m=1}^{M} p_m\, 1(y \in b_m)\right) + \left(\sum_{m=1}^{M} p_m^2\right), \tag{3}$$

as noted by Czado et al. (2009). Rankings by Q must match rankings by B, because one is a positive monotonic transformation of the other. Both B and Q are strictly proper scoring rules under weak conditions.

2.3 Ranked Score

The ranked score (Epstein, 1969) is,

$$R(p,y) = \sum_{m=1}^{M}\left(P_m - 1(y \le b_{m+})\right)^2,$$

where $P_m = \sum_{h=1}^{m} p(b_h)$ is the cdf of the density forecast p, defined on bins $b_m = [b_{m-}, b_{m+}]$, $m = 1, ..., M$. R effectively proceeds by comparing realizations to the cdf forecast rather than the density forecast. R is strictly proper under weak conditions.

2.4 Multiple Forecasters and Time Periods

Let us now modify the notation to identify the specific forecaster, k. Thus far there has been no need, as we have considered just one forecaster, but shortly we will want to consider a set of forecasters, $k = 1, ..., K$. This is just a notational change, inserting "k" subscripts in the relevant places. In addition let us write the scores for a set of periods, $t = 1, ..., T$, rather than for just one period. This just involves summing over time.

We have:

$$L_k(\mathbf{p_k}, \mathbf{y}) = \sum_{t=1}^{T}\left(-\log\left(\sum_{m=1}^{M} p_{mkt}\, 1(y_t \in b_m)\right)\right),\ k = 1, ..., K$$

$$B_k(\mathbf{p_k}, \mathbf{y}) = \sum_{t=1}^{T} \left(\frac{1}{M} \sum_{m=1}^{M} \left(p_{mkt} - 1(y_t \in b_m) \right)^2 \right), \; k = 1, ..., K$$

$$R_k(\mathbf{p_k}, \mathbf{y}) = \sum_{t=1}^{T} \left(\sum_{m=1}^{M} \left(P_{mkt} - 1(y_t \leq b_{m+}) \right)^2 \right), \; k = 1, ..., K,$$

where $\mathbf{p_k} = (p_{k1}, ..., p_{kT})$ is the sequence of density forecasts over time for forecaster k, and $\mathbf{y} = (y_1, ..., y_T)$ is the sequence of realizations over time.

2.5 Discussion

Thus far we have implicitly emphasized the differences among the L, B, and R scores, but there are also many similarities.

B, for example, might appear linked to Gaussian environments, because it is a mean-squared error analog, unlike L, which is based directly on the likelihood and therefore valid under great generality. But it is not; indeed its "Q version" (3),

$$Q = -2L + \left(\sum_{m=1}^{M} p_m^2 \right),$$

reveals the intimate relationship between B and L. Moreover, B remains a strictly proper scoring rule regardless of distributional environment.

Now consider R. First, it is interesting to note that R is a generalization of absolute-error loss to density forecasts, just as B is a generalization of squared-error loss to density forecasts. In particular, Gneiting and Raftery (2007) show that R is driven by $E_p|Y - y|$:

$$R(p, y) = E_p|Y - y| - \frac{1}{2} E_p|Y - Y'|,$$

where Y and Y' are independent copies of a random variable with distribution p.

Second, R's generalization of absolute-error loss (MAE) to density forecasts also makes it a generalization of the Diebold and Shin (2017) stochastic error distance (SED), because MAE and SED rankings must agree, as shown by Diebold and Shin (2017). Moreover, and interestingly, SED is based on cdf divergences, just as is R.

Finally, although R might appear linked to a particular (Laplace) distributional environment, because it is an absolute-error analog, it is not. R is a strictly proper scoring rule regardless of distributional environment.

3 Penalties

Our goal is to produce mixtures of density forecasts,

$$c(\omega) = \sum_{k=1}^{K} \omega_k p_k,$$

with regularized mixture weights $\omega = (\omega_1, ..., \omega_K)'$. We score mixtures in the same way as we scored individual density forecasts. The only difference is that we now score the mixture, $c(\omega)$, rather than an individual forecast, p_k.

Thus far we have focused on appropriate objectives for regularized mixture weight estimation, $objective(c(\omega), y)$, and we emphasized use of strictly proper density forecast scoring rules. Now we consider appropriate constraints for regularized mixture weight estimation, $penalty(\omega)$. As we shall see, imposition of the unit simplex constraint (i.e., imposing that mixture weights be non-negative and sum to one: $\omega_i \geq 0 \ \forall i$ and $\sum_{i=1}^{K} \omega_i = 1$) provides essential regularization. In addition, however, simultaneous imposition of other regularization constraints may also be helpful.

3.1 Simplex

The unit simplex constraint has two parts: non-negativity and sum-to-one. For point forecasts we can relax both parts and potentially achieve better combined point-forecasting performance, as recognized by Granger and Ramanathan (1984) and done routinely ever since. As first recognized in the pioneering work of Brodie et al. (2009), it turns out that density forecasts are different: *When combining density forecasts it is crucial to impose (both parts of) the simplex constraint.*

First consider non-negativity. For point forecasts, allowing negative combining weights can improve performance, in a fashion analogous to allowing short positions in a financial asset portfolio. For density forecasts, in contrast, negative weights are unambiguously problematic, producing pathologies even if sum-to-one holds, because negative mixture weights can drive parts of the mixture density negative.

Now consider sum-to-one. Immediately, sum-to-one is required for the mixture combination to be a valid probability density.[7] Moreover, and separately, the solution to the mixture weight estimation problem can be pathological without imposition of sum-to-one.

[7] See also Yao et al. (2018), who briefly discuss issues related to the imposition of convex mixture weights.

To see this, consider a simple example with two continuous density forecasts and a log score objective. We have

$$\hat{\omega} = \arg\min_{\omega_1,\omega_2} \left(-\sum_{t=1}^{T} \log(\omega_1 f_{1,t}(y_t) + \omega_2 f_{2,t}(y_t)) \right),$$

where $f_{k,t}(y_t)$ is forecaster i's density forecast evaluated at the realization, y_t. Without the sum-to-one constraint, the optimal solution is not well defined: either $\omega_1 \to \infty$ or $\omega_2 \to \infty$ leads to the smallest possible objective function value, because $f_{1,t}$ and $f_{2,t}$ are non-negative for any y_t.

For all of the above reasons, we henceforth impose both the non-negativity and sum-to-one parts of the simplex constraint. Interestingly, moreover, their imposition is not only necessary to eliminate pathologies, but also desirable to provide regularization. In particular, the simplex constraint clearly imposes a particular L^1 "parameter budget"; it is effectively a special case of LASSO.

Assembling everything, the basic regularized estimator with log score objective (Geweke and Amisano, 2011; Amisano and Geweke, 2017) is[8]

$$\arg\min_{\omega} \left(-\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} \omega_k f_{k,t}(y_t) \right) \right) \tag{4}$$

$$\text{s.t. } \omega_k \in (0,1), \ \sum_{k=1}^{K} \omega_k = 1.$$

The methodological question remains, however, of how to provide additional, and more flexible, regularization, as does the substantive situation-specific empirical question of whether and where additional regularization is helpful. In the remainder of this paper we work toward answering both questions.

3.2 Simplex+Ridge

L^1 simplex regularization is a special case of L^1 LASSO regularization, corresponding to a specific choice of LASSO regularization parameter. Hence we cannot introduce additional L^1 regularization.

[8]Other objectives may of course be used, as discussed earlier in section 2. Note that for a histogram forecast we have $f_{k,t}(y_t) = \sum_{m=1}^{M} p_{mkt} 1(y_t \in b_m)$.

Additional regularization of some other type may nevertheless be useful for a variety of reasons. One reason is that the sparsity promoted by the simplex constraint may not be desirable (Giannone et al., 2021), so we may want to shrink all K mixture weights away from 0, thereby "undoing" the selection implicit in the LASSO-style L^1 penalty, allowing for non-zero mixture weights on all forecasts. We focus in particular on introducing shrinkage toward an equally-weighted mixture (i.e., shrinkage of all K weights toward $1/K$).

Consider, for example, introducing L^2 regularization. Immediately, incorporating an L^2 penalty in addition to the simplex constraint, we have:[9]

$$\hat{\omega} = \arg\min_{\omega} \left(\underbrace{-\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} \omega_k f_{k,t}(y_t) \right)}_{\text{log score}} + \underbrace{\lambda \left(\sum_{k=1}^{K} \left(\omega_k - \frac{1}{K} \right)^2 \right)}_{L^2 \text{ penalty}} \right) \tag{5}$$

$$\text{s.t. } \omega_k \in [0,1], \ \sum_{k=1}^{K} \omega_k = 1.$$

This parallels the egalitarian ridge estimator of Diebold and Shin (2019), with an additional simplex constraint imposed. Note that, due to the simplex constraint, the solution may discard some forecasters (setting some weights approximately if not exactly to zero), but that situation becomes progressively less likely as λ grows, pulling the weights toward equality.

We can re-write (5) as

$$\hat{\omega} = \arg\min_{\omega} \left(\underbrace{-\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} \omega_k f_{k,t}(y_t) \right)}_{\text{log score}} + \underbrace{\lambda_1 \left(\sum_{k=1}^{K} |\omega_k| - 1 \right)}_{L^1 \text{ simplex/LASSO penalty}} + \underbrace{\lambda_2 \left(\sum_{k=1}^{K} \left(\omega_k - \frac{1}{K} \right)^2 \right)}_{L^2 \text{ ridge penalty}} \right), \tag{6}$$

$$\text{s.t. } \omega_k \in [0,1],$$

which emphasizes that simplex+ridge regularization involves a combination of L^1 and L^2 penalties.[10] Note, however, that we are not free to choose λ_1, because the sum-to-one constraint must bind; equations (5) and (6) instead coincide for "large enough" λ_1.

Equation (6) in turn reveals that simplex+ridge regularization is closely related to the

[9] For transparency we make most of our arguments using a log score objective.

[10] Equation (6) also reveals that simplex+ridge is closely related to an additive-penalty version of partial egalitarian LASSO (Diebold and Shin, 2019), but with the egalitarian penalty done in L^2 (ridge) form rather than L^1 (LASSO) form.

elastic net of Zou and Hastie (2005). The elastic net penalty is

$$Penalty(\omega) \quad = \quad \underbrace{\alpha \sum_{k=1}^{K} |\omega_k|}_{L^1 \text{ LASSO penalty}} \quad + \quad \underbrace{(1-\alpha) \sum_{k=1}^{K} \omega_k^2}_{L^2 \text{ ridge penalty}},$$

where $\alpha \in [0,1]$ is a parameter, so that elastic net also involves combinations of L^1 and L^2 (that is, LASSO/simplex and ridge) penalties. Elastic net is well known to work well for regularization problems with many correlated predictors, exactly the situation of relevance for the large sets of economic forecasts on which we focus.

3.3 Simplex+Divergence

Here we move from simplex+ridge to simplex plus a general penalty based on the divergence between two discrete probability measures. As we will see, the divergence penalty includes simplex+ridge as a special case, but it also introduces a rich variety of new possibilities. Write the estimator as

$$\hat{\omega} = \arg\min_{\omega} \left(\underbrace{-\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} \omega_k f_{k,t}(y_t) \right)}_{\text{log score}} + \underbrace{\lambda \, D\left(\omega, \omega^*\right)}_{\text{penalty}} \right) \tag{7}$$

$$\text{s.t. } \omega_k \in [0,1], \ \sum_{k=1}^{K} \omega_k = 1,$$

where $D(\omega, \omega^*)$ is a measure of divergence between w and w^*. The key insight is that once the simplex restriction is imposed, ω can be interpreted as a discrete probability measure on $\{1, 2, ..., K\}$. If we let ω^* be the uniform probability mass function with weight $1/K$ on each outcome, then the penalized optimization (7) shrinks the solution toward equal weights.

Maintaining uniform ω^* throughout, but using different divergence measures $D(\omega, \omega^*)$, we obtain new regularized estimators. For example:

1. The L^2 norm,

$$D(\omega, \omega^*) = \sum_{k=1}^{K} \left(\omega_k - \frac{1}{K} \right)^2,$$

produces the simplex plus egalitarian ridge penalty given in (5) and (6).

2. The L^1 norm (total variation),

$$D(\omega, \omega^*) = \sum_{k=1}^{K} \left| \omega_k - \frac{1}{K} \right|,$$

produces a simplex plus egalitarian LASSO penalty (Diebold and Shin, 2019).

