Federal Reserve Bank of Philadelphia

Working Papers
Research Department

WP 20-31
Revised August 2020
https://doi.org/10.21799/frbp.wp.2020.31

Probability Forecast Combination via Entropy Regularized Wasserstein Distance

Ryan Cumings-Menon
U.S. Census Bureau

Minchul Shin
Federal Reserve Bank of Philadelphia Research Department

ISSN: 1962-5361

Probability Forecast Combination via Entropy Regularized Wasserstein Distance*

Ryan Cumings-Menon† Minchul Shin‡

August 6, 2020

Abstract

We propose probability and density forecast combination methods that are defined using the entropy regularized Wasserstein distance. First, we provide a theoretical characterization of the combined density forecast based on the regularized Wasserstein distance under the Gaussian assumption. Second, we show how this type of regularization can improve the predictive power of the resulting combined density. Third, we provide a method for choosing the tuning parameter that governs the strength of regularization. Lastly, we apply our proposed method to the U.S. inflation rate density forecasting, and illustrate how the entropy regularization can improve the quality of predictive density relative to its unregularized counterpart.

Key words: Entropy regularization, Wasserstein distance, optimal transport, density forecasting, model combination.

JEL codes: C53, E37

*__Acknowledgement:__ We thank Frank Diebold, Roger Koenker and Frank Schorfheide for their insightful comments. **Disclaimer**: The views expressed in these papers are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia, the Federal Reserve System, or the Census Bureau. Any errors or omissions are the responsibility of the authors. There are no sensitive data in this paper.

†The US Census Bureau, 4600 Silver Hill Rd, Suitland-Silver Hill, MD 20746. e-mail: ryan.r.cumings@gmail.com

‡Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106. e-mail: visiblehand@gmail.com.

1 Introduction

In this paper, we study a class of density forecast combination methods based on a Wasserstein metric. In the univariate case, an equally weighted centroid defined by a Wasserstein metric corresponds to a quantile averaging or vincentized center where quantiles of forecast densities are averaged. The resulting combined density tends to be narrower than the linear opinion rule (Geweke and Amisano, 2011; Lichtendahl et al., 2013; Busetti, 2017), which may or not be desirable, depending on the context.

We propose to use the entropy regularized Wasserstein metric to construct a combined density forecast. Like its unregularized counterpart, this combined probability/density can be defined by an optimization problem, but the optimization problem in this case includes an additional regularization term that penalizes densities with low entropy, which ensures the combined density forecast is smooth. One advantage of this approach is that the entropy regularized Wasserstein barycenter can be found in a much more computationally efficient manner than its unregularized counterpart when the input densities are multi-dimensional (Benamou et al., 2015). While computational efficiency is the most commonly cited reason for using entropy regularization, this paper demonstrates that there is an additional advantage of regularization when it comes to the density combination problem. It provides a way to tune the degree of dispersion of the combined density forecast. To the best of our knowledge, this regularized metric has not been explored in the context of the density forecasting combination problem.

We proceed as follows. Section 2 formulates a density forecast combination problem with a general metric. Several existing aggregation methods in the literature can be formulated with the choice of a specific metric within this unified framework. After discussing these existing approaches, we introduce our proposal of using the entropy regularized Wasserstein barycenter. Section 3 provides theoretical results that describe the impact of entropy regularization on the combined density under a Gaussian assumption and discusses how this helps improve the quality of the combined density prediction. Section 4 discusses how to set the strength of the entropy regularization in practice and shows that our proposed selection rule achieves a certain notion of optimality. Section 5 provides an empirical exercise that illustrates how entropy regularization improves the quality of density prediction of the U.S. inflation rate relative to the unregularized combined density forecast. Section 6 concludes.

2 Regularized Wasserstein barycenter for density forecast combination

This section introduces the density combination problem; see, for example, Timmermann (2006). We assume that agent $i \in \{1, \ldots, N\}$ at time $t \in \mathbb{N}_+$ provides a forecast of the density function $p_{it} : \mathbb{R}^d \to \mathbb{R}_+$, with distribution function denoted by $P_{it} : \mathbb{R}^d \to \mathbb{R}_+$, of the random variable y_{t+h} with $h \in \mathbb{N}_+$. We are interested in aggregating information contained in the N agents' forecasts to generate a better predictive distribution for y_{t+h}.

Throughout the paper, we shall focus on density combinations that can be viewed as a type of average over probability densities. Specifically, those that can be defined as,

$$\overline{p}_t = \arg\min_{p_t \in \mathcal{P}} \sum_{i=1}^{N} D(p_{it}, p_t), \tag{1}$$

where $D(p_i, p_j)$ is a measure of the discrepancy between the densities p_i and p_j. When $D(\cdot)$ satisfies the usual properties of a distance metric, which is the case when $D(\cdot)$ is defined as Euclidean or an unregularized Wasserstein metric, then $\overline{p}_t$ is known as a Fréchet mean, which is a generalization of the average for real numbers. We will refer to $\overline{p}_t$ as a barycenter to also encompass the more general case in which $D(\cdot)$ is not a metric. As described in Eqn (1), we restrict our attention to the case in which $\overline{p}_t$ is a density forecast with each input density having equal weight, which is known to perform quite well as a combination forecast (Clemen, 1989).

A specific choice of metric, $D(p_i, p_j)$, will lead to a different combined density, $\overline{p}_t$. Before introducing our proposed definition of $D(\cdot)$, the entropy regularized Wasserstein metric, the next two sections introduce choices for $D(p_i, p_j)$ that lead to well-known density forecast combination methods.

2.1 Equal-weighted linear opinion rule

As a starting point let us consider $D(p_i, p_j) := \|p_i - p_j\|_2^2$. Then, Eqn (1) becomes

$$\overline{p}_t = \arg\min_{p_t \in \mathcal{P}} \sum_{i=1}^{N} \int (p_{it} - p_t)^2, \tag{2}$$

which results in the following solution,

$$\overline{p}_t = \frac{1}{N}\sum_{i=1}^{N} p_{it}. \tag{3}$$

This can be derived using the first-order condition with respect to p_t, which is $\sum_{i=1}^{N}(p_{it}-\overline{p}_t) = 0$.

This solution is known as the linear opinion rule with equal-weighting. This is the prototypical aggregation methods both in the forecasting literature and in practice; see, for example, Geweke and Amisano (2011). This is a particularly tractable density combination method, as it is equivalent to a mixture density, and it has the additional advantage of being computationally tractable to compute. However, one disadvantage is that it does not preserve the shape of the individual forecast densities. For example, when combining two uni-modal densities, the resulting solution is generally bi-modal.

2.2 Quantile aggregation and the Wasserstein barycenter

In this section we consider the case in which $D(\cdot)$ is defined as the $p-$Wasserstein metric, which is defined as

$$W_p(p_{it}, p_{jt}) = \left(\inf_{\varphi\in\Omega(p_{it},p_{jt})}\int ||z_i - z_j||^p \mathrm{d}\varphi(z_i, z_j)\right)^{1/p}, \tag{4}$$

where $\Omega(p_{it}, p_{jt})$ is the set of all joint distributions $\varphi(z_i, z_j)$ that have marginal densities given by p_{it} and p_{jt}, respectively. Formally, we write

$$\Omega(p_{it}, p_{jt}) = \left\{\varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^1_+ | \forall A \subset \mathbb{R}^d,\ \varphi(A, \mathbb{R}^d) = p_{it}(A) \text{ and } \varphi(\mathbb{R}^d, A) = p_{jt}(A)\right\}. \tag{5}$$

In other words, each $\varphi \in \Omega(p_{it}, p_{jt})$ is a coupling between the distributions p_{it} and p_{jt}. In the optimal transport literature, the minimizer of (4) is also known as the optimal transport plan. This is because, for any $A, B \subset \mathbb{R}^d$, $\varphi(A, B)$ can be interpreted as the amount of mass that is moved from A to B in order to minimize $E\left(\|z_i - z_j\|_p^p\right)$ where $z_i \sim p_{it}$ and $z_j \sim p_{jt}$. For more detail on the field of optimal transport, see Villani (2003) and Galichon (2018).

A special case of this Wasserstein barycenter has a close relation to a recently proposed probability/density forecast combination method in the forecasting literature. More specifically, suppose that input densities are univariate, and $\overline{p}_t$ is defined as the squared Wasserstein

metric, denoted by $D(\cdot) := W_2^2(\cdot)$; in this case, we have,

$$\overline{P}_t^{-1}(\tau) = \frac{1}{N}\sum_{i=1}^{N} P_{it}^{-1}(\tau), \quad \text{for all } \tau \in (0,1), \tag{6}$$

where $P_{it}^{-1}(\cdot)$ and $\overline{P}_t^{-1}(\cdot)$ are the quantile function of agent i and of the combination method, respectively. This forecast aggregation rule is also known as "quantile aggregation" or "Vincentized distribution" (Ratcliff, 1979; Lichtendahl et al., 2013; Busetti, 2017).[1]

The Wasserstein barycenter is known to preserve the shape of input densities, such as log-concavity (Genest, 1992). For example, Agueh and Carlier (2011) show that the Wasserstein barycenter of the inputs, $N(\mu_1, S_1)$ and $N(\mu_2, S_2)$, is $N((\mu_1+\mu_2)/2,\ S)$, where S is the solution of,

$$S = \left(S^{1/2}S_1S^{1/2}\right)^{1/2}/2 + \left(S^{1/2}S_2S^{1/2}\right)^{1/2}/2; \tag{7}$$

see also, Knott and Smith (1994). This is different than the linear opinion rule, which leads to two-normal mixture density with mean $(\mu_1+\mu_2)/2$ and variance $\frac{\sigma_1^2+\sigma_2^2}{2} + \frac{(\mu_1-\mu_2)^2}{4}$, which, in contrast, can be expected to be bi-modal whenever $\mu_1 \neq \mu_2$.

Another difference between these two aggregation methods is that the variance of the Wasserstein barycenter is smaller than that of the combined density resulting from a linear opinion rule. This holds for a more general class of input densities as shown in Lichtendahl et al. (2013) in the univariate case. Of course, a narrow (i.e., sharp) predictive density can be good or bad depending on the underlying distribution of the target variable. It may be desirable to have an ability to flexibly adjust the dispersion of the combined density.

2.3 Regularized Wasserstein barycenter

Now, we turn to our proposal. In this paper, we use a regularized Wasserstein distance (Cuturi, 2013; Peyré and Cuturi, 2019) to combine individual probability forecasts. The regularization term used in this approximation of the Wasserstein metric is given by the negative differential entropy, which, when φ is an absolutely continuous measure, we will define as, $h(\varphi) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log\left(\frac{\mathrm{d}\varphi}{\mathrm{d}\lambda}\right)\mathrm{d}\varphi$, where λ is the Lebesgue measure, and infinity otherwise.

[1] We prefer the representation of Eqn (1) because this definition can be easily extended to higher dimensional densities or mixed data types (when some inputs are continuous and others are discrete) unlike quantile aggregation.

We will use $h(\varphi)$ to define the regularized Wasserstein metric as

$$W_{p,\gamma}(p_{it}, p_{jt}) = \left(\inf_{\varphi \in \Omega(p_{it}, p_{jt})} \int ||z_i - z_j||^p \mathrm{d}\varphi(z_i, z_j) + \gamma h(\varphi) \right)^{1/p}, \tag{8}$$

where $\gamma > 0$ controls a strength of regularization. Note that φ is constrained by the same two marginal restrictions as its unregularized counterpart, as described in the definition of $\Omega(p_{it}, p_{jt})$. This form of regularization is originally introduced by Cuturi (2013) in order to estimate the Wasserstein metric in a computationally efficient manner using the iterative proportional fitting procedure (IPFP) provided by Sinkhorn (1967).

When $\gamma = 0$, there is no regularization, so we have $W_{p,0}(p_{it}, p_{jt}) = W_p(p_{it}, p_{jt})$. One can also show that the optimal coupling, say $\varphi^\star_\gamma$, satisfies $\lim_{\gamma \to 0^+} \varphi^\star_\gamma = \varphi^\star_0$ when $\varphi^\star_0$ is uniquely defined, and otherwise this limiting value is given by the element of the set of optimal unregularized couplings with maximum entropy. Higher values of γ place more weight on the second term in the objective function, which results in optimal couplings that are smoother and more dispersed than their unregularized counterparts.

Defining $D(p_{it}, p_{jt})$ by $W^2_{2,\gamma}(p_{it}, p_{jt})$ results in the combined density

$$\overline{p}_t = \arg\min_{p_t \in \mathcal{P}} \sum_{i=1}^{N} W^2_{2,\gamma}(p_{it}, p_{jt}), \tag{9}$$

which is known as the regularized Wasserstein barycenter. Benamou et al. (2015) provided a generalization of the IPFP procedure to find this barycenter in a computationally efficient manner. While computational efficiency is the commonly cited reason for using entropy regularization, as we will see in the later sections, our motivation for regularization is not entirely computational.

For the rest of the paper, we study this regularized Wasserstein barycenter, which is $\overline{p}_t$ defined in Eqn (1) using (8). First, we present analytical results under a parametric assumption that broadens our understanding about the role of the regularization in forecast density combination. Then, we discuss how one can empirically choose the strength of the regularization that would achieve a certain notion of optimality.

3 Analytical results: The impact of entropy regularization

In this section we provide analytical results that describe the impact of entropy regularization on the shape of the barycenter. To better compare this barycenter with its unregularized counterpart in the Gaussian case, as defined above, we will focus on the regularized barycenter when p_1 and p_2 are d-dimensional multivariate Gaussians ($d \geq 1$). The regularized Wasserstein barycenter in this case is defined as

$$\overline{p} \in \arg\min_{q} \left(\mathcal{W}_\gamma^2(p_1, q) + \mathcal{W}_\gamma^2(p_2, q) \right). \tag{10}$$

The following theorem completely characterizes the resulting barycenter in this case. Like the unregularized case, the theorem shows that regularization does not impact the mean of the barycenter; however, it does have an impact on its variance-covariance matrix.

Theorem 1: *Let p_1 and p_2 be Gaussian density functions with means $\mu_1, \mu_2 \in \mathbb{R}^d$, and variance matrices, $S_1, S_2 \in \mathbb{R}^{d \times d}$. The regularized Wasserstein barycenter between p_1 and p_2 is given by the density function of $N(\mu_B, S_B)$, where $\mu_B \in \mathbb{R}^d$ and $S_B \in \mathbb{R}^{d \times d}$ are defined by,*

$$\begin{aligned}
\mu_B :=& (\mu_1 + \mu_2)/2 \\
S_B :=& \left(V/\gamma + I\right)^{-1} \left(V/2 + I\gamma/2 + S_2\right) \left(V/\gamma + I\right)^{-1} \\
=& \left(-V/\gamma + I\right)^{-1} \left(-V/2 + I\gamma/2 + S_1\right) \left(-V/\gamma + I\right)^{-1},
\end{aligned}$$

where $V \in \mathbb{R}^{d \times d}$ is the unique symmetric matrix that satisfies these equalities and $-I\gamma < V < I\gamma$.

Also, the iterates of the following series converge to V when $V^{(0)} := \mathbf{0}_{d \times d}$,

$$V^{(k+1)} = S_2 - S_1 + S_1 \left(S_1 + I\gamma/2 - V^{(k)}/2\right)^{-1} S_1 - S_2 \left(S_2 + I\gamma/2 + V^{(k)}/2\right)^{-1} S_2.$$

The proof of this result is included in the Appendix. We prove a slightly more general version of the theorem where the objective function in Eqn (10) is a weighted average of $\mathcal{W}_\gamma^2(p_1, q)$ and $\mathcal{W}_\gamma^2(p_2, q)$. The proof first finds a system of equations that holds only in the case in which the regularized barycenter is Gaussian. Afterward, a fixed point theorem provided by Ran and Reurings (2004) for mappings on partially ordered sets is used to show that this

system of equations has a unique solution, so an implication is that the system characterizes the regularized barycenter when both input densities are Gaussian.

Now, we discuss our theoretical results and their implication to the density forecast combination problem.

Remark 1 (on location). Regularization does not affect the mean of the resulting barycenter. The joint entropy of the multivariate normal does not depend on the location, and we conjecture that this is the reason why the regularization only affects the covariance matrix in a Gaussian case.