3. Kullback-Leibler divergence (entropy) from ω to ω^*,

$$D(\omega, \omega^*) = -\log K - \sum_{k=1}^{K} \log \omega_k,$$

produces a "simplex+entropy" penalty, $-\sum_{k=1}^{K} \log \omega_k$. In Appendix A we formally show that the simplex+entropy regularized estimator,

$$\hat{\omega} = \arg\min_{\omega} \left(\underbrace{-\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} \omega_k f_{k,t}(y_t) \right)}_{\text{log score}} + \underbrace{\lambda \left(-\sum_{k=1}^{K} \log(\omega_k) \right)}_{\text{entropy penalty}} \right) \tag{8}$$

$$\text{s.t. } \omega_k \in (0,1), \ \sum_{k=1}^{K} \omega_k = 1,$$

arises as the posterior mode in a Bayesian analysis with a log score (pseudo-) log likelihood and a Dirichlet prior. It puts positive probability only on the unit simplex and also shrinks weights toward equality for a certain hyperparameter configuration.

4. Rényi divergence of order α from ω to ω^*,

$$D_\alpha(\omega^* || \omega) = \frac{1}{\alpha - 1} \log \left(\sum_{k=1}^{K} \frac{1/K^\alpha}{\omega_k^{\alpha-1}} \right),$$

encompasses various statistical divergences including Kullback-Leibler divergence (α=1) and Hellinger distance (α=2), and can be used to produce still more interesting regularized estimators.[11]

All of the above divergence functions shrink the density mixture weights toward equality,

[11] Rényi divergence, moreover, is equivalent to Cressie-Read discrepancy up to an affine transformation.

thereby promoting inclusion of more forecasters in the regularized mixture. Importantly, the optimization that defines the regularized estimator (7) is convex so long as $D(\omega, \omega^*)$ is a convex function of ω, because the log score and simplex constraints are convex functions of ω. This makes numerical computation of the estimator straightforward.

3.4 Partially-Egalitarian Ridge and Subset Averaging

One might want a density forecast version of the "partially egalitarian" penalizations developed for point forecasts case by Diebold and Shin (2019). Consider, for example, the simplex-constrained partially egalitarian ridge problem:

$$\hat{\omega} = \min_{w} \left(-\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} w_k f_{k,t}(y_t) \right) + \lambda \sum_{k=1}^{K} \left(w_k - \frac{1}{\delta(w)} \right)^2 \right) \tag{9}$$

$$\text{s.t. } w_k \in [0,1], \ \sum_{k=1}^{K} w_k = 1,$$

where $\delta(\omega)$ is the number of non-zero elements in ω. Recall that a LASSO-type L^1 penalty is also implicitly embedded in the partially egalitarian ridge estimator (9) via the simplex constraint, which promotes selection of some coefficients to 0. Hence partially-egalitarian ridge discards some forecasters and then shrinks the survivors' weights toward equality ($1/\delta(\omega)$), in contrast to the simplex+ridge estimator (5), which discards some forecasters but then shrinks *all* weights toward equality ($1/K$).

Computation of the partially-egalitarian ridge solution (9) is possible in principle, as follows. First we note that there are C_κ^K possible density forecasters mixtures, where $\kappa \in [1, 2, 3, ..., K]$) is the number of forecasters included in the mixture. Then, for the jth such mixture ($j = 1, 2, ..., C_\kappa^K$), we solve

$$L^*(\kappa, j) = \min_{w^j} \left(-\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} w_k^j f_{k,t}(y_t) \right) + \lambda \sum_{k=1}^{K} \left(w_k^j - \frac{1}{\delta(w)} \right)^2 \right) \tag{10}$$

$$\text{s.t.} w_k^j \in [0,1], \ \sum_{k=1}^{K} w_k^j = 1,$$

where w_k^j is zero if the kth forecaster is not selected in the jth mixture. In this case, some

of the mixture weights are forced to zero, so the penalty term is reduced to

$$\lambda \sum_{k=1}^{K} \left(w_k^j - \frac{1}{\delta(w)} \right)^2 = \lambda \sum_{k \in \mathcal{N}} \left(w_k^j - \frac{1}{\kappa} \right)^2,$$

where $\mathcal{N} = \{k : w_k^j \neq 0\}$. This is just a partially egalitarian ridge mixture for a particular set of forecasters. The solution to the full partially egalitarian ridge problem (9) is then $\arg\min_{\kappa,j} L^*(\kappa, j)$.

Unfortunately, however, if partially-egalitarian ridge mixtures are possible in principle, they are nevertheless infeasible in practice. The computational cost is huge, because the penalized optimization (10) must be solved numerically $n_K = \sum_{\kappa=1}^{K} C_\kappa^K$ times. For example, when K=20, n_K=1048575.

But there is one very important exception. As $\lambda \to \infty$ in equation (9), the partially egalitarian estimator converges to a direct subset averaging procedure in the spirit of Elliott (2011), which is simple to compute and automatically imposes the simplex constraint. The subset averaging idea is trivial: At each time, rolling forward, we simply find and use the historically best-performing average.

The basic subset average is "best N-average". We exogenously select N, the number of included forecasters, and then we determine the historically best-performing N-forecast average and use it. We might, for example, select N=3, in which case at any time t we use data through time t, determine the historically best-performing 3-average, and use it to predict t+1. Then at time t+1 we update and use data through time t+1, determine the historically best-performing 3-average, use it to predict t+2, and so on, proceeding through the sample.

A natural extension of "best N-average" is "best $\leq N_{max}$-average", which eliminates the choice of N, instead requiring only choice of a maximum number of included forecasters, N_{max}. At each time t we determine the historically best-performing N-average such that $N \leq N_{max}$, and we use it to predict t+1, proceeding through the sample.

Although subset average computation is much less demanding than full partial egalitarian ridge computation, it can nevertheless be substantial, depending on K and N (or N_{max}). Finding the best $\leq N_{max}$-Average from among K forecasters, for example, requires computing ${}_KC_{Nmax} + {}_KC_{Nmax-1} + ... + {}_KC_1$ simple averages and then sorting them to determine the minimum, each period. Fortunately, however, the relevant K and N_{max} are quite small in typical economic combinations. In our subsequent empirical work, for example, $N_{max} \leq 4$ appears adequate, and we have K=19. Best $\leq$4-average combination requires evaluating

and sorting just ${}_{19}C_4 +_{19} C_3 +_{19} C_2 +_{19} C_1 = 5035$ averages per period.

3.5 Discussion

Having now considered both regularization objectives and constraints (penalties), and the resulting estimators that balance them, some additional discussion is warranted.

3.5.1 On the Novelty of Density Forecast Mixtures

It is important to note that our regularized mixtures of density forecasts are not just straightforward adaptations of existing methods of combining point forecasts. They differ in important and interesting ways. First, the objective function changes. Things like "forecast errors" and the "sum of squared errors" are ill-defined in the density case. Appropriate density forecast scoring rules must be used. We have emphasized several, including the log score, the Brier score, and the ranked score.

Second, the penalty function changes. When forming mixtures of density forecasts, the unit simplex constraint *must* be imposed, and it has the side benefit of proving some regularization. Mixtures of density forecasts nevertheless admit new regularization penalties that are intimately connected to the maintained simplex constraint, by viewing the mixture weights as a discrete probability distribution. We introduced several such penalties, emphasizing Kullback-Leibler distance (entropy).

Finally – and we have not yet noted this – it is generally unnecessary to center regularization penalties around equal weights once the simplex constraint is imposed. Shrinkage toward equal weights will be induced either way. Consider, for example, the simplex+ridge penalty in equation (5), and consider centering around equal weights, as written, vs centering around 0. There is no difference, because

$$\begin{aligned} \sum_{k=1}^{K} \left(\omega_k - \frac{1}{K} \right)^2 &= \sum_{k=1}^{K} \omega_k^2 - \frac{2}{K} \sum_{k=1}^{K} \omega_k + \frac{1}{K} \\ &= \sum_{k=1}^{K} \omega_k^2 + \frac{1-2K}{K}, \end{aligned} \tag{11}$$

where the last equality is due to the sum-to-one restriction embedded in the simplex constraint.[12] The intuition is simply that shrinkage toward 0 is *impossible* when maintaining

[12] In fact this equivalence holds as long as all weights are centered on the same value (it does not have to be $1/K$) and the weights are constrained to sum to a bounded real value (it does not have to be 1).

the sum-to-one restriction, and equal weights are as close to 0 as one can get.

3.5.2 On Our Use of Linear Mixtures

Our focus on *linear* mixtures is intentional. There are several reasons. First, linear combinations are simple to compute, both in absolute terms and relative to nonlinear combinations.

Second, linear combination is typically the default option in practice. For example, when central bank density forecast surveys summarize their results, they construct a representative predictive density based on linear combination (as with the ECB-SPF surveys that we study below).

Finally, linear combinations turn out to have provably good properties. For example, as noted by Geweke and Amisano (2011), if the process of combination is to commute with any possible marginalization of the distributions involved, then the combination must be linear.

It is worth noting, however, that although we emphasize linearity for the above reasons, our core approach – regularized mixtures of predictive densities – is potentially equally relevant in nonlinear environments. Nonlinear classes of mixture aggregation rules include logarithmic weighted averaging of PDFs (Kascha and Ravazzolo, 2010; Wallis, 2011); weighted averaging of CDFs, or "quantile aggregation" (Busetti, 2017); and the EMOS approach popular in weather forecasting (Gneiting et al., 2005).[13] Exploration of nonlinear approaches is, however, beyond the scope of the present paper.

3.5.3 On the Relationship of our Mixture Combinations of Predictive Densities to Mixture-of-Experts Models

The mixtures-of-experts model refers to a large class of mixture models in statistics, computer science, machine learning and related disciplines (Yuksel et al., 2012; Gormley and Frühwirth-Schnatter, 2019), and it offers a flexible way to approximate an arbitrary function by mixing several density functions. More specifically, in the mixtures-of-experts framework, the mixture models aim to learn about the conditional distribution of the output variable conditional on input variables (i.e., covariates). In such mixture models, location, scale, and mixing probabilities are allowed to be different across individual densities (i.e., experts). This type of model is quite flexible and offers a theoretical guarantee to consistently estimate the unknown data generating process (i.e., a true conditional density of the output variable given covariates) under some conditions (Jiang and Tanner, 1999; Norets, 2010).

[13] See also Ranjan and Gneiting (2010), Gneiting and Ranjan (2013), Billio et al. (2013), Kapetanios et al. (2015), McAlinn and West (2019), and Takanashi and McAlinn (2020).

Our mixture combinations of predictive distributions approach and the mixtures-of-experts approach are related insofar as both aim to approximate the data generating process by combining several density functions. However, they are different in an important way. Our approach takes each mixture component as given while the mixtures-of-experts model lets the user parameterize and estimate each component in the mixtures. Moreover, in our framework oftentimes we do not even know the exact conditioning set (or, conditioning variables) that each forecaster used to derive their probability/density forecast, while the user of the mixtures-of-experts model selects which covariates to include in the mixture component. Therefore, our approach and problem is about how to aggregate information contained in each density forecast to produce a better predictive density, whereas the mixtures-of-experts approach and problem is about how to represent and estimate the conditional probability density function using the mixtures of densities.

4 Monte Carlo

We now explore the potential of our regularized mixture estimators via a small Monte Carlo experiment.