In the generic case this property would require that the domain is unbounded. This property about the location does not hold for some input densities. For example, if the domains of p_1 and p_2 are both $[0, 1]$, then the optimal coupling between p_1 and q converges to the distribution on $[0, 1] \times [0, 1]$ with maximum entropy that has one marginal equal to p_1 as γ diverges. This implies the other marginal, which defines q, is the uniform distribution on $[0, 1]$. Similar logic applies to the optimal coupling between p_2 and q. Thus, $\lim_{\gamma\to\infty} E_{x\sim q}(x) = 1/2$, regardless of the means of the input densities. Then, an immediate implication of differentiability of $\frac{dW^2_{2,\gamma}(p_i,q)}{dq}$ with respect to γ is that γ impacts the mean of the barycenter except for some special cases when p_1 and p_2 are such that the resulting barycenter is already centered around $1/2$.

Remark 2 (on dispersion). Regularization tends to smooth the resulting barycenter, leading to a more dispersed combined density. To understand this point, let us consider a simple example below.

Simple example. Consider a case with univariate $p_{it} = N(\mu_{it}, \sigma^2)$ and $N = 2$. Then, the original Wasserstein barycenter (quantile averaging) is $\overline{p}_t = N((\mu_{1t} + \mu_{2t})/2,\ \sigma^2)$. On the other hand the regularized Wasserstein barycenter is $\overline{p}_t(\gamma) = N((\mu_{1t} + \mu_{2t})/2,\ \sigma^2 + \gamma/2)$.

As this case exemplifies the strength of the regularization controls a dispersion of the combined density. The heavier the regularization the greater dispersed (or, the smoother) density we obtain. This result highlights that the entropy regularization offers an extra flexibility to control the dispersion of the combined density. In the next section, we propose a data-driven way to select the value of γ, the strength of the regularization.

Remark 3. The normality assumption that we made to obtain the closed-form solution for the barycenter is not needed in practice. The regularized barycenter of probability/density forecasts is well-defined and computationally tractable for a broader context. One can have multiple inputs, non-Gaussian densities, discrete/continuous/mixed distribution. This includes many interesting and empirically relevant situations in economic forecasting such as macroeconomic and financial forecasting. Benamou et al. (2015) and Solomon et al. (2015) describe the generalizations of IPFP that are used to calculate the regularized barycenter in practice.

4 On choosing the strength of the regularization

This section discusses how to choose the strength of the penalization. Our empirical strategy is to select γ by the value that most accurately fits the observed data.[2] To do so, we regard the regularized barycenter computed at time t, $\overline{p}_t$, as a predictive likelihood for y_{t+1}. This predictive likelihood interpretation of the barycenter can be formally justified by the principal-agent framework similar to the one developed by Del Negro et al. (2016). Suppose we have collected the regularized barycenters and the realized value of the target variable from the initial period (1) to present (t). We write this collection as $\mathcal{I}_t$. Then, we can define a maximum likelihood estimator for γ at t with $\mathcal{I}_t$ as

$$\widehat{\gamma}_{1:t}^{mle} \in \arg\max_{\gamma \geq 0} \sum_{\tau=1}^{t-1} \log \overline{p}_\tau(y_{\tau+1}; \gamma), \tag{11}$$

and, the combined density prediction for y_{t+1} at time t is

$$\widehat{p}(y_{t+1}|\mathcal{I}_t) = \overline{p}_t(y_{t+1}; \widehat{\gamma}_{1:t}^{mle}). \tag{12}$$

There is a notion in which this combined density with $\widehat{\gamma}$ is optimal. Suppose that $y_t \sim_{i.i.d.} p^*(y)$, and assume that forecasters report a sequence of predictive densities, $p_i(y)$ for y_t, $t = 1, 2, ..., T$ and $i = 1, 2, ..., N$. These forecasts are reported before the realization of y_t, and the barycenter $\overline{p}(y; \gamma)$ is defined by $p_i(y)$'s and $\gamma > 0$. Then, the following can be shown

[2] To economize our notation we restrict our discussion to the 1-step-ahead prediction (i.e., $h = 1$).

under regularity conditions,

$$\frac{1}{T}\sum_{t=1}^{T}\log\overline{p}(y_t;\gamma) \to_p \int \log\overline{p}(y;\gamma)p^*(y)dy \quad \text{as } T\to\infty,$$

for $\gamma \in \Gamma \in \mathbb{R}_+$. In turn, a maximizer of the left-hand-side term also converges to the maximizer of the right-hand-side term, which is a minimizer of

$$KL(\overline{p}(y;\gamma), p^*(y)) = -\int \log\overline{p}(y)p^*(y)dy + \int \log(p^*(y))p^*(y)dy.$$

Therefore, $\widehat{\gamma}$ converges to a point so-called the pseudo-true parameter that minimizes Kullback-Leibler (KL) divergence from the regularized barycenter to the true data generating process. In other words, we find γ that makes the resulting barycenter close to the true data generating process in the limit. This asymptotic thought experiment can be justifiable under quite general conditions allowing for a range of serial dependence in y_t as well as a flexible form of the regularized Wasserstein barycenter implied by $p_{i,t-1}(y_t)$'s. We can operationalize this by recognizing that $\overline{p}_{t-1}(y;\gamma)$ can be viewed as a predictive likelihood for y_t formed at time $t-1$. Then, quasi-MLE theory can be invoked (e.g., White, 1982; Bollerslev and Wooldridge, 1992). We provide a simple example in which the true data generating process follows the autoregressive (AR) process.

Simple example. Suppose that forecaster 1 and 2 use mean-zero Gaussian AR(1) process to construct their density prediction. The two forecasts differ only by the mean reversion parameter. That is, the means of predictive distribution for forecaster 1 and 2 are $\mu_{1t} = \rho_1 y_{t-1}$ and $\mu_{2t} = \rho_2 y_{t-1}$, respectively. Based on our theory in the previous section, the barycenter is $\overline{p}_{t-1}(y;\gamma) = N(\overline{\mu}_t,\ \sigma^2 + \gamma/2)$ where $\overline{\mu}_t = (\mu_{1t} + \mu_{2t})/2$, and the log density of the regularized barycenter at τ for $y_{\tau+1}$ is

$$\log(\overline{p}_\tau(y_{\tau+1};\gamma)) = -1/2\log(2\pi) - 1/2\log(\sigma^2+\gamma/2) - 1/2\left(\frac{y_{\tau+1}-\overline{\mu}_{\tau+1}}{\sqrt{\sigma^2+\gamma/2}}\right)^2, \tag{13}$$

and the ML estimator for γ at time t is

$$\widehat{\gamma}_{1:t}^{mle} \in \arg\max_{\gamma\geq 0}\sum_{\tau=1}^{t-1}\left(-1/2\log(2\pi) - 1/2\log(\sigma^2+\gamma/2) - 1/2\left(\frac{y_{\tau+1}-\overline{\mu}_{\tau+1}}{\sqrt{\sigma^2+\gamma/2}}\right)^2\right), \tag{14}$$

which leads to

$$\widehat{\gamma}_{1:t}^{mle} = 2 \times \max\left(\frac{1}{(t-1)}\sum_{\tau=1}^{t-1}(y_{\tau+1} - \overline{\mu}_{\tau+1})^2 - \sigma^2,\ 0\right). \tag{15}$$

Now, suppose that the actual data generating process is

$$y_t = \rho_* y_{t-1} + v_t, \quad v_t \sim_{i.i.d.} N(0, \sigma_*^2). \tag{16}$$

When the simple average of both forecasters' autoregressive parameter equals ρ_*, the ML estimate for γ depends on the true conditional variance, σ_*^2, and forecasters' conditional variance. If the sample variance is larger than that of the forecasters, then γ is chosen so that the resulting regularized barycenter has the same variance as the sample variance. On the other hand, if the sample variance is smaller than that of the forecasters, then γ is set to 0. Note that there is an asymmetry in adjusting the variance of the barycenter. This is natural in that the regularization only makes the resulting density smoother. In practice, this may not be a problem if the practitioner's concern is the combined density being too sharp (e.g., relative to the linear opinion rule).

Note that $\widehat{\gamma}_{1:t}^{mle}$ converges in probability to $\gamma_\infty = 2\max(\sigma_*^2 - \sigma^2, 0)$. The KL divergence between $\overline{p}(y_{t+1}; \gamma)$ and the true conditional density of y_{t+1} at t is minimized at $\gamma = \gamma_\infty$. This confirms that our selection rule for γ aims to fit the data well by shaping the regularized barycenter as close as possible to the data generating process.