4.1 Data-Generating Process and Forecasts

The data-generating process (DGP), which we assume to be known by the forecasters, is:

$$\begin{aligned} y_t &= x_t + \sigma_y e_t, \quad e_t \sim \text{iid}\, N(0,1) \\ x_t &= \phi_x x_{t-1} + \sigma_x v_t, \quad v_t \sim \text{iid}\, N(0,1), \end{aligned} \tag{12}$$

where e and v are orthogonal at all leads and lags. y is the variable to be forecast, and x_t can be interpreted as the long-run component of y_t. Individual forecasters receive heterogeneous independent noisy signals about x_t. For forecaster k we have

$$z_{kt} = x_t + \sigma_{zk}\eta_{kt}, \quad \eta_{kt} \sim \text{iid}\, N(0,1), \tag{13}$$

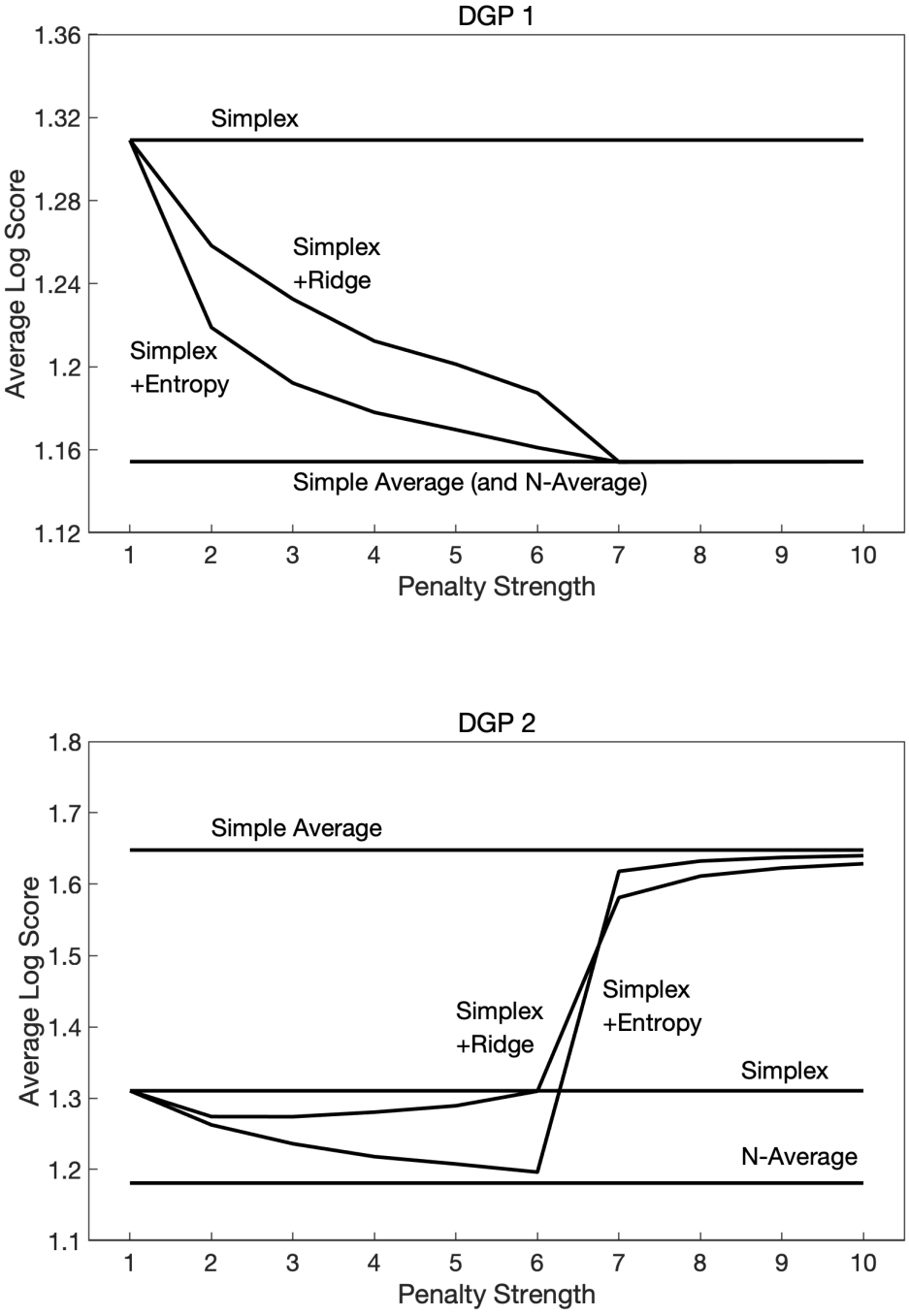
where η_k and $\eta_{k'}$ are orthogonal at all leads and lags for all forecasters k and k'. This signal is noisy yet leading in that z_{kt+1} is available to forecaster k at time t. Assume that forecasters have a strong belief that the 1-step-ahead predictive density is Gaussian with variance σ_y^2, but that they don't know its mean, and that forecaster k therefore uses z_{kt+1}, resulting in

Table 1: Average Log Scores

	DGP 1			DGP 2		
Regularization group	L	#	λ^*	L	#	λ^*
Simplex	1.31	5.27	NA	1.31	4.75	NA
Simplex+Ridge	1.15	20.00	2511.25	1.20	8.66	15.00
Simplex+Entropy	1.15	20.00	5.22	1.27	20.00	0.10
Subset Averages	L	#	λ^*	L	#	λ^*
Best N-Average:						
$N = 1$	2.65	1.00	NA	2.86	1.00	NA
$N = 2$	1.60	2.00	NA	1.61	2.00	NA
$N = 3$	1.38	3.00	NA	1.35	3.00	NA
$N = 4$	1.29	4.00	NA	1.27	4.00	NA
$N = 5$	1.23	5.00	NA	1.24	5.00	NA
$N = 6$	1.22	6.00	NA	1.20	6.00	NA
$N = 7$	1.21	7.00	NA	1.19	7.00	NA
$N = 8$	1.20	8.00	NA	1.18	8.00	NA
$N = 9$	1.18	9.00	NA	1.18	9.00	NA
$N = 10$	1.18	10.00	NA	1.19	10.00	NA
$N = 15$	1.16	15.00	NA	1.46	15.00	NA
$N = 20$	1.15	20.00	NA	1.65	20.00	NA
Best $\leq$ 2-Average	1.61	2.00	NA	1.62	2.00	NA
Best $\leq$ 3-Average	1.42	2.84	NA	1.40	2.81	NA
Best $\leq$ 5-Average	1.34	3.63	NA	1.35	3.44	NA
Best $\leq$ 10-Average	1.34	3.71	NA	1.34	3.49	NA
Best $\leq$ 15-Average	1.34	3.71	NA	1.34	3.49	NA
Best $\leq$ 20-Average	1.34	3.71	NA	1.34	3.49	NA
Comparisons	L	#	λ^*	L	#	λ^*
Best	0.24	1	NA	0.27	1	NA
75%	0.53	1	NA	0.99	1	NA
Median	1.66	1	NA	5.37	1	NA
25%	4.16	1	NA	33.73	1	NA
Worst	12.20	1	NA	193.02	1	NA
Simple Average	1.15	20	NA	1.65	20	NA

Notes: L is the average log score, # is the average number of forecasters selected, λ^* is the ex post optimal penalty parameter, and K is the total number of forecasters. We perform 10,000 Monte Carlo replications.

Figure 1: Monte Carlo Estimates of Expected Mixture Performance vs Penalty Strength



Note: We perform 10,000 Monte Carlo replications.

the predictive density

$$p_{kt}(y_{t+1}) = N(z_{kt+1}, \sigma_y^2). \tag{14}$$

Note that in this environment, forecasters' predictive densities differ only by their locations (means).

4.2 Results

We consider two parameterizations:

1. DGP 1: σ_{zk}=1 for all k
2. DGP 2: σ_{zk}=1 for $k = 1, 2, ..., \frac{K}{2}$ and σ_{zk}=5 for $k = \frac{K}{2}+1, ..., K$,

where each DGP has common parameters ϕ_x=0.9, σ_x=1, σ_y=0.5. The two DGPs differ only by the quality of the signals that forecasters receive. Under DGP 1 the simple average should be preferred, because all signals are of the same quality, while under DGP 2 the linear opinion rule should be preferred (at least asymptotically, so that estimation error vanishes), giving more weight to forecasters $k = 1, 2, ..., \frac{K}{2}$, who receive better signals.

To cohere with our subsequent empirical work, we explore K=T=20. We generate data, estimate mixture weights, generate 1-step-ahead mixture densities, and evaluate them using the log score objective. We repeat this 10,000 times and compute the average LPS for several methods:

1. Simple average
2. Simplex (equation (4))
3. Simplex+ridge (equation (5))
4. Simplex+entropy (equation (8))
5. Subset averaging (equation (9) with $\lambda \to \infty$).

For each of simplex+ridge and simplex+entropy, we explore 20 penalization strengths. For simplex+ridge, we choose 10 equispaced points in [1e-15,10] and 10 equispaced points in [15,10000]. For simplex+entropy we choose 10 equispaced points in [1e-15,0.2] and 10 equispaced points in [0.3,20].

Results appear in Table 1 and Figure 1. In Table 1 we present the optimized average log score for each method under DGPs 1 and 2, respectively. In Figure 1 we show how the

optimized score varies with regularization penalty strength under DGPs 1 and 2, respectively. Under DGP 1, simple averaging performs well, and unregularized simplex performs poorly, as expected. As the strength of shrinkage gets heavier, the performance of both simplex+entropy and simplex+ridge improves monotonically until they perform as well as the simple average (full shrinkage). In addition, the performance of simplex+entropy improves more quickly than that of simplex+ridge as shrinkage strength increases and dominates throughout. Finally, subset averaging performs admirably under DGP 1, and as expected the optimal "subset" includes all forecasters.

Under DGP 2, simplex is expected to perform well, and simple averaging is expected to perform poorly. Simplex does indeed outperform simple averaging. Moreover, both simplex+ridge and simplex+entropy behave as expected. For little shrinkage (toward the left), their performance is similar to that of simplex, and for heavy shrinkage (toward the right), their performance is similar to that of the simple average. In between, for moderate amounts of shrinkage, they outperform simplex. In that region, regularized simplex improves on unregularized simplex, because the large unregularized simplex estimation error makes it likely that some relevant forecasters are dropped from the pool, and regularization brings them back. Importantly, subset averaging continues to perform admirably under DGP 2, but now the optimal average involves only 10 or so forecasters, as expected.

4.3 Discussion

First, note that we endow our forecasters with incompletely-rational predictive densities. In particular, the mean of of each forecaster's predictive density is simply the observed signal z_{kt+1}, whereas replacing z_{kt+1} with $E[x_{t+1}|z_{kt+1}, z_{kt}, ..., z_{k1}]$ would result in better mean forecast. We do this intentionally, in an effort to achieve a more realistic Monte Carlo design, as many real-world forecasters appear incompletely rational. Nevertheless it is also of interest to explore fully-rational forecasts, which are an important and obvious benchmark. Hence we repeat the Monte Carlo with fully-rational forecasts. The results, which appear in Appendix B, are qualitatively identical.

Second, note that the performance documented in Table 1 and Figure 1 is almost surely not achievable in practice, because it is based on use of ex post optimal penalty parameters (λ's). Nevertheless the results are informative, because they document what can be achieved *in principle*, even if not necessarily in practice. Practical performance is an empirical matter, to which we now turn, in applications to Eurozone inflation and real interest rates. The empirical analysis will elucidate, among other things, practical implementation issues such

as the tuning of regularization parameters, with surprising results.

5 Eurozone Inflation and Real Interest Rate Forecasts

Here we use our methods to construct regularized mixtures of density forecasts for Eurozone inflation and real interest rates. Expected inflation is a key driver of the bond market via its direct impact on nominal interest rates. Expected inflation may also negatively impact real growth, and hence the stock market, insofar as it "puts sand in the Walrasian gears", as classically emphasized by Bresciani-Turroni (1937). High inflation, moreover, also tends to be volatile inflation (Friedman, 1977), which puts additional sand in the gears.[14] Expected inflation is also a key part of the ex ante real interest rate, which in turn is a key guide to intertemporal allocation and a key link between macroeconomic fundamentals and financial markets. From a variety of angles, then, inflation forecasts are central to financial markets, the macroeconomy, and the interface.

5.1 Data

Following the pathbreaking work of Conflitti et al. (2015), we study inflation density forecasts from the European Central Bank Survey of Professional Forecasters (ECB-SPF), which has been undertaken since 1999. Participants are surveyed quarterly, in January, April, July, and October.[15] Our forecast sample contains 83 quarterly surveys, starting in 1999Q1 and ending in 2019Q3.

The precise Euro-area inflation variable about which the ECB-SPF asks is the percentage change in the Harmonized Index of Consumer Prices (HICP), for the year following the forecast.[16] For example, when the survey was conducted in October 2017 (2017Q4), HICP inflation data were available up to September 2017, so the 2017Q4 survey asks for a forecast for the year from October 2017 through September 2018. Our realization sample, matched to our forecast sample, contains 83 quarterly observations, starting in December 1999 and ending in June 2020.

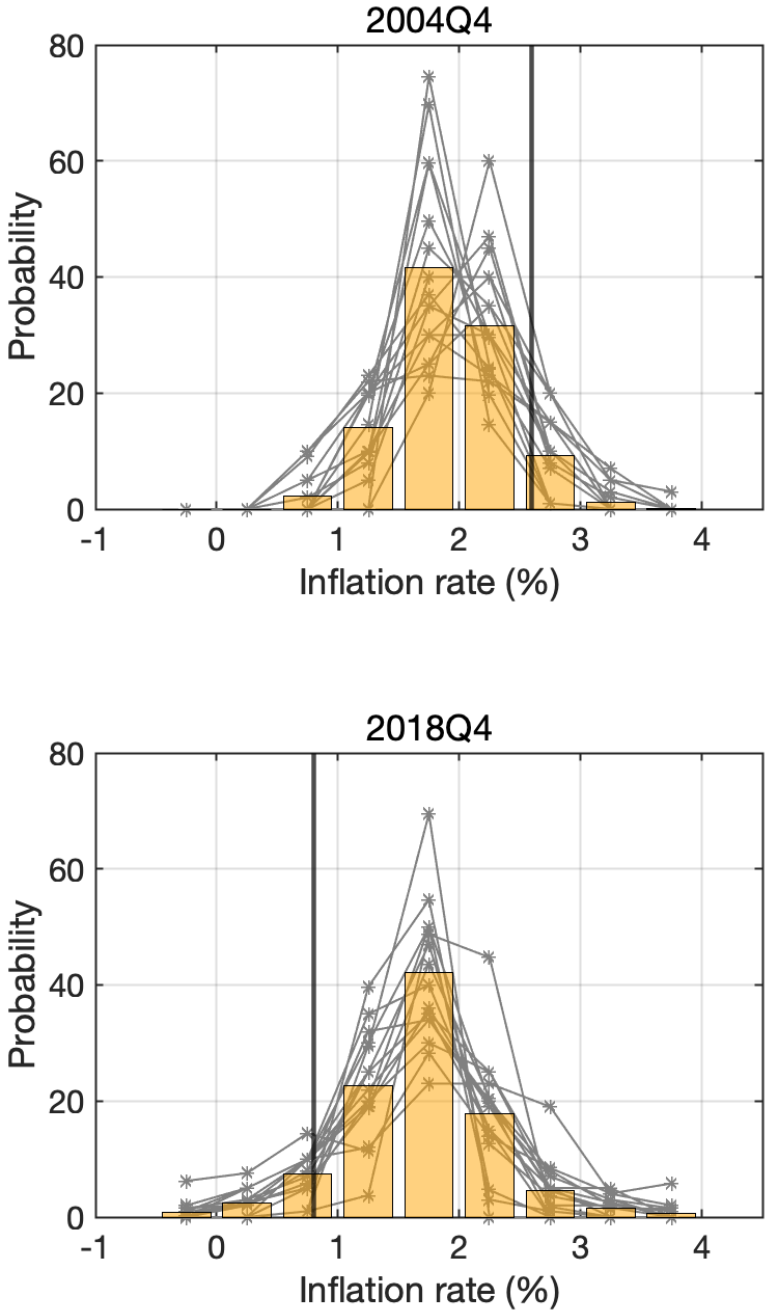
As an entrée into the density forecast data, in Figure 2 we show all individual survey

[14] See also Chen et al. (1986).

[15] See `https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/index.en.html`.

[16] Eurostat, Harmonized Index of Consumer Prices: All Items for Euro area (19 countries) [CP0000EZ19M086NEST], Retrieved from FRED, Federal Reserve Bank of St. Louis; `https://fred.stlouisfed.org/series/CP0000EZ19M086NEST`.

Figure 2: Eurozone Inflation: Individual Density Forecasts, Simple Average Mixture Forecasts, and Realizations



Notes: We show the individual survey inflation forecasts in gray (as frequency polygons), the simple average mixture forecast in orange (as a histogram), and realized inflation as a black vertical line.

inflation forecasts expressed as frequency polygons, and the simple average mixture forecast expressed as a histogram, for two illustrative surveys: 2004Q4 (early in the sample), 2018Q4 (late in the sample). For comparison we also show the 2004Q4 and 2018Q4 realizations as vertical lines in each panel.

At each date there is substantial variation of the individual forecasts around the average forecast. Moreover, substantial differences in the average forecasts are apparent at the two survey dates. The average forecast in 2004Q4, for example, puts 2.3% probability on the event that the inflation rate is less than 1%, whereas in 2018Q4 it puts 10.5% probability on the same event. More generally, the average forecast shifts systematically, from right-skewed in 2004Q4 to left-skewed in 2018Q4, and interestingly, the realization is indeed in the right tail of the mixture density for 2004Q4 and in the left tail for 2018Q4.

In Figure 3 we show the complete time series of simple average mixture forecasts and realizations. We show the average forecast densities as frequency polygons in the top panel, and as a heat map in the bottom panel, which also contains a time-series plot of the realizations. Large temporal movements in both location and scale of the average density forecasts are evident, and the realizations display similarly large variation.

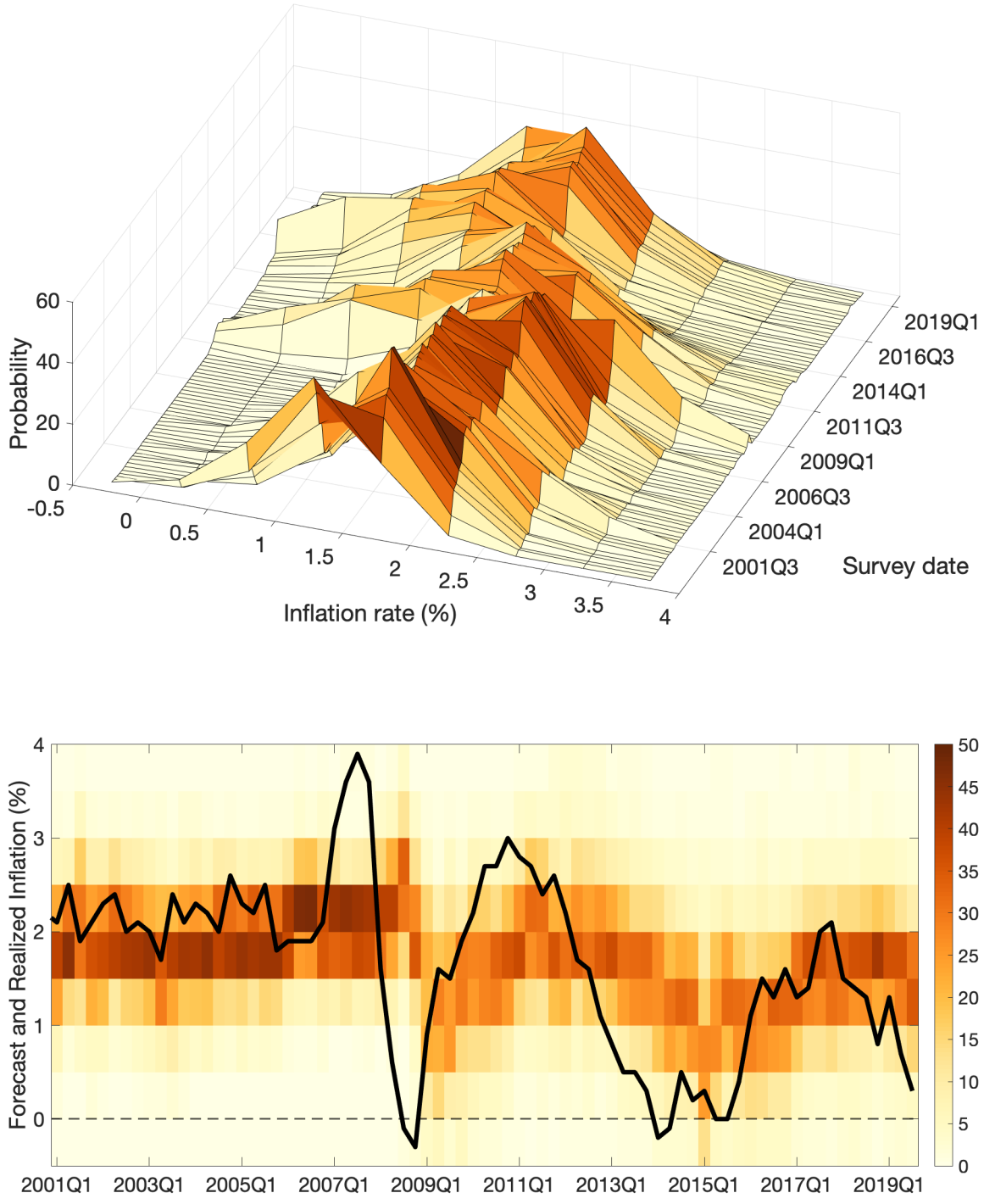
We will soon obtain mixture densities using the log score objective and several regularizations, including simplex, simplex+ridge, simplex+entropy, and subset averaging. Before proceeding to empirical results, however, we address several issues.

5.1.1 Survey Entry and Exit

First, forecasters can enter and exit the survey pool. There are 103 unique forecasters between 1999Q1 and 2019Q4, and no forecaster appears in the pool continuously. Following Genre et al. (2013), we proceed by first excluding forecasters who miss more than four consecutive surveys, which leaves 18 forecasters. Then we interpolate the remaining gaps based on historical performance.[17]

[17]More precisely, we fill in the gaps in the first survey (t=1, 1999Q1) with the average of non-missing forecasts from all other available forecasters. Then we calculate the ranked score for each forecaster and divide them into five mutually exclusive groups based on the score, and move to the second survey. At each of the following rounds ($t = 2, 3, ..., T$), we set the missing observations of a particular forecaster to the average of non-missing forecasts from her group, and then using the full set of forecasts we re-calculate ranked scores and update the group structure for use in the next round.

Figure 3: Eurozone Inflation: Simple Average Mixture Forecasts and Realizations



Notes: In the top panel we show simple average mixture densities over time, expressed as frequency polygons. In the bottom panel we show heat maps of the simple average mixture densities over time, with the realized Eurozone inflation rate superimposed. The dashed line indicates zero.

5.1.2 Time-Varying Bin Definitions

Second, outcome bin definitions vary over time. Although bin definitions have been stable for mid-range "standard" inflation values, extreme tail bins have become finer over time, as realizations have fallen in the tails. For example, for high inflation, there was originally a >3.5 bin, but it was eventually split into 3.5-4 and >4 bins.[18] We proceed by merging extreme tail bins sufficiently to produce 11 bin definitions, fixed for the entire sample: $(-\infty, -0.5]$, $(-0.5, 0]$, $(0, 0.5]$, ..., $(3.5, 4]$, $(4, \infty]$.

5.1.3 Zero-Probability Realizations

Finally, complications can arise with the log-score objective. Consider, for example, the survey forecast:

$$y \in \begin{cases} (-\infty, 1.5] & w.p. = 0 \\ (1.5, 2.0] & w.p. = .3 \\ (2.0, 2.5] & w.p. = .5 \\ (2.5, 3.0] & w.p. = .2 \\ (3.0, \infty] & w.p. = 0. \end{cases} \tag{15}$$

The zero probabilities assigned to the leftmost and rightmost bins obviously create a problem (infinite loss) for the log-score objective, due to its use of logs, if a realization occurs that was assigned zero probability.

Zero-probability realizations rarely, but occasionally, appear in our data. Sometimes they occur in edge bins (e.g., $(4, \infty]$), because forecasters sometimes fail to put positive probability on those bins. In addition to the edge-bin phenomenon, some forecasters' histograms are simply too sharp, and they sometimes put zero probability on an interior bin that eventually contains the realization.

One can address the log score "zero problem" by requiring the survey bin into which the realization falls to have been assigned at least some small probability, say 1%. We achieve this by assigning 1% probability to the bin containing the realization if it had originally been assigned 0, where the 1% is taken in equal shares from the bins originally assigned non-zero probability.[19]

[18] During our sample period the number of bins started at 9, peaked at 14 during the Great Recession, and eventually dropped to 12.