5 Empirical illustration

In this section, we illustrate our proposed method using macroeconomic data for the U.S. We consider 14 hypothetical forecasters who produce their own 1-step-ahead forecast about the U.S. inflation rate based on the following vector autoregression (VAR) with three variables,

$$Y_t = \Phi_0 + \sum_{i=1}^{4} \Phi_i Y_{t-i} + e_t, \quad e_t \sim_{i.i.d} N(0, \Sigma), \tag{17}$$

where Y_t is a 3×1 vector that consists three quarterly macroeconomic variables, Φ_0 is a 3×1 vector, $\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Sigma$ are 3×3 matrices. The first two elements of Y_t are common to all 14 forecasters: the annualized quarter-over-quarter inflation rate and real GDP growth rate. They differ by the third element of Y_t. We assign each forecaster a different macroeconomic variable from the FRED-QD database by McCracken and Ng (2020). A detailed description

Table 1: Variable used in empirical exercise

Y_t	Used by	Variable description	FRED-QD Mnemonic
Variable 1	All	Inflation rate	GDPCTPI
Variable 2	All	Real GDP growth rate	GDPC1
Variable 3	Forecaster 1	Real Personal Consumption Expenditures	PCECC96
	Forecaster 2	Industrial Production Index	INDPRO
	Forecaster 3	All Employees: Total Nonfarm	PAYEMS
	Forecaster 4	Housing Starts: Total Privately Owned Housing Units Started	HOUST
	Forecaster 5	Real Manufacturing and Trade Industries Sales	CMRMTSPLx
	Forecaster 6	Real Crude Oil Prices: West Texas Intermediate (WTI)	OILPRICEx
	Forecaster 7	Real Average Hourly Earnings: Manufacturing	CES3000000008x
	Forecaster 8	10-Year Treasury Constant Maturity Minus 3-Month Treasury Bill	GS10TB3Mx
	Forecaster 9	Real Commercial and Industrial Loans	BUSLOANSx
	Forecaster 10	Real Total Assets of Households and Nonprofit Organizations	TABSHNOx
	Forecaster 11	U.S. / U.K. Foreign Exchange Rate	EXUSUKx
	Forecaster 12	Consumer Sentiment (University of Michigan)	UMCSENTx
	Forecaster 13	S&P's Common Stock Price Index: Composite	S&P 500
	Forecaster 14	Real Disposable Business Income	CNCFx

Note: All variables are obtained from the FRED-QD database (McCracken and Ng, 2020). Inflation rate is computed as a log difference of the GDP deflator (GDPCTPI). Real GDP growth rate is computed as a log difference of the real GDP (GDPC1). All other variables are transformed following McCracken and Ng (2020). We use the 2019-11 vintage data.

of the variable used in this exercise is in Table 1.

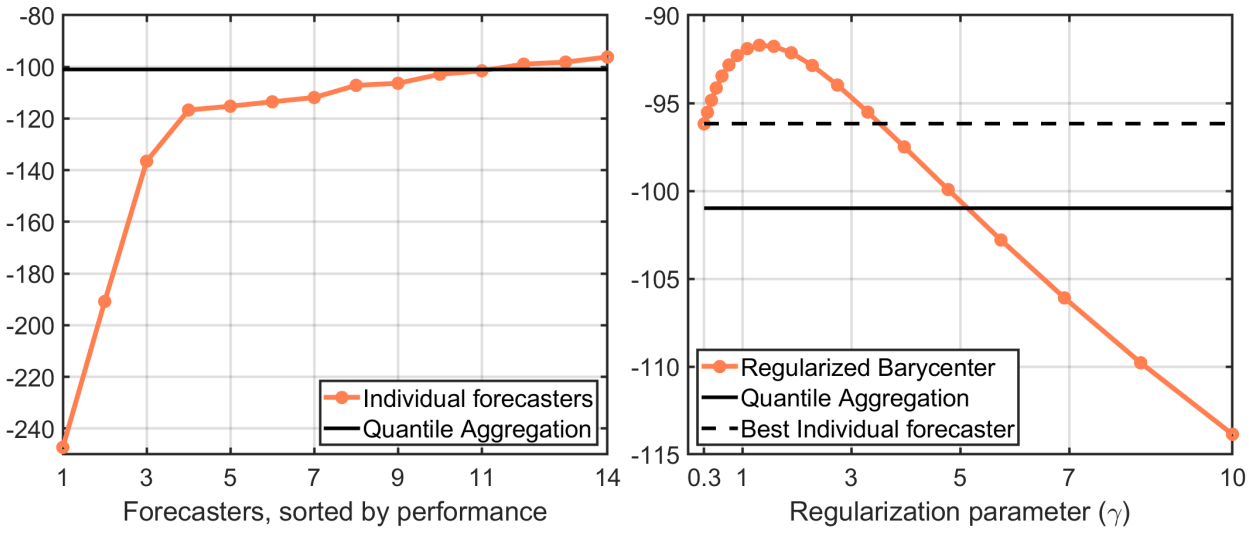
We compute each forecasters' 1-step-ahead predictive distribution for the inflation rate at time t as $\pi_{t+1|t} \sim N([\mu_{t+1|t}]_{(1,1)}, [\Sigma_{t+1|t}]_{(1,1)})$ where $[x]_{(i,j)}$ denotes (i,j) element of vector/matrix x. These forecasters assume that the 1-step-ahead predictive distribution of Y_{t+1} at t is Gaussian, and they use their best guess about the predictive mean and variance to construct the predictive distribution. More specifically, they set these two moments as,

$$\mu_{t+1|t} = \widehat{\Phi}_{0,t} + \sum_{p=1}^{4} Y'_{t-p+1}\widehat{\Phi}_{p,t}, \quad \text{and} \quad \Sigma_{t+1|t} = \widehat{\Sigma}_t, \tag{18}$$

where $(\widehat{\Phi}_{0,t}, \widehat{\Phi}_{1,t}, \widehat{\Phi}_{2,t}, \widehat{\Phi}_{3,t}, \widehat{\Phi}_{4,t}, \widehat{\Sigma}_t)$ is the posterior mean of $p(\Phi_0, \Phi_1, \Phi_2, \Phi_3, \Phi_4, \Sigma | Y_{t:(t-R+1)})$ with a flat prior. We set $R = 80$, meaning that they also use the most recent 20 years of data to construct the predictive distribution.

We let the forecasters to generate their 1-step-ahead predictive distribution for the inflation rate from 2001Q1 to 2018Q4. This leaves us 72 quarters for a forecast evaluation sample. At each point in time, we also combine these 14 predictive densities based on the regularized Wasserstein barycenter with 20 different values of the regularization parameter γ on $[0.3, 10]$. As we explained in the previous section, a larger value of this parameter implies

Figure 1: Sum of log predictive score for U.S. inflation rate (2000Q1-2018Q4)



a stronger regularization, and the resulting combined predictive density becomes smoother with a larger variance. We also compute the combined density with $\gamma = 0$, which leads to "quantile aggregation" or "Vincentized distribution."

We evaluate each forecasters, and other forecast aggregation methods by the sum of log predictive score, which is a logarithm of the predictive density evaluated at the actualized value, over the evaluation sample. These results are presented in Figure 1. The left panel presents the sum of the log score for individual forecasters sorted by their performance. There is a sizeable difference in their historical performance. The solid line represents the performance based on the quantile aggregation, which aggregates all forecasters in the pool. As found by other research papers (e.g., Lichtendahl et al., 2013; Busetti, 2017) the quantile aggregation method generates a decent predictive distribution, which performs slightly better than the ex-post top 4 forecaster.

The right panel in Figure 1 shows the historical performance of our proposed approach with various choices of regularization parameter, γ. For a wide range of values for γ the regularized barycenter performs better than the quantile aggregation. And it does even better than the best individual. This is interesting because we cannot identify the best forecaster a priori.

The optimal value of γ defined in Eqn (11) at the end of the evaluation sample would be the value of γ that corresponds to the peak of the curve, which is about $\widehat{\gamma}_{2018Q4} \approx 1.3$. If we were to use this value at the beginning of the evaluation sample, then the mean difference

in the log predictive score between the regularized Wasserstein barycenter and the quantile aggregation would have been 0.12 with the heteroscedasticity and autocorrelation consistent (HAC) standard error being 0.07. This implies that the difference in the peak of the curve and the solid line is statistically significant at 10% confidence level.