[19] One could of course switch to another objective, but the log score objective is simple and deservedly popular, which is why we have used it throughout this paper as a leading case for both our theory and Monte Carlo. We will continue to use it for our empirical work, where it is also deservedly popular, despite the zero

Table 2: Log Scores for Eurozone Inflation

Regularized Mixtures	L	#	ECB/SPF	L	#
Simplex	1.88	3.52	Best	2.02	1
Simplex+Ridge	1.86	5.00	90%	2.04	1
Simplex+Entropy	1.87	19.00	70%	2.13	1
Best 4-Average:	1.87	4	Median	2.17	1
Best $\leq$4-Average	1.90	2.24	Worst	2.56	1
Simple Average	1.98	18			

Notes: We show log scores for 1-year-ahead Eurozone inflation density forecasts, made quarterly, using a 20-quarter rolling estimation window. The burn-in sample is 1999Q1-2000Q4, and the forecast evaluation sample is 2001Q1-2019Q3 (75 quarters). There are 18 ECB-SPF density forecasters in the pool, plus a 19th forecaster whose predictive density is constant and uniform, for a total of 19 forecasters. L is the log score, and # is the average number of forecasters selected. Results for simplex+ridge and simplex+entropy are based on ex post optimal penalty parameters. See text for details.

5.2 Empirical Results for Inflation

There are 18 ECB-SPF density forecasters in the pool. We also include a fictitious 19th forecaster whose predictive density is constant and uniform, in rough parallel to including a constant in point forecast combining regressions, for a total of 19 forecasters. Doing so appears *a priori* desirable, in the tradition of Granger and Ramanathan (1984). Moreover, it constrains the mixture density to put positive probability on each histogram bin as long as the uniform forecaster gets a non-zero mixture weight, in which case the earlier-discussed log score "zero problem" vanishes.

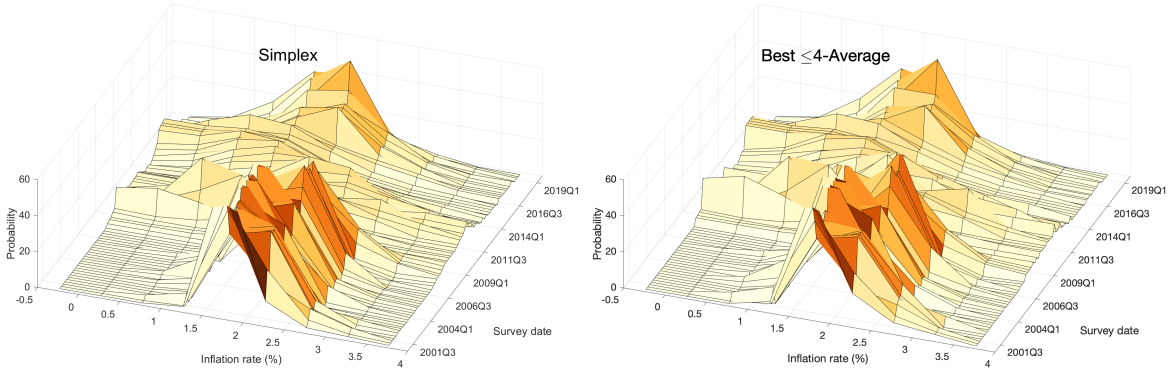
Results appear in Table 2. Strikingly, each regularized mixture outperforms each ECB/SPF individual forecaster (even the ex post *best* forecaster). To get a feel for the size of the improvement, note that the log score of the best $\leq$4-average, for example, is approximately 15% better than that of the median individual forecaster, and 7% better than that of the ex post best individual forecaster. Each regularized mixture also outperforms the simple average, which in turn outperforms the ECB/SPF forecasts.

Table 2 also reveals that the average number of forecasters selected after regularization is always small, regardless of the regularization method.[20] Simultaneously, both the log scores in Table 2 and the graphs in Figure 4 reveal that the simplex and best average regularized

problem.

[20]simplex+entropy selects all 19 forecasters, but simplex+entropy *must* select all 19 forecasters, because $\log(\omega_k) \to \infty$ as $\omega_k \to 0$. All regularizations capable of selecting only a few forecasters do in fact select only a few.

Figure 4: Simplex and Best ≤4-Average Mixture Forecasts Over Time, Eurozone Inflation



Notes: We show density forecast mixtures expressed as frequency polygons. The forecasts are quarterly, from 1999Q1 to 2019Q3.

mixtures are almost identical, suggesting that the simplex solution is effectively dropping all but a few forecasts and simply averaging the survivors, producing something very close to a best ≤4-average.

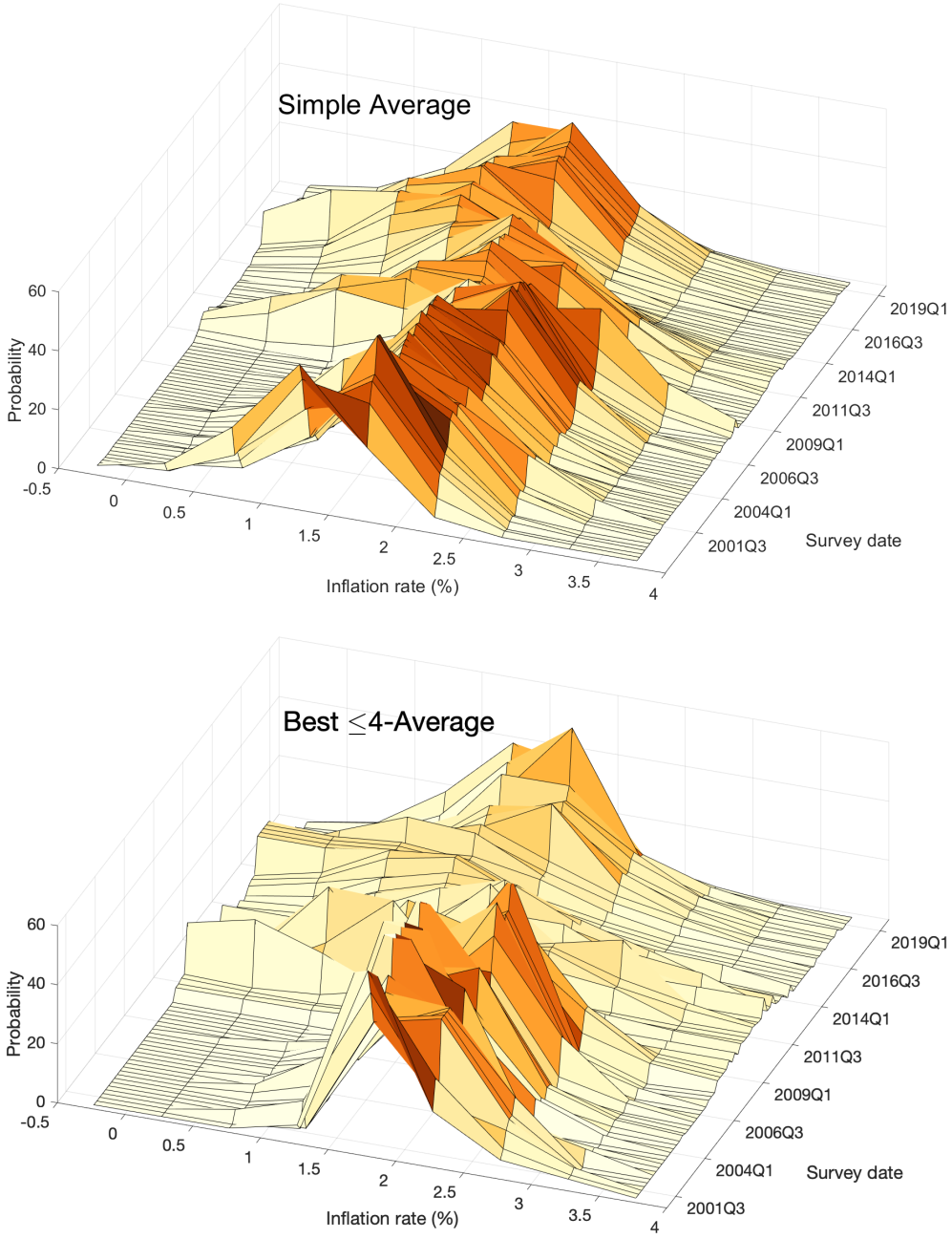
The good performance of both simplex and best average is particularly noteworthy insofar as neither requires tuning.[21] That is, quite remarkably, the simplex and best average regularizations perform as well as those requiring choice of tuning parameters (simplex+ridge and simplex+entropy), despite the fact that we evaluate the latter in Table 2 using ex post optimal tuning parameters, which is not feasible in real time.

If the effects of simplex and best ≤4-average regularization are almost identical, both are nevertheless *very* different from a simple average mixture. This is revealed clearly in Figure 5, in which we compare the time series of simple average mixture densities (top panel) and best-average mixture densities (bottom panel).

The bottom panel of Figure 5 also reveals that the effects of best-average regularization differ strikingly before and after the onset of the Great Recession. Before the onset of the Great Recession, best-average regularization moves probability mass upward toward higher inflation relative to simple averaging, particularly from the 1.0%-1.5% range to the 1.5%-2.5% range, mostly adjusting density forecast location and symmetry. After that, however, best-average regularization spreads probability mass from the center into both tails of the distribution, from the 1.0%-2.5% range outward to below 0.5% and above 3.0%, mostly adjusting density forecast dispersion and kurtosis.

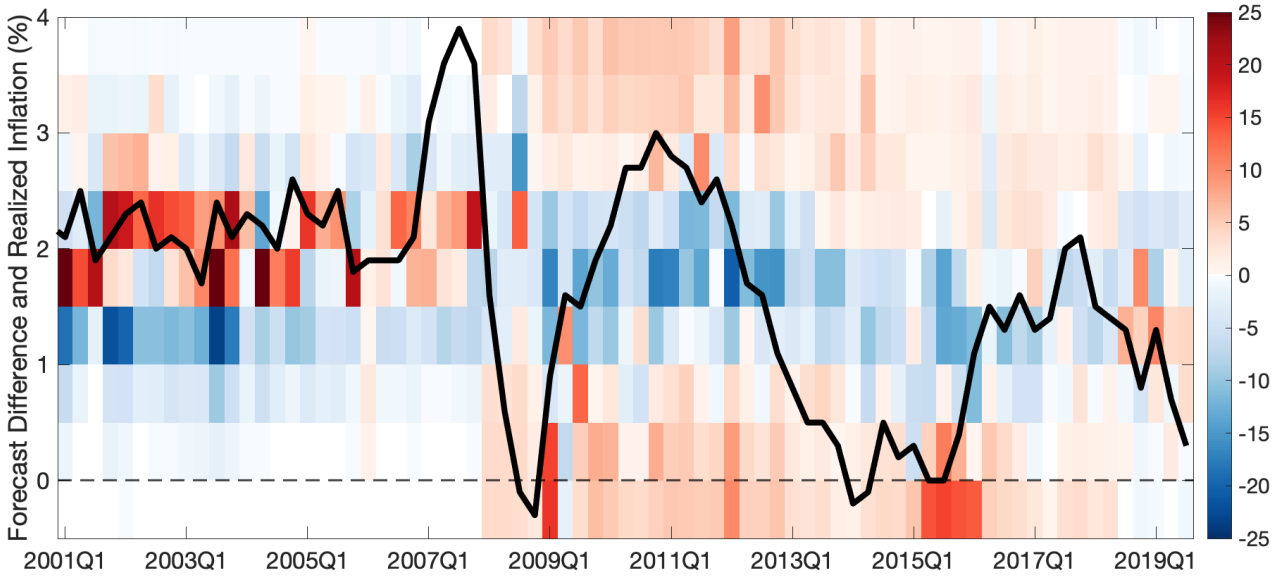
[21] Strictly speaking, best average procedures require some slight tuning – a choice of N – although we are comfortable with simply always adopting $N = 4$.

Figure 5: Simple Average and Best ≤4-Average Mixture Forecasts Over Time, Eurozone Inflation



Notes: We show density forecast mixtures expressed as frequency polygons. The forecasts are quarterly, from 1999Q1 to 2019Q3.

Figure 6: Difference Between Best ≤4-Average and Simple Average Mixture Forecasts, Eurozone Inflation, with Superimposed Inflation Realizations

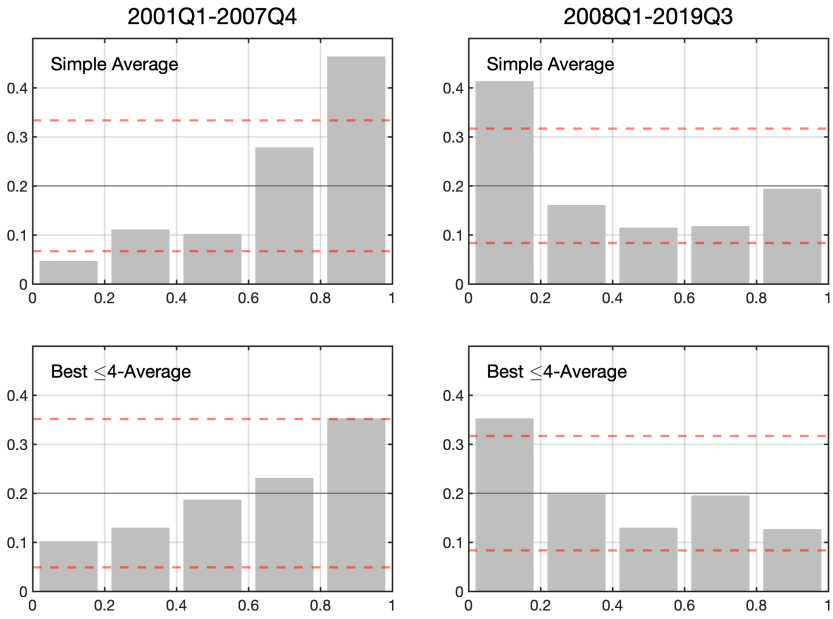


Notes: We show a heat map of the bin-by-bin differences between the best ≤4-average and simple average mixture densities (best ≤4-average minus simple average). Red bin shading indicates that best ≤4-average adds probability to the bin relative to the simple average, and blue bin shading indicates that best ≤4-average subtracts probability from the bin relative to the simple average. We also superimpose the realized Eurozone inflation rate. The dashed line indicates zero.

The effects of best-average regularization, and their structural shift at the onset of the Great Recession, are revealed even more clearly in Figure 6, where we show a heat map of the bin-by-bin differences between the best ≤4-average and simple average mixture densities (best ≤4-average minus simple average). Red bin shading indicates that best ≤4-average adds probability to the bin relative to the simple average, and blue bin shading indicates that best ≤4-average subtracts probability from the bin relative to the simple average. We also superimpose the realized Eurozone inflation rate.

Before the Great Recession, regularization clearly subtracts probability from low interest rate bins (blue shading, roughly from 0 to 1.5%) and adds it to higher interest rate bins (red shading, roughly from 1.5 to 2.5%), a good thing to do during that period as the realized inflation rate held close to 2%. After the Great Recession, in sharp contrast, regularization clearly subtracts probability from mid-range low interest rate bins (blue shading, roughly from 0.5 to 2.5%) and adds it to both higher and lower tail bins (red shading, roughly down to bins in 0.5 to -0.5%, and up to bins in 2.5-4%), reflecting increased awareness of tail risk. This was also a good adjustment, as post-2007 inflation volatility clearly increased, with

Figure 7: *PIT* Histograms, Simple Average and Best ≤4-Average Mixture Forecasts, Eurozone Inflation



Notes: We show histograms of discrete probability integral transforms (PITs) for simple average and best ≤4-average mixture forecasts. In red we show the pointwise binomial confidence bands that hold when $PIT \sim iid\, U(0,1)$. See text for details.

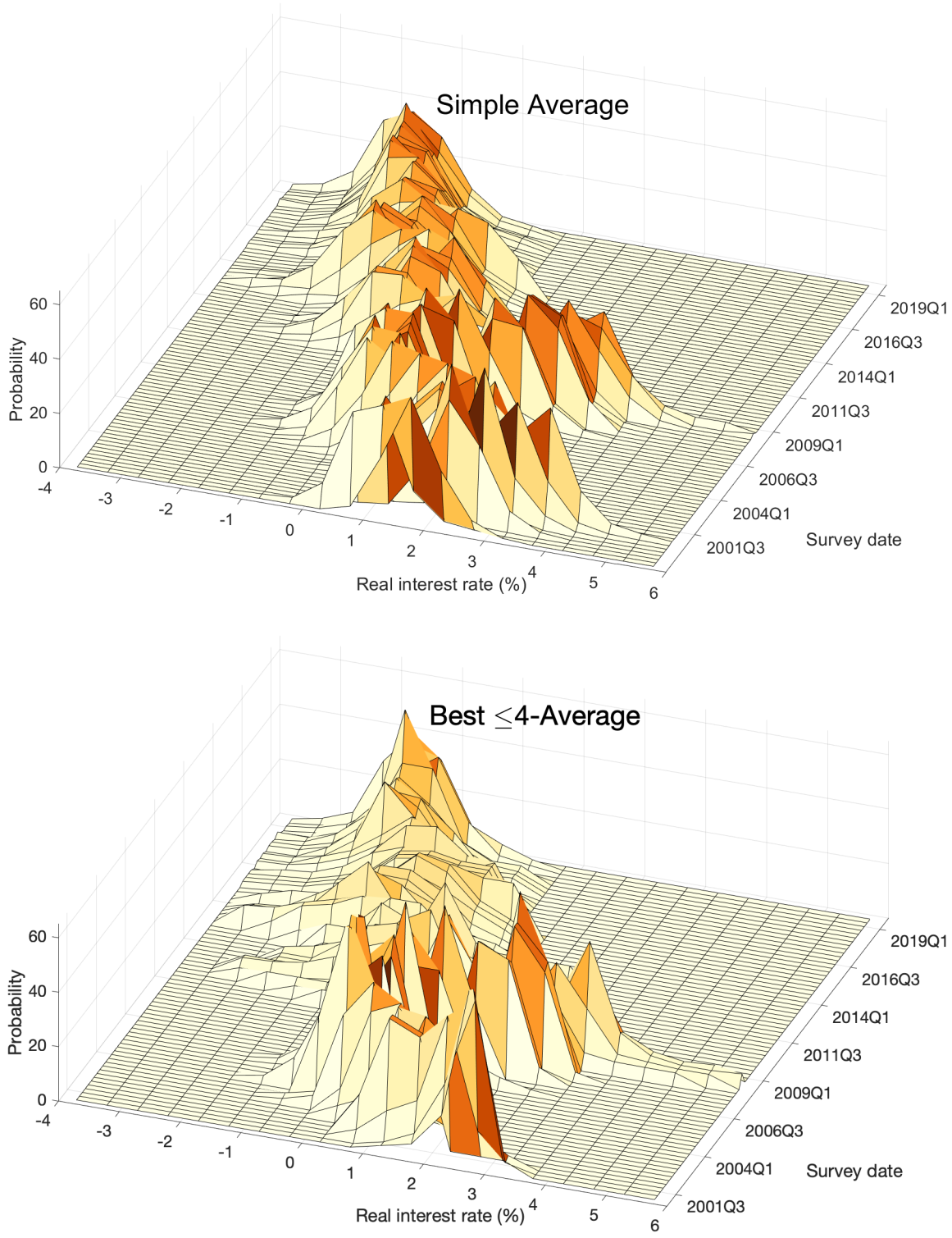
realized inflation taking wide swings.

It is also informative to examine and compare probability integral transforms (PITs) for various mixtures. Diebold et al. (1998) consider the continuous case, in which the PIT is defined as $PIT_t = \int_{-\infty}^{y_t} p_t(u)du$, and show that correct conditional calibration of density forecasts implies that $PIT \sim iid\, U(0,1)$. Czado et al. (2009) extend the evaluation framework to the discrete case and show that the result still holds for an appropriate discrete PIT definition. To assess uniformity, and any patterns in deviations from uniformity, in Figure 7 we show histograms of the Czado et al. (2009) discrete PIT for the simple average and best ≤4-average mixtures.

The PIT histograms reveal problems with the simple average mixture, which match our discussion of the two regimes in Figures 5 and 6. In particular, the simple average PIT histograms show noticeable deviations from uniformity in both subsamples, and the shapes of the deviations are very different.

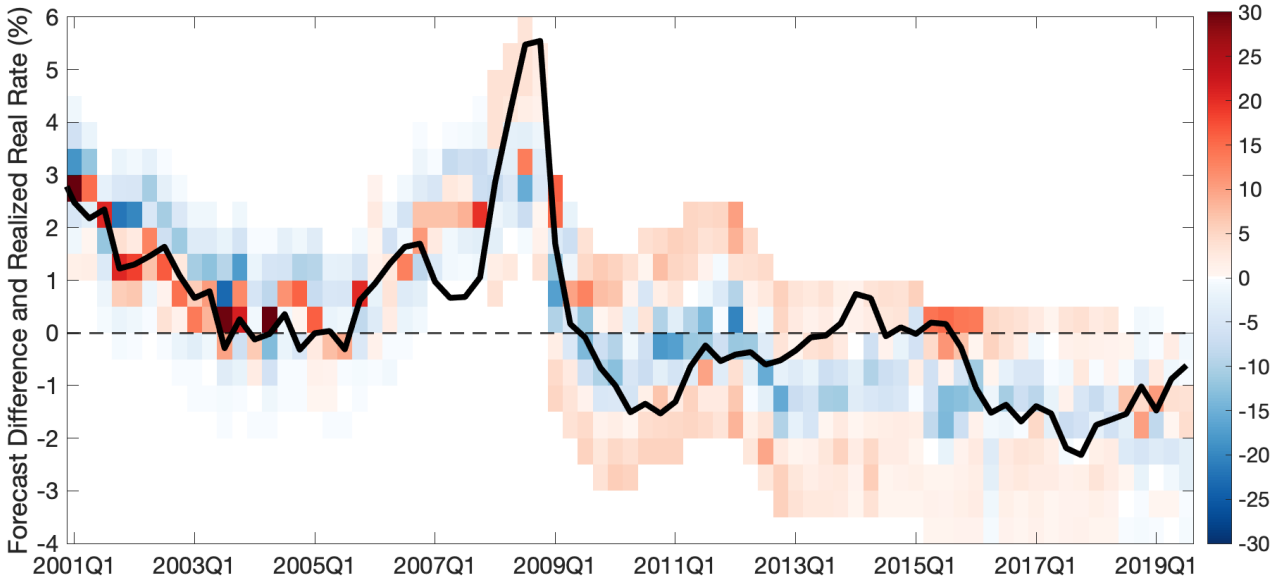
In the first subsample, the simple average PIT histogram is highly skewed as shown in the upper-left panel of Figure 7, with far too little probability mass near 0 and far too much

Figure 8: Simple Average and Best ≤4-Average Mixture Forecasts Over Time, Eurozone Real Interest Rate



Notes: We show density forecast mixtures expressed as frequency polygons. The forecasts are quarterly, from 1999Q1 to 2019Q3.

Figure 9: Difference Between Best ≤4-Average and Simple Average Mixture Forecasts, Eurozone Real Interest Rate, with Superimposed Real Interest Rate Realizations



Notes: We show a heat map of the bin-by-bin differences between the best ≤4-average and simple average mixture densities (best ≤4-average minus simple average). Red bin shading indicates that best ≤4-average adds probability to the bin relative to the simple average, and blue bin shading indicates that best ≤4-average subtracts probability from the bin relative to the simple average. We also superimpose the realized Eurozone real interest rate. The dashed line indicates zero.

near 1, again indicating too many large inflation realizations relative to the simple average density forecasts. Regularization, however, shifts the densities upward as discussed earlier, producing an improved (if still imperfect) best-average PIT as seen in the bottom left panel of Figure 7.

In the second subsample the simple average PIT histogram is more U-shaped, as shown in the upper-right panel of Figure 7. In this regime the regularization spreads out the densities as discussed earlier, better accommodating the tail realizations and producing an improved best ≤4-average PIT as seen in the bottom right panel of Figure 7.

5.3 Empirical Results for Real Interest Rates

Finally, in parallel to our earlier examination of ECB/SPF inflation density forecasts, we now examine real interest rate density forecasts. The real interest rate density is a simple

sign change and location shift of the inflation density:

$$f(r_{t,t+1}) = i_{t,t+1} - f(\pi_{t,t+1}), \tag{16}$$

where r denotes the real interest rate, i denotes the nominal interest rate, and π denotes inflation. Real interest rate densities are of course driven by the inflation densities via equation (16), but it is nevertheless interesting to make the translation from inflation into the real cost of borrowing.

In Figure 8 we show the simple average and best $\leq$4-average real interest rate density forecasts, and in Figure 9 we show the differences between them, together with the realizations.[22] One is immediately struck by the high probability assigned to negative real rates through much of the sample. $P(r_{t,t+1})<0$ is, for example, routinely greater than 1/2 since the end of the Great Recession, and the realized real rates often *are* negative.