To make the γ selection fully adaptive, we also compute the optimal γ sequentially from the beginning to the end of the evaluation sample. Even in this case the regularized Wasserstein barycenter performs better than the best individual forecaster and the quantile aggregation. More specifically, the sum of the log predictive score is -93.09, and the mean difference in the log predictive score with the quantile aggregation is 0.11 with the HAC standard error being 0.06. This suggests that the regularized Wasserstein barycenter with the adaptively chosen γ performs statistically better that its unregularized counterpart, the quantile aggregation, at 10% significance level.

6 Concluding remarks

This paper proposes to use the entropy regularized Wasserstein barycenter to combine several probability and density forecasts. The entropy regularization smooths the resulting combined forecast, and it offers a flexible way to adjust the dispersion of the predictive density when it is needed. We study the effect of the regularization on the combined density forecast and provide an exact relationship between the strength of the regularization and the variance-covariance matrix of the combined density when input densities are Gaussian. We then provide a way to select the strength of regularization by choosing the regularized barycenter that most closely matches the data. We apply our proposed methodology to the U.S. inflation density forecasting and show how the entropy regularization can improve the quality of the density forecast relative to its unregularized counterpart.

In this article, we restrict weights of each input densities on the final combined density to be equal. This choice was intentional to focus on studying the role of entropy regularization. In practice, however, it is possible that a subset of input densities might be superior to others, and one may wish to put different weights on each input density. Alternatively, it is desirable to include only a subset of input densities into the combined density and set other weights to zero (see, for example, Diebold and Shin, 2019). For those cases, it is fruitful to develop a data-dependent method that chooses both the regularization strength and those weights simultaneously, which is a topic for future research.

Appendix

Ran and Reurings (2004) provide the following fixed point theorem, which we will use in the proof of Theorem 1.

Lemma 1 (Ran and Reurings, 2004): *Let T be a partially ordered set such that every pair $x, y \in T$ has a lower bound and an upper bound. Furthermore, let d be a metric on T such that (T, d) is a complete metric space. If $\mathcal{F} : T \to T$ is a continuous, monotone (e.g., either order-preserving or order-reversing) map from T into T such that,*

$$\exists\, c \in (0,1) :\ d(\mathcal{F}(x), \mathcal{F}(y)) < c d(x, y),\ \forall x > y$$

and

$$\exists\, x_0 \in T :\ \mathcal{F}(x_0) > x_0 \text{ or } \mathcal{F}(x_0) > x_0,$$

then $\mathcal{F}$ has a unique fixed point, $x^\star \in T$. Also, for all $x \in T$,

$$\lim_{n\to\infty} \mathcal{F}^n(x) = x^\star.$$

The following result follows from Lemma 1.

Lemma 2: *Suppose $\lambda \in (0,1)$, $T \subset \mathbb{R}^{d\times d}$ is the set of symmetric matrices with all eigenvalues in the range $\left(\frac{-\gamma}{2\lambda}, \frac{\gamma}{2(1-\lambda)}\right)$, and $S_1, S_2 \in \mathbb{R}^{d\times d}$ are positive definite matrices. Then there is a unique $V^\star \in T$ such that $\mathcal{F}(V^\star) = V^\star$, where*

$$\mathcal{F}(V) := S_2 - S_1 + S_1 \left(S_1 + I\gamma/2 - V(1-\lambda)\right)^{-1} S_1 - S_2 \left(S_2 + I\gamma/2 + V\lambda\right)^{-1} S_2.$$

Also, for any $V \in T$, $\lim_{n\to\infty} \mathcal{F}^n(V) = V^\star$.

Proof: Suppose $A, B \in T$ and $A > B$. First we will establish that $\mathcal{F}(\cdot)$ is order-preserving, which is equivalent to $\mathcal{F}(A) > \mathcal{F}(B)$. Note that,

$$S_1 \left(\left(S_1 + I\gamma/2 - A(1-\lambda)\right)^{-1} - \left(S_1 + I\gamma/2 - B(1-\lambda)\right)^{-1}\right) S_1 > 0 \iff$$

$$\left(S_1 + I\gamma/2 - A(1-\lambda)\right)^{-1} > \left(S_1 + I\gamma/2 - B(1-\lambda)\right)^{-1} \iff$$

$$-A < -B \iff A > B.$$

Similar logic implies that for all such $A, B \in T$,

$$S_2 \left((S_2 + I\gamma/2 + B\lambda)^{-1} - (S_2 + I\gamma/2 + A\lambda)^{-1} \right) S_2 > 0 \iff A > B,$$

and since $\mathcal{F}(A) - \mathcal{F}(B)$ is the sum of both of these order-preserving functions, $\mathcal{F}(\cdot)$ is also order-preserving.

Clearly our bounds on the eigenvalues imply that $\mathcal{F}(V)$ is continuous for all $V \in T$. To show that $\mathcal{F}$ is a mapping from T into T, note that matrix symmetry is preserved over addition and inversion, so $\mathcal{F}(V)$ is symmetric for all $V \in T$. Also, note that,

$$\mathcal{F}(-I\gamma/(2\lambda)) = -S_1 + S_1 \left(S_1 + I\gamma/(2\lambda) \right)^{-1} S_1 > -I\gamma/(2\lambda) \iff$$

$$-S_1^{1/2} \left(I - \left(I + S_1^{-1}\gamma/(2\lambda) \right)^{-1} \right) S_1^{1/2} > -I\gamma/(2\lambda) \iff$$

$$S_1^{1/2} \left(I - \left(I + S_1^{-1}\gamma/(2\lambda) \right)^{-1} \right) S_1^{1/2} < I\gamma/(2\lambda) \iff$$

$$\left(I + S_1 2\lambda/\gamma \right)^{-1} < S_1^{-1}\gamma/(2\lambda) \iff I > 0.$$

Similar logic can be used to show that $\mathcal{F}(I\gamma/(2(1-\lambda)) < I\gamma/(2(1-\lambda)$. This also implies the final requirement of Lemma 1.

The only remaining requirement of Lemma 2 is the penultimate, which we will establish for $A, B \in T$ such that $A > B$, and using the norm, $d(A, B) = \mathrm{Tr}(A - B)$. Also, let $\alpha := \{1, -1\}$, $\beta := \{\lambda - 1, \lambda\}$, and $\|C\|$ denote the spectral norm of $C \in \mathbb{R}^{d \times d}$. We will use the property $\mathrm{Tr}(CD) \leq \|C\| \, \mathrm{Tr}(D)$, where $C, D \in \mathbb{R}^{d \times d}$ and $C, D > 0$; see for example, Ran and Reurings (2004). Note that,

$$\mathrm{Tr}(\mathcal{F}(A) - \mathcal{F}(B)) =$$

$$\textstyle\sum_i \alpha_i \mathrm{Tr}(S_i \left((S_i + I\gamma/2 + A\beta_i)^{-1} - (S_i + I\gamma/2 + B\beta_i)^{-1} \right) S_i) =$$

$$\textstyle\sum_i \alpha_i \beta_i \mathrm{Tr}(S_i \left(S_i + I\gamma/2 + A\beta_i \right)^{-1} (B - A) \left(S_i + I\gamma/2 + B\beta_i \right)^{-1} S_i) =$$

$$\textstyle\sum_i \alpha_i \beta_i \mathrm{Tr} \left((S_i + I\gamma/2 + B\beta_i)^{-1} S_i S_i \left(S_i + I\gamma/2 + A\beta_i \right)^{-1} (B - A) \right) \leq$$

$$\textstyle\sum_i \alpha_i \beta_i \left\| (S_i + I\gamma/2 + B\beta_i)^{-1} S_i S_i \left(S_i + I\gamma/2 + A\beta_i \right)^{-1} \right\| \mathrm{Tr} \left(B - A \right) <$$

$$c\mathrm{Tr} \left(B - A \right) \textstyle\sum_i \alpha_i \beta_i = c\mathrm{Tr} \left(A - B \right),$$

where $c \in (0, 1)$. The second inequality follows from the matrix $S_i \left(S_i + I\gamma/2 - A\beta_i \right)^{-1}$ (respectively, $S_i \left(S_i + I\gamma/2 - B\beta_i \right)^{-1}$) being similar to a symmetric matrix, and with eigenvalues contained in $(0, 1)$ because $A \in T$ ($B \in T$) implies $I\gamma/2 - A\beta_i > 0$ ($I\gamma/2 - B\beta_i > 0$).