Nevertheless our earlier inflation patterns and lessons remain firmly intact, because real interest rate density forecasts are driven by inflation density forecasts. There are two clear real interest rate "regularization regimes," demarcated by the onset of the Great Recession. In the first, the best-average regularization pushes real interest rate densities downward, because, as discussed earlier, regularization pushes inflation densities upward. In the second, the regularization adds dispersion to real interest rate densities, because regularization adds dispersion to inflation densities.

5.4 Discussion

Although our earlier Monte Carlo of section 4 suggested the possibility of modest gains from additional regularization beyond simplex when using the ex post optimal regularization parameter, there is no guarantee of achieving those gains in practical applications like our ECB-SPF analysis, where determination of a suitable regularization parameter would require data-driven methods (e.g., dynamic cross validation) that may perform poorly in small samples.

The remarkable thing that emerges in both our Monte Carlo and in our ECB-SPF analysis is that the two simplest regularizations – simplex and best-average – do almost as well as the more sophisticated regularizations with ex post optimal regularization parameters, *and*

[22]There is no need to show regularized estimation results for real interest rates, because the log score is invariant to the switch from inflation to real interest rate density forecasts defined by equation (16). There is similarly no need to show real interest rate *PIT* histograms, because they are exact mirror images of the inflation *PIT* histograms in Figure 7, as revealed by equation (16).

they do not require tuning. Hence both simplex and best-average appear highly appealing for practical work.

Simplex has no regularization parameter because it achieves its regularization by direct imposition of two hard constraints, non-negativity and sum-to-one. There is no issue of choosing the "strength" of the regularization – non-negativity and sum-to-one are simply imposed and must hold exactly.

Best-average does have an implicit regularization parameter, the number of forecasters kept for averaging (equivalently, the number of zero weights imposed on various forecasters), which it selects in real time by brute-force determination of the historically best-performing average – effectively performing precisely the sort of dynamic cross validation mentioned above.

Moreover, in our empirical work, although there is no reason why it should be the case in general, the best-average and simplex solutions are nearly identical – simplex assigns zero weight to many forecasters and then equally weights those remaining. This precisely parallels the nearly-identical best-average and egalitarian lasso point forecast combinations in Diebold and Shin (2019).

The upshot is that, at least for the analyses undertaken in this paper, there is little gain from regularization beyond simplex or best-average.

6 Concluding Remarks and Directions for Future Research

We have proposed methods for constructing regularized mixtures of density forecasts, exploring a variety of objectives and penalties, which we used in substantive explorations of Eurozone inflation and real interest rate survey density forecasts. All individual survey forecasters (even the ex post best forecaster) are outperformed ex ante by our regularized mixtures. The log scores of the simplex and best-average mixtures, for example, are approximately 15% better than that of the ex post median forecaster, and 7% better than that of the ex post best individual forecaster.

Before the Great Recession, regularization tends to correct for bias, shifting inflation density locations upward toward higher inflation, and hence real interest rate density locations downward toward low or negative real rates. From the Great Recession onward, the situation is very different – regularization tends to correct for overconfidence, moving probability mass from the centers to the tails of both inflation and real interest rate density forecasts.

A variety of avenues for future research are apparent. First, our empirical work did not emphasize mixture regularization methods that require hyperparameter selection (simplex+ridge or simplex+entropy), because at least in this paper's empirical analyses they added little to the simpler methods that impose sparsity (simplex, best-average). But the benefits of sparsity may be illusory, as emphasized by Giannone et al. (2021), so the more sophisticated regularizations may prove useful in other contexts and clearly represent additional empirical exploration. An obvious issue is the efficacy of feasible real-time hyperparameter selection.

Second, one could use the probability integral transform as a regularized mixture estimation objective, minimizing a goodness-of-fit statistic (e.g., Kolmogorov-Smirnov) for testing the joint hypothesis of an $iid\,U(0,1)$ probability integral transform.

Third, one could broaden our mixture approach to allow flexibly time-varying mixture weights as in Jore et al. (2010), and mixture weights that vary over regions of density support, as in Kapetanios et al. (2015).

Finally, one could explore non-mixture approaches to density forecast combination. For example, one could combine conditional mean point forecasts in the usual way (e.g., Diebold and Shin (2019)) and then build up the full conditional density from the combining-regression residuals. This is related to the work of Gneiting et al. (2005). It may also be challenging, however, as indicated by more recent work such as Hounyo and Lahiri (2021).

Appendices

A Derivation of the Simplex+Entropy Regularized Estimator

The simplex+entropy estimator solves the optimization problem:

$$\hat{\omega} = \arg\min_{\omega} \left(\underbrace{-\sum_{t=1}^{T} \log\left(\sum_{k=1}^{K} \omega_k f_{k,t}(y_t)\right)}_{\text{log score}} + \underbrace{(\alpha - 1)\left(-\sum_{k=1}^{K} \log(\omega_k)\right)}_{\text{entropy penalty}} \right) \tag{A1}$$

$$\text{s.t. } \omega_k \in (0,1), \ \sum_{k=1}^{K} \omega_k = 1.$$

As we will show, this arises as the posterior mode in a Bayesian analysis with (1) log likelihood given by the log score, and (2) Dirichlet prior, which puts positive probability only on the unit simplex but also shrinks toward equal weights for a certain hyperparameter configuration.[23] In particular, the K-dimensional Dirichlet prior is governed by K hyperparameters, and when they are equal, the prior mean is $1/K$. Hence the simplex+entropy regularization (8) with equal prior hyperparameters does the same thing as simplex+ridge (5): Impose simplex and shrink toward equal weights.

A.1 Prior

The Dirichlet prior on $\omega = (\omega_1, \omega_2, ..., \omega_K)$ with hyperparameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$ is

$$f_D(\omega; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \omega_k^{\alpha_k - 1},$$

where $B(\cdot)$ is the beta function, $\alpha_k > 0 \ \forall k \in 1, ..., K$, and the support of ω is $\omega_k \in (0,1)$ with $\sum_{k=1}^{K} \omega_k = 1$.

As is well known, the Dirichlet mean and variance are:

[23] Formally, we should say "Bayesian-inspired" rather than Bayesian, "pseudo-likelihood" rather than likelihood (because a scoring rule is not a likelihood), and "pseudo-posterior" rather than posterior. We will refrain from doing so, in an effort to avoid tedious verbiage.

$$E(\omega_i) = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}$$

and

$$var(\omega_i) = \frac{\frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}\left(1 - \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}\right)}{1 + \sum_{k=1}^{K} \alpha_k}.$$

Hence when $\alpha_1 = \alpha_2 = ... = \alpha_K = \alpha$, we have

$$E[\omega_k] = 1/K$$

and

$$Var(\omega_k) = \frac{K-1}{\alpha K^3 + \ K^2},$$

for all $k = 1, ..., K$. That is, the prior is centered on equal weights $1/K$, and $var(\omega_k) \to 0$ as $\alpha \to \infty$, so that α governs prior precision, with larger α producing heavier shrinkage toward $1/K$.

A.2 Posterior

The posterior distribution is

$$f_D(\omega|y;\boldsymbol{\alpha}) = \underbrace{\prod_{t=1}^{T}\left(\sum_{k=1}^{K}\omega_k f_{k,t}(y_t)\right)}_{\text{likelihood}} \times \underbrace{\frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\omega_k^{\alpha-1}}_{\text{prior}} \times \frac{1}{p(y)},$$

so the log posterior is

$$\log f_D(\omega;\boldsymbol{\alpha}) = \sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}\omega_k f_{k,t}(y_t)\right) + (\alpha-1)\sum_{k=1}^{K}\log(\omega_k) - \log B(\boldsymbol{\alpha}) - \log p(y).$$

Because $B(\boldsymbol{\alpha})$ and $p(y)$ do not depend on $\boldsymbol{\omega}$, we can drop the last two term, so the posterior mode is

$$\hat{\omega} = \arg\min_{\omega}\left(\underbrace{-\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}\omega_k f_{k,t}(y_t)\right)}_{\text{Log score}} + \underbrace{(\alpha-1)\left(-\sum_{k=1}^{K}\log(\omega_k)\right)}_{\text{penalty}}\right) \tag{A2}$$

$$\text{s.t. } \omega_k \in (0,1), \ \sum_{k=1}^{K} \omega_k = 1.$$

A.3 Understanding the Penalty Term

One way to understand the penalty term is to recall the solution to the empirical likelihood maximization problem of Owen (2001),

$$\arg\min_{\omega} \left(-\sum_{k=1}^{K} \log(\omega_k) \right)$$

$$\text{s.t. } \omega_k \in (0,1), \ \sum_{k=1}^{K} \omega_k = 1,$$

which is equal weights, $\omega_k{=}1/K$, $\forall k$. Hence we see that the penalty part of (A2) is minimized at $\omega_k{=}1/K$, which yields a clear interpretation of the penalty term. Larger α means a tighter prior on ω, with heavier shrinkage toward equal weights. Several interesting limiting cases emerge. First, for $\alpha{\to}\infty$, the penalty term dominates, and the optimal solution is equal weights. Second, for $\alpha{\to}1$, the penalty term vanishes, and the optimal solution matches that of the optimal linear pool, with simplex constraint imposed. Third, there is an upper bound for $var(\omega_k)$: as $\alpha{\to}0$, $var(\omega_k){\to}(K-1)/K^2$.

A.4 Remarks

1. The entropy regularization optimization problem is convex, because both the log-score and the penalty are convex. A closed form may not exist for the regularized ω, but convexity makes numerical computation straightforward.

2. Entropy regularization has a clear parallel to ridge regularization. As is well known, ridge regularization emerges as the posterior mode in a Bayesian analysis with Gaussian prior, and as we have shown, entropy regularization emerges as posterior mode in a Bayesian analysis with Dirichlet prior. Both regularizations, moreover, are governed by a single parameter linked to prior precision.

3. If the effects of the ridge and entropy penalties are very similar in certain respects (imposition of simplex and shrinkage toward $1/K$), their full Bayesian interpretations are nevertheless different. In particular, the ridge (Gaussian) and entropy (Dirichlet)

priors differ, even if their means are the same $(1/K)$, and so the posteriors differ. For $\alpha < 1$ the Dirichlet prior distribution may not even have a single mode.

B Monte Carlo Results for Fully-Rational Forecasts

Recall the DGP used in the Monte Carlo experiment reported in the main text:

$$y_t = x_t + \sigma_y e_t \tag{B1}$$

$$x_t = \phi_x x_{t-1} + \sigma_x v_t$$

$$z_{kt} = x_t + \sigma_{zk}\eta_{kt},$$

where y_t is the variable to be forecast, x_t is the long-run component of y_t, the z_{kt} are heterogeneous independent noisy signals about x_t received by the forecasters, and all shocks are iid $\mathrm{N}(0,1)$. The signal is leading in that z_{kt+1} is available to forecaster k at time t.

In the main text we provide our forecasters with incompletely-rational forecasts. In particular, the mean of each forecaster's predictive distribution is the observed signal z_{kt+1}. However, given that x_t is serially correlated and hence z_{kt-1} could provide additional information about x_t over z_{kt}, construction of fully-rational mean forecasts of x_t requires incorporating information from lagged signals. As a result, we replace z_{kt+1} with $\hat{x}_{kt+1} = E\left[x_{t+1} \mid z_{kt+1}, z_{kt}, \ldots, z_{k1}\right]$ and repeat the Monte Carlo with these fully-rational forecasts.

To obtain the optimal extraction of x_{t+1} given all observed signals for forecaster k, we apply the Kalman filter to the state-space system:

$$z_{kt} = x_t + \sigma_{zk}\eta_{kt} \tag{B2}$$

$$x_t = \phi_x x_{t-1} + \sigma_x v_t,$$

where all shocks are iid $\mathrm{N}(0,1)$. We run the Kalman filter using the true values of σ_{zk}, ϕ_x, and σ_x, because we assume our forecasters know the underlying DGP. Assuming forecasters have a strong belief that the 1-step-ahead predictive density is Gaussian with variance σ_y^2, forecaster k's fully-rational predictive density is then

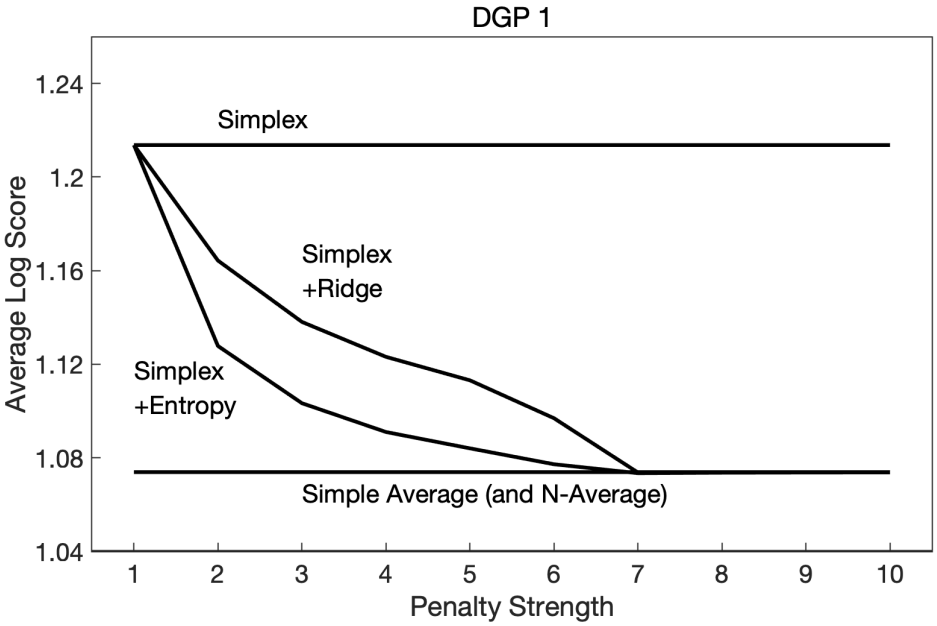
$$p_{kt}(y_{t+1}) = N(\hat{x}_{kt+1}, \sigma_y^2). \tag{B3}$$

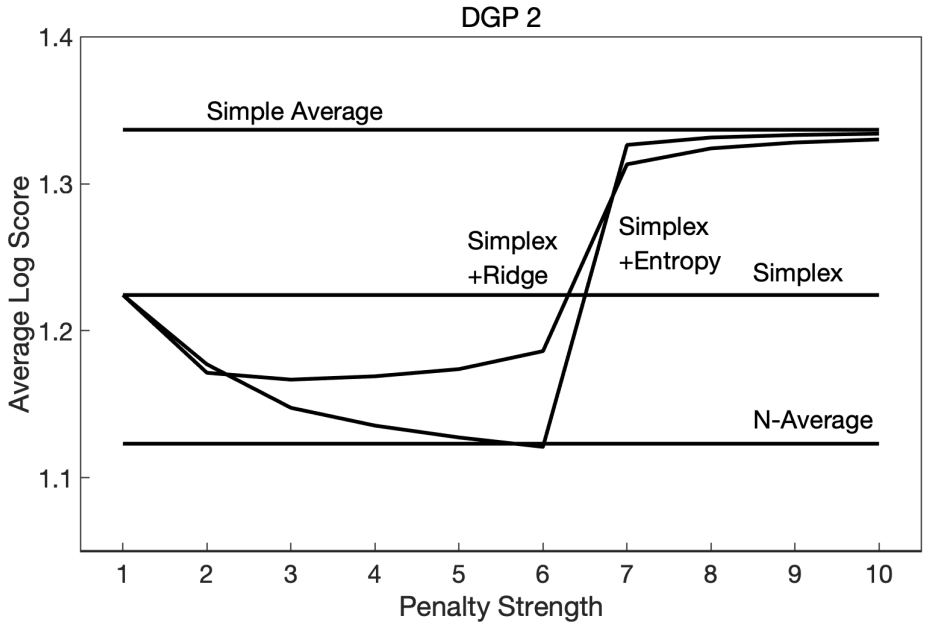
The simulation results, which appear below in Table B1 and Figure B1, are qualitatively identical to those reported in the main text.

Table B1: Average Log Scores, Fully-Rational Forecasts

	DGP 1			DGP 2		
Regularization group	L	#	λ^*	L	#	λ^*
Simplex	1.21	4.70	NA	1.22	4.43	NA
Simplex+Ridge	1.07	20.00	2511.25	1.12	9.86	15.00
Simplex+Entropy	1.07	20.00	5.22	1.17	20.00	0.10
Subset Averages	L	#	λ^*	L	#	λ^*
Best N-Average:						
$N = 1$	1.91	1.00	NA	1.99	1.00	NA
$N = 2$	1.35	2.00	NA	1.39	2.00	NA
$N = 3$	1.25	3.00	NA	1.24	3.00	NA
$N = 4$	1.18	4.00	NA	1.17	4.00	NA
$N = 5$	1.14	5.00	NA	1.16	5.00	NA
$N = 6$	1.13	6.00	NA	1.14	6.00	NA
$N = 7$	1.11	7.00	NA	1.13	7.00	NA
$N = 8$	1.11	8.00	NA	1.12	8.00	NA
$N = 9$	1.10	9.00	NA	1.13	9.00	NA
$N = 10$	1.10	10.00	NA	1.14	10.00	NA
$N = 15$	1.08	15.00	NA	1.24	15.00	NA
$N = 20$	1.07	20.00	NA	1.34	20.00	NA
Best $\leq$ 2-Average	1.37	1.98	NA	1.40	1.98	NA
Best $\leq$ 3-Average	1.28	2.77	NA	1.28	2.73	NA
Best $\leq$ 5-Average	1.23	3.38	NA	1.26	3.27	NA
Best $\leq$ 10-Average	1.23	3.44	NA	1.26	3.31	NA
Best $\leq$ 15-Average	1.23	3.44	NA	1.26	3.31	NA
Best $\leq$ 20-Average	1.23	3.44	NA	1.26	3.31	NA
Comparisons	L	#	λ^*	L	#	λ^*
Best	0.28	1	NA	0.30	1	NA
75%	0.62	1	NA	0.81	1	NA
Median	1.44	1	NA	2.73	1	NA
25%	2.89	1	NA	6.83	1	NA
Worst	6.94	1	NA	17.77	1	NA
Simple Average	1.07	20	NA	1.34	20	NA

Figure B1: Monte Carlo Estimates of Expected Mixture Performance vs Penalty Strength, Fully-Rational Forecasts





References

Aastveit, K.A., J. Mitchell, F. Ravazzolo, and H.K. van Dijk (2019), "The Evolution of Forecast Density Combinations in Economics," *Oxford Research Encyclopedia of Economics and Finance*, https://doi.org/10.1093/acrefore/9780190625979.013.381.

Amisano, G. and J. Geweke (2017), "Prediction Using Several Macroeconomic Models," *Review of Economics and Statistics*, 99, 912–925.

Askanazi, R., F.X. Diebold, F. Schorfheide, and M. Shin (2018), "On the Comparison of Interval Forecasts," *Journal of Time Series Analysis*, 39, 953–965.

Bates, J.M. and C.W.J. Granger (1969), "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451–468.

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, 213–232.

Brehmer, J.R. and T. Gneiting (2021), "Scoring Interval Forecasts: Equal-Tailed, Shortest, and Modal Interval," *Bernoulli*, 27, 1993–2010, also arXiv:2007.05709 [math.ST], https://arxiv.org/abs/2007.05709.

Bresciani-Turroni, C. (1937), *The Economics of Inflation*, Allen and Unwin.

Brier, G.W. (1950), "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3.

Brodie, J., I. Daubechies, C. De Mol, D. Giannone, and I. Loris (2009), "Sparse and Stable Markowitz Portfolios," *Proceedings of the National Academy of Sciences*, 106, 12267–12272.

Busetti, F. (2017), "Quantile Aggregation of Density Forecasts," *Oxford Bulletin of Economics and Statistics*, 79, 495–512.

Chen, N.-F., R. Roll, and S. Ross (1986), "Economic Forces and the Stock Market," *Journal of Business*, 383–403.

Conflitti, C., C. De Mol, and D. Giannone (2015), "Optimal Combination of Survey Forecasts," *International Journal of Forecasting*, 31, 1096–1103.

Czado, C., T. Gneiting, and L. Held (2009), "Predictive Model Assessment for Count Data," *Biometrics*, 65, 1254–1261.

Diebold, F.X. (1991), "A Note on Bayesian Forecast Combination Procedures," In P. Hackl and A. Westlund (eds.), *Economic Structural Change: Analysis and Forecasting,* 225-232, Springer-Verlag.

Diebold, F.X., T. Gunther, and A. Tay (1998), "Evaluating Density Forecasts, with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

Diebold, F.X. and M. Shin (2017), "Assessing Point Forecast Accuracy by Stochastic Error Distance," *Econometric Reviews*, 36, 588–598.

Diebold, F.X. and M. Shin (2019), "Machine Learning for Regularized Survey Forecast Combination: Partially-Egalitarian LASSO and its Derivatives," *International Journal of Forecasting*, 35, 1679–1691.

Elliott, G. (2011), "Averaging and the Optimal Combination of Forecasts," Manuscript, Department of Economics, UCSD.

Elliott, G. and A. Timmermann (2016), *Economic Forecasting*, Princeton University Press.

Epstein, E.S. (1969), "A Scoring System for Probability Forecasts of Ranked Categories," *Journal of Applied Meteorology*, 8, 985–987.

Friedman, M. (1977), "Nobel Lecture: Inflation and Unemployment," *Journal of Political Economy*, 85, 451–472.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013), "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting*, 29, 108–121.

Geweke, J. and G. Amisano (2011), "Optimal Prediction Pools," *Journal of Econometrics*, 164, 130–141.

Giannone, D., M. Lenza, and G.E. Primiceri (2021), "Economic Predictions with Big Data: The Illusion of Sparsity," *Econometrica*, 89, 2409–2437, in press.

Gneiting, T. and A.E. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., A.E. Raftery, A.H. Westveld, and T. Goldman (2005), "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation," *Monthly Weather Review*, 133, 1098–1118.

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

Good, I.J. (1952), "Rational Decisions," *Journal of Royal Statistical Society: Series B*, 14, 107–114.

Gormley, I.C. and S. Frühwirth-Schnatter (2019), "Mixture of Experts Models," .

Granger, C.W.J. and R. Ramanathan (1984), "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3, 197–204.

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

Hounyo, U. and K. Lahiri (2021), "Estimating the Variance of a Combined Forecast: Bootstrap-Based Approach," CREATES Research Paper 2021-14, Department of Economics and Business Economics, Aarhus University.

Jiang, W. and M.A. Tanner (1999), "Hierarchical Mixtures-of-Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation," *Annals of Statistics*, 27, 987–1011.

Jore, A.S., J. Mitchell, and S.P. Vahey (2010), "Combining Forecast Densities from VARs with Uncertain Instabilities," *Journal of Applied Econometrics*, 25, 621–634.

Kapetanios, G., J. Mitchell, S. Price, and N. Fawcett (2015), "Generalised Density Forecast Combinations," *Journal of Econometrics*, 188, 150–165.

Kascha, C. and F. Ravazzolo (2010), "Combining Inflation Density Forecasts," *Journal of Forecasting*, 29, 231–250.

McAlinn, K. and M. West (2019), "Dynamic Bayesian Predictive Synthesis in Time Series Forecasting," *Journal of Econometrics*, 210, 155–169.

Norets, A. (2010), "Approximation of Conditional Densities by Smooth Mixtures of Regressions," *Annals of Statistics*, 38, 1733–1766.

Owen, A. (2001), *Empirical Likelihood*, Chapman and Hall.

Ranjan, R. and T. Gneiting (2010), "Combining Probability Forecasts," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 71–91.

Takanashi, K. and K. McAlinn (2020), "Predictive Properties and Minimaxity of Bayesian Predictive Synthesis," Preprint, RIKEN and Temple University.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Timmermann, A. (2006), "Forecast Combinations," *Handbook of Economic Forecasting*, 135–196.

Wallis, K.F. (2011), "Combining Forecasts–Forty Years Later," *Applied Financial Economics*, 21, 33–41.

Winkler, R.L. and A.H. Murphy (1968), "'Good' Probability Assessors," *Journal of Applied Meteorology*, 7, 751–758.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman (2018), "Using Stacking to Average Bayesian Predictive Distributions," *Bayesian Analysis*, 13, 917–1003.

Yuksel, S.E., J.N. Wilson, and P.D. Gader (2012), "Twenty Years of Mixture of Experts," *IEEE Transactions on Neural Networks and Learning Systems*, 23, 1177–1193.

Zou, H. and T. Hastie (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 67, 302–320.