□

Next we will establish Theorem 1, which is restated below. This is a slightly more general version of the theorem in the main text where the objective function in Eqn (10) is a weighted average of $\mathcal{W}_\gamma^2(p_1, q)$ and $\mathcal{W}_\gamma^2(p_2, q)$.

Theorem 1: *Let $\lambda \in (0,1)$ and p_1 and p_2 be Gaussian density functions with means $\mu_1, \mu_2 \in \mathbb{R}^d$, and variance matrices, $S_1, S_2 \in \mathbb{R}^{d\times d}$. The regularized Wasserstein barycenter between p_1 and p_2 is given by the density function of $N(\mu_B, S_B)$, where $\mu_B \in \mathbb{R}^d$ and $S_B \in \mathbb{R}^{d\times d}$ are defined by,*

$$\begin{aligned}\mu_B :=& \lambda\mu_1 + (1-\lambda)\mu_2 \\ S_B :=& \left(V2\lambda/\gamma + I\right)^{-1}\left(V\lambda + I\gamma/2 + S_2\right)\left(V2\lambda/\gamma + I\right)^{-1} \\ =& \left(V2(\lambda-1)/\gamma + I\right)^{-1}\left(V(\lambda-1) + I\gamma/2 + S_1\right)\left(V2(\lambda-1)/\gamma + I\right)^{-1},\end{aligned}$$

where $V \in \mathbb{R}^{d\times d}$ is the unique symmetric matrix that satisfies these equalities and $-I\gamma/(2\lambda) < V < I\gamma/(2(1-\lambda))$.

Also, the iterates of the following series converge to V when $V^{(0)} := \mathbf{0}_{d\times d}$,

$$V^{(k+1)} = S_2 - S_1 + S_1\left(S_1 + I\gamma/2 - V^{(k)}(1-\lambda)\right)^{-1}S_1 - S_2\left(S_2 + I\gamma/2 + V^{(k)}\lambda\right)^{-1}S_2.$$

Proof: Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be defined as, $\phi(z) := \exp(-\left\|z\right\|_2^2/\gamma)$, and, for a given function $f : \mathbb{R}^d \to \mathbb{R}$, we will denote the convolution of $f(z)$ and $\phi(z)$ as, $f(z) \circledast \phi(z) := \int_{\mathbb{R}^d} f(t)\phi(z-t)\mathrm{d}t$. When there is little risk of confusion, we will omit the input $z \in \mathbb{R}^d$ of functions supported on $\mathbb{R}^d$ in the remainder of the proof.

We will characterize the barycenter using the fact that it is the minimizer of the following optimization problem.

$$\min_q \lambda\mathcal{W}_\gamma^2(q, p_1) + (1-\lambda)\mathcal{W}_\gamma^2(q, p_2). \tag{A.1}$$

To do so, note that $\mathcal{W}_\gamma^2(q, p_i)$ can be defined by instead solving the dual of (8), which is,

$$\mathcal{W}_\gamma^2(q, p_i) = \max_{w_i, u_i} E_{p_i}(\log(w_i)) + E_q(\log(u_i)) - \gamma\int_{\mathbb{R}^d\times\mathbb{R}^d} w_i(z_1)u_i(z_2)\exp(-\left\|z_1 - z_2\right\|^2/\gamma)\mathrm{d}z_1\mathrm{d}z_2, \tag{A.2}$$

and the optimal coupling can be defined in terms of the dual variables as, $\varphi_i(z_1, z_2) = u_i(z_1)\phi(z_1)\phi(z_2)w_i(z_2)$. The first order conditions of (A.2) are,

$$p_i = w_i \left(u_i \circledast \phi\right) \tag{A.3}$$

$$q = u_i \left(w_i \circledast \phi\right). \tag{A.4}$$

Also, since the objective function of (A.2) is differentiable, an application of the envelope theorem implies,

$$\tfrac{\delta \mathcal{W}_\gamma^2(q,p_i)}{\delta q} = \log(u_i).$$

Thus, the optimum of (A.1) can be characterized by the following functional derivative being zero.

$$\frac{\delta}{\delta q}\left(\lambda \mathcal{W}_\gamma^2(q, p_1) + (1-\lambda)\mathcal{W}_\gamma^2(q, p_2)\right) = 0 \implies$$
$$\lambda \log(u_1) + (1-\lambda)\log(u_2) = 0$$

After combining this equality with (A.3-A.4), we have that the barycenter can be characterized by the system,

$$p_1 = w_1\left(u_1 \circledast \phi_{\gamma/2}\right),\ p_2 = w_2\left(u_2 \circledast \phi_{\gamma/2}\right)$$

$$q = u_1\left(w_1 \circledast \phi_{\gamma/2}\right) = u_2\left(w_2 \circledast \phi_{\gamma/2}\right), \text{ and } 1 = u_1^{\lambda} u_2^{1-\lambda}.$$

This system can be reduced to two equalities after noting that, $p_i = w_i\left(u_i \circledast \phi_{\gamma/2}\right)$ and $q = u_i\left(w_i \circledast \phi_{\gamma/2}\right)$ implies

$$q = u_i\left(\tfrac{p_i}{u_i \circledast \phi_{\gamma/2}} \circledast \phi_{\gamma/2}\right).$$

After combining both equalities, and noting $u_1 = u_2^{(\lambda-1)/\lambda}$, we have

$$q = u_2^{(\lambda-1)/\lambda}\left(\frac{p_1}{u_2^{(\lambda-1)/\lambda} \circledast \phi_{\gamma/2}} \circledast \phi_{\gamma/2}\right) = u_2\left(\frac{p_2}{u_2 \circledast \phi_{\gamma/2}} \circledast \phi_{\gamma/2}\right) \tag{A.5}$$

Let $\mathcal{G}$ be defined as the set of functions $g : \mathbb{R}^d \to \mathbb{R}_+^1$ of the form

$$g(z) = a \exp(-(z-\mu_g)^\top V_g^{-1}(z-\mu_g)/2),$$

where $\mu_g \in \mathbb{R}^d$, $V_g \in \mathbb{R}^{d\times d}$ is a symmetric and invertible matrix, and $a \in \mathbb{R}^1_{++}$. It will also be convenient to let $\mathcal{C} : \mathcal{G} \to \mathbb{R}^{d\times d}$ be defined so that $\mathcal{C}(g) = V_g$ and $\mathcal{M} : \mathcal{G} \to \mathbb{R}^d$ be defined so that $\mathcal{M}(g) = \mu_g$. It is well known that if $g, h \in \mathcal{G}$ are Gaussian density functions, then $g^b, cg, g \circledast h, gh \in \mathcal{G}$, where $b, c \in \mathbb{R}^1$ and $b \neq 0$, and it is also straightforward to show

$$\mathcal{C}(g^b) = V_g/b,\ \mathcal{C}(cg) = V_g,$$

$$\mathcal{C}(gh) = \left(V_g^{-1} + V_h^{-1}\right)^{-1}, \text{ and } \mathcal{C}(g \circledast h) = V_g + V_h.$$

Likewise, in the case of $\mathcal{M}(\cdot)$, we will also use the properties

$$\mathcal{M}(g^b) = \mu_g,\ \mathcal{M}(cg) = \mu_g,\ \mathcal{M}(gh) = \mathcal{C}(gh)\left(V_g^{-1}\mu_g + V_h^{-1}\mu_h\right), \text{ and } \mathcal{M}(g \circledast h) = \mu_g + \mu_h.$$

Note that $V_g^{-1} + V_h^{-1} > 0$ is the necessary and sufficient condition for $g \circledast h$ to be well defined, and it is straightforward to verify that the properties above also hold over all pairs of $g, h \in \mathcal{G}$ when this is the case; for the case of normal density functions, see for example, (Bromiley, 2003).

Next, we will suppose that u_2 is in $\mathcal{G}$, which, due to (A.5), also implies $q, u_1, w_1, w_2 \in \mathcal{G}$, and then show that there exists a unique $u_2 \in \mathcal{G}$ that satisfies (A.5). Since (A.1) is a strictly convex optimization problem, when a solution to (A.1) exists, it can be characterized uniquely by its first-order conditions.[3] Thus, after providing $u_2 \in \mathcal{G}$ that solves (A.5), we will have also shown that this solution is unique even when not restricted to $\mathcal{G}$.

Since ϕ, p_1, and p_2 are elements of $\mathcal{G}$, and $\mathcal{G}$ is closed under multiplication, division, convolution, and exponentiation to the (non-zero) power of $(\lambda - 1)/\lambda$, if $u_2 \in \mathcal{G}$ then the functions on both sides of the equality (A.5) will also be elements of $\mathcal{G}$. Let $U_i := \mathcal{C}(u_i)$ and $\mu_u := \mathcal{M}(u_2)$. As noted above, the convolutions in (A.5) are only well defined if the following matrix inequalities hold, so we will also require the solution to satisfy these inequalities.[4]

$$I2/\gamma + U_i^{-1} > 0 \text{ and } I2/\gamma + U_i^{-1}(\lambda - 1)/\lambda > 0,$$

[3]Note that, for any pair u_i, w_i that solves (A.1), we have that $u_i a, w_i/a$, where $a \in \mathbb{R}^1_{++}$, are also solutions. We avoid complications from this issue by placing the additional restriction on these dual variables that $w_i(0) = 1$, as this ensures strict convexity over this set of dual functions. To see that this is also without loss of generality, note that rescaling the dual variables by $u_i a, w_i/a$ would not impact the objective function in (A.2) because $\int_{\mathbb{R}^d} q(z)\mathrm{d}z = \int_{\mathbb{R}^d} p_i(z)\mathrm{d}z = 1$. Also, a would not impact the first order conditions (A.3-A.4), so it would also not have an impact on q.

[4]It is straightforward to verify that these inequalities are identical to the ones that ensure the optimal coupling is integrable, as this coupling is given by, $\varphi_i(z_1, z_2) = u_i(z_1)\phi(z_1)\phi(z_2)w_i(z_2)$. Thus, Fubini's theorem implies that they are also sufficient conditions for q to be integrable.

which hold if and only if,

$$-2/\gamma I < U_2^{-1} < 2\lambda/(\gamma(1-\lambda))I. \tag{A.6}$$

We can find S_B by applying $\mathcal{C}(\cdot)$ to (A.5), which implies,

$$\begin{aligned} S_B^{-1} =& U_2^{-1} + \left(\left(S_2^{-1} - (U_2 + I\gamma/2)^{-1}\right)^{-1} + I\gamma/2\right)^{-1} && \text{(A.7)} \\ =& U_2^{-1}(\lambda-1)/\lambda + \left(\left(S_1^{-1} - (U_2\lambda/(\lambda-1) + I\gamma/2)^{-1}\right)^{-1} + I\gamma/2\right)^{-1}. && \text{(A.8)} \end{aligned}$$

Let $b_i \in \{\lambda/(\lambda-1), 1\}$. After three applications of the matrix inversion lemma and simplifying we have that, for each $i \in \{1, 2\}$,

$$\begin{aligned} S_B^{-1} - U_2^{-1}/b_i =& \left(\left(S_i^{-1} - (U_2 b_i + I\gamma/2)^{-1}\right)^{-1} + I\gamma/2\right)^{-1} \\ =& \left(\left(S_i^{-1} - I2/\gamma + 4/\gamma^2\left(U_2^{-1}/b_i + I2/\gamma\right)^{-1}\right)^{-1} + I\gamma/2\right)^{-1} \\ =& I2/\gamma - 4/\gamma^2\left(S_i^{-1} + 4/\gamma^2\left(U_2^{-1}/b_i + I2/\gamma\right)^{-1}\right)^{-1} \\ =& I2/\gamma - 4/\gamma^2 S_i + 4/\gamma^2 S_i\left(\gamma^2/4U_2^{-1}/b_i + I\gamma/2 + S_i\right)^{-1} S_i. \end{aligned} \tag{A.9}$$

This, along with equations (A.7-A.8), implies that U_2 can be characterized by,

$$\begin{aligned} \gamma^2/4U_2^{-1} - S_2 + S_2\left(\gamma^2/4U_2^{-1} + I\gamma/2 + S_2\right)^{-1} S_2 = \\ \gamma^2/4U_2^{-1}(\lambda-1)/\lambda - S_1 + S_1\left(\gamma^2/4U_2^{-1}(\lambda-1)/\lambda + I\gamma/2 + S_1\right)^{-1} S_1. \end{aligned}$$

After defining V as $\gamma^2/(4\lambda)U_2^{-1}$, this implies,

$$V = S_2 - S_1 + S_1\left(S_1 + I\gamma/2 - V(1-\lambda)\right)^{-1} S_1 - S_2\left(S_2 + I\gamma/2 + V\lambda\right)^{-1} S_2.$$

Note that our requirement that U_2^{-1} satisfy (A.6) can be written in terms of V as, $-\gamma/(2\lambda)I < V < \gamma/(2(1-\lambda))I$, and Lemma 2 implies that there is a unique solution that satisfies these conditions.

The functional form for S_B from the statement of this theorem follows from an alternative

ordering of the matrix inversion theorem. Specifically, starting from (A.9),

$$\begin{aligned}
S_B^{-1} - U_2^{-1}/b_i &= I2/\gamma - 4/\gamma^2 \left(S_i^{-1} + 4/\gamma^2 \left(U_2^{-1}/b_i + I2/\gamma \right)^{-1} \right)^{-1} \\
&= -U_2^{-1}/b_i + 4/\gamma^2 \left(\gamma^2/4U_2^{-1}/b_i + I\gamma/2 \right) \left(\gamma^2/4U_2^{-1}/b_i + I\gamma/2 + S_i \right)^{-1} \\
&\times \left(\gamma^2/4U_2^{-1}/b_i + I\gamma/2 \right) \\
&= -U_2^{-1}/b_i + \left(2\lambda/(\gamma b_i)V + I \right) \left(\lambda/b_i V + \gamma/2I + S_i \right)^{-1} \left(2\lambda/(\gamma b_i)V + I \right).
\end{aligned}$$

Thus,

$$\begin{aligned}
S_B^{-1} &= \left(V2\lambda/\gamma + I \right)^{-1} \left(S_2 + V\lambda + I\gamma/2 \right) \left(V2\lambda/\gamma + I \right)^{-1} \\
&= \left(V2(\lambda-1)/\gamma + I \right)^{-1} \left(S_1 + V(\lambda-1) + I\gamma/2 \right) \left(V2(\lambda-1)/\gamma + I \right)^{-1}
\end{aligned}$$

After applying $\mathcal{M}(\cdot)$ to both sides of (A.5), we have

$$\mathcal{M}\left(u_2^{b_i} \left(\frac{p_i}{u_2^{b_i} \circledast \phi_{\gamma/2}} \circledast \phi_{\gamma/2} \right) \right) =$$

$$S_B \left(U_2^{-1}\mu_u/b_i + \left(S_B^{-1} - U_2^{-1}/b_i \right) \left(S_i^{-1} - \left(U_2 b_i + I\gamma/2 \right)^{-1} \right)^{-1} \left(S_i^{-1}\mu_i - \left(U_2 b_i + I\gamma/2 \right)^{-1} \mu_u \right) \right). \tag{A.10}$$

To simplify this expression, we will first establish three intermediate equalities. First, equations (A.7-A.8) imply

$$\begin{aligned}
S_B^{-1} - U_2^{-1}/b_i = \left(\left(S_i^{-1} - \left(U_2 b_i + I\gamma/2 \right)^{-1} \right)^{-1} + I\gamma/2 \right)^{-1} \implies \\
\left(S_B^{-1} - U_2^{-1}/b_i \right) \left(S_i^{-1} - \left(U_2 b_i + I\gamma/2 \right)^{-1} \right)^{-1} \left(U_2 b_i + I\gamma/2 \right)^{-1} \\
= \left(U_2 b_i + \gamma/2 \left(U_2 b_i + I\gamma/2 \right) S_i^{-1} \right)^{-1} \\
= \left(I + \gamma/2 \left(I + \gamma/(2b_i) U_2^{-1} \right) S_i^{-1} \right)^{-1} U_2^{-1}/b_i.
\end{aligned} \tag{A.11}$$

Second, (A.11) in turn implies

$$\begin{aligned}
&\left(S_B^{-1} - U_2^{-1}/b_i \right) \left(S_i^{-1} - \left(U_2 b_i + I\gamma/2 \right)^{-1} \right)^{-1} S_i^{-1} \\
&= \left(I + \gamma/2 \left(I + \gamma/(2b_i) U_2^{-1} \right) S_i^{-1} \right)^{-1} \left(I + \gamma/(2b_i) U_2^{-1} \right) S_i^{-1}.
\end{aligned} \tag{A.12}$$

Third, after an application of the matrix inverse identity to (A.7-A.8),

$$S_B^{-1} = U_2^{-1}/b_i + \left(\left(S_i^{-1} - (U_2 b_i + I\gamma/2)^{-1}\right)^{-1} + I\gamma/2\right)^{-1} \tag{A.13}$$

$$= U_2^{-1}/b_i + I2/\gamma - I4/\gamma^2 \left(I2/\gamma + S_i^{-1} - (U_2 b_i + I\gamma/2)^{-1}\right)^{-1}, \tag{A.14}$$

which implies

$$\begin{aligned} S_B &= \left(U_2^{-1}/b_i + I2/\gamma - 4/\gamma^2 \left((U_2 b_i + I\gamma/2)\left(I2/\gamma + S_i^{-1}\right) - I\right)^{-1} (U_2 b_i + I\gamma/2)\right)^{-1} \\ &= \left(U_2^{-1}/b_i + I2/\gamma - 4/\gamma^2 \left(I2/\gamma + (I + U_2^{-1}\gamma/(2b_i))S_2^{-1}\right)^{-1} U_2^{-1}/b_i \, (U_2 b_i + I\gamma/2)\right)^{-1}. \end{aligned}$$

Thus,

$$S_B = \left(U_2^{-1}/b_i + I2/\gamma\right)^{-1} \left(I - \left(I + \gamma/2\left(I + \gamma/(2b_i)U_2^{-1}\right) S_i^{-1}\right)^{-1}\right)^{-1}. \tag{A.15}$$

We will start with the coefficient on μ_u in (A.10). The equalities (A.11) and (A.15) imply that this term is equal to

$$\begin{aligned} & S_B \left(U_2^{-1}/b_i - \left(S_B^{-1} - U_2^{-1}/b_i\right)\left(S_i^{-1} - (U_2 b_i + I\gamma/2)^{-1}\right)^{-1} (U_2 b_i + I\gamma/2)^{-1}\right) \mu_u \\ &= \left(U_2^{-1}/b_i + I2/\gamma\right)^{-1} \left(I - \left(I + \gamma/2\left(I + \gamma/(2b_i)U_2^{-1}\right) S_i^{-1}\right)^{-1}\right)^{-1} \\ &\quad \times \left(I - \left(I + \gamma/2\left(I + \gamma/(2b_i)U_2^{-1}\right) S_i^{-1}\right)^{-1}\right) U_2^{-1}/b_i \mu_u \\ &= \left(U_2^{-1}/b_i + I2/\gamma\right)^{-1} U_2^{-1}/b_i \mu_u \\ &= \left(I + U_2 2b_i/\gamma\right)^{-1} \mu_u. \end{aligned}$$

The equalities (A.12) and (A.15) imply that the coefficient on μ_i in (A.10) can be written as,

$$\begin{aligned} & S_B \left(S_B^{-1} - U_2^{-1}/b_i\right)\left(S_i^{-1} - (U_2 b_i + I\gamma/2)^{-1}\right)^{-1} S_i^{-1} \mu_i \\ &= \left(U_2^{-1}/b_i + I2/\gamma\right)^{-1} \left(I - \left(I + S_i^{-1}\gamma/2 + U_2^{-1} S_i^{-1}/b_i\gamma^2/4\right)^{-1}\right)^{-1} \\ &\quad \times \left(I + \gamma/2\left(I + \gamma/(2b_i)U_2^{-1}\right) S_i^{-1}\right)^{-1} \left(I + \gamma/(2b_i)U_2^{-1}\right) S_i^{-1} \mu_i \\ &= \left(U_2^{-1}/b_i + I2/\gamma\right)^{-1} \left(\gamma/2\left(I + \gamma/(2b_i)U_2^{-1}\right) S_i^{-1}\right)^{-1} \left(I + \gamma/(2b_i)U_2^{-1}\right) S_i^{-1} \mu_i \\ &= \left(U_2^{-1}\gamma/(2b_i) + I\right)^{-1} \mu_i. \end{aligned}$$

After combining these terms, we can define (A.10) as the solution to

$$\begin{aligned}
\mu_q = (I + U_2 2b_i/\gamma)^{-1} \mu_u + \left(U_2^{-1}\gamma/(2b_i) + I\right)^{-1} \mu_i \implies \\
(I + U_2 2b_i/\gamma) \left(\mu_q - \left(U_2^{-1}\gamma/(2b_i) + I\right)^{-1} \mu_i\right) = \mu_u \implies \\
(I + U_2 2b_1/\gamma) \left(\mu_q - \left(U_2^{-1}\gamma/(2b_1) + I\right)^{-1} \mu_1\right) = (I + U_2 2/\gamma) \left(\mu_q - \left(U_2^{-1}\gamma/2 + I\right)^{-1} \mu_2\right).
\end{aligned}$$

Since the matrix inverse identity also implies,

$$\left(U_2^{-1}\gamma/(2b_i) + I\right)^{-1} = I - (U_2 2b_i/\gamma + I)^{-1},$$

we have,

$$\begin{aligned}
(I + U_2 2/\gamma)\, \mu_q - U_2 2/\gamma \mu_2 = (I + U_2 2b_1/\gamma)\, \mu_q - U_2 2b_1/\gamma \mu_1 \implies \\
(1 - b_1)\mu_q = \mu_2 - b_1 \mu_1 \implies \\
(1 + \lambda/(1 - \lambda))\mu_q = \mu_2 + \lambda/(1 - \lambda)\mu_1 \implies \\
\mu_q = \mu_2(1 - \lambda) + \lambda \mu_1.
\end{aligned}$$

□

References

Agueh, M. and G. Carlier (2011): "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, 43, 904–924.

Benamou, J., G. Carlier, M. Cuturi, L. Nenna, and G. Peyre (2015): "Iterative Bregman projections for regularized transportation problems," *SIAM Journal on Scientific Computing*, 37, 1111–1138.

Bollerslev, T. and J. M. Wooldridge (1992): "Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances," *Econometric reviews*, 11, 143–172.

Bromiley, P. (2003): "Products and convolutions of Gaussian probability density functions," *Tina-Vision Memo*, 3, 1.

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

Clemen, R. (1989): "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 5, 559–583.

Cuturi, M. (2013): "Sinkhorn distances: Lightspeed computation of optimal transport," 2292–2300.

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

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.

Galichon, A. (2018): *Optimal transport methods in economics*, Princeton University Press.

Genest, C. (1992): "Vincentization revisted," *The Annals of Statistics*, 20, 1137–1142.

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

Knott, M. and C. S. Smith (1994): "On a generalization of cyclic monotonicity and distances among random vectors," *Linear algebra and its applications*, 199, 363–371.

Lichtendahl, K. C., Y. Grushka-Cockayne, and R. Winkler (2013): "Is it better to average probabilities or quantiles," *Management Science*, 59, 1594–1611.

McCracken, M. and S. Ng (2020): "FRED-QD: A quarterly database for macroeconomic research," *Working Paper*.

Peyré, G. and M. Cuturi (2019): "Computational optimal transport: with applications to data science," *Foundations and Trends® in Machine Learning*, 11, 355–607.

Ran, A. C. and M. C. Reurings (2004): "A fixed point theorem in partially ordered sets and some applications to matrix equations," *Proceedings of the American Mathematical Society*, 1435–1443.

Ratcliff, R. (1979): "Group reaction time distributions and an analysis of distribution statistics," *Psychological Bulletin*, 86, 446–461.

Sinkhorn, R. (1967): "Diagonal equivalence to matrices with prescribed row and column sums," *The American Mathematical Monthly*, 74, 402–405.

Solomon, J., F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas (2015): "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Transactions on Graphics (TOG)*, 34, 66.

Timmermann, A. (2006): "Forecast combinations," in *Handbook of Economic Forecasting*, Elsevier, vol. 1, 135–196.

Villani, C. (2003): *Topics in optimal transportation*, vol. 58, American Mathematical Soc.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 1–25.