Measuring disagreement in probabilistic and density forecasts*

Ryan Cumings-Menon† Minchul Shin‡ Keith Sill§

Abstract

In this paper, we introduce and study a class of disagreement measures for probability distribution forecasts based on the Wasserstein metric. We describe a few advantageous properties of this measure of disagreement between forecasters. After describing alternatives to our proposal, we use examples to compare these measures to one another in closed form. We provide two empirical illustrations. The first application uses our measure to gauge disagreement among professional forecasters about output growth and inflation rate in the Euro zone. The second application employs our measure to gauge disagreement among *multivariate* predictive distributions generated by different forecasting methods.

Key Words: Wasserstein metric; optimal transport; dispersion; disagreement; Survey of Professional Forecasters; point forecasts; density forecasts; heterogeneity.

1. Introduction

Measures of disagreement between forecasters play a few important roles in economics and business. First, they can serve as estimates of economic uncertainty (Zarnowitz and Lambros, 1987; Lahiri et al., 1988; Bomberger, 1996; Giordani and Söderlind, 2003; Rich and Tracy, 2004; Liu and Lahiri, 2004; Boero et al., 2008; Lahiri and Sheng, 2010b; Bruine de Bruin et al., 2011; Boero et al., 2014; Abel et al., 2016; Glas, 2020). Second, they are also used to understand the behavior of forecasters (Mankiw et al., 2003; Lahiri and Sheng, 2008; Patton and Timmermann, 2010; Lahiri and Sheng, 2010a; Coibion and Gorodnichenko, 2012; Clements, 2014; Andrade et al., 2016). While there has been a great deal of work in economics on measures of discrepancy between the point predictions of forecasters, there has been less work on measuring the discrepancy between probability and density predictions, with the exception of papers that consider the cross-sectional variation of higher moments (variance, range, skewness) of predictive distributions produced by the professional forecasters (D'Amico and Orphanides, 2008; Bruine de Bruin et al., 2011; Boero et al., 2008; Clements, 2014; Li and Tay, 2017) and of papers that measure disagreement among predictive distributions based on entropy or statistical divergence (Shoja and Soofi, 2017; Lahiri and Wang, 2019; Bajgiran et al., 2020; Rich and Tracy, 2020).

*__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.

§Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106. e-mail: keith.sill@phil.frb.org.

In this note, we use the Wasserstein metric, a distance metric between probability distributions, to motivate a class of measures for the dispersion between either probability densities or distributions.[1] Our proposed measure of disagreement is based on the Fréchet variance in the q–Wasserstein metric space. This can be viewed as a natural extension of cross-sectional variance of point forecasts, which is the Fréchet variance in Euclidean space, to probability/density forecasts.

After introducing our notation, the next section introduces additional notation, including the Wasserstein metric and our proposed measure of dispersion. Then, we study its properties in the context of measuring disagreement in probability and density predictions. Section 3 describes other possible robust measures of disagreement based on the Wasserstein metric that are potentially robust to the outlying predictive distribution. Section 4 explores alternative measures of dispersion that are also based on Fréchet variance, but use either distance metrics or divergence measures other than the q–Wasserstein metric, and compare our proposed measure to other measures of disagreement used in the economic forecasting literature. Two empirical applications are provided in Section 5.

2. Dispersion based on optimal transport

We will denote the set of probability distributions with support in $\mathbb{R}^d$ as $\mathcal{P}$. We will assume that in time period $t \in \mathbb{N}_+$ each agent $i \in \{1, \dots, N\}$ provides the probability distribution function forecast $P_{it} \in \mathcal{P}$ of the random variable y_{t+h}, where $h \in \mathbb{N}_+$. When this distribution is assumed to have a well defined density function, we will denote this function by $p_{it} : \mathbb{R}^d \to \mathbb{R}_+$. Also, when our discussion is limited to a particular time, we will omit the time subscript, t, on these functions.

Using this notation, the $q-$Wasserstein metric is defined as

$$W_q(P_i, P_j) = \left(\inf_{\varphi \in \Omega(P_i, P_j)} \int \|z_i - z_j\|_q^q \mathrm{d}\varphi(z_i, z_j) \right)^{1/q}, \tag{1}$$

where $\Omega(P_i, P_j)$ is the set of all couplings between the distributions P_i and P_j, which can also be defined more formally as

$$\Omega(P_i, P_j) = \left\{ \varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+ \mid \forall A \subset \mathbb{R}^d,\ \varphi(A, \mathbb{R}^d) = P_i(A) \text{ and } \varphi(\mathbb{R}^d, A) = P_j(A) \right\}. \tag{2}$$

The minimizer of (1) is known as the optimal transport plan because, for any $A, B \subset \mathbb{R}^d$, $\varphi(A, B)$ can be interpreted as the probability mass that is mapped, or transported, from A to B in order to minimize $E\left(\|z_i - z_j\|_q^q\right)$ where $z_i \sim P_i$ and $z_j \sim P_j$. This distance metric has the advantage of always being well defined for distributions with support in $\mathbb{R}^d$, including in cases in which these distributions are not absolutely continuous. For more detail on the field of optimal transport, see (Villani, 2003; Galichon, 2018; Panaretos and Zemel, 2019).

[1] It is worth to note that Rich and Tracy (2020) also consider a disagreement measure based on the Wasserstein metric. Our proposed measure is different from theirs, and we discuss their measure in Section 3.

Since $W_q(\cdot)$ is a distance metric, it is straightforward to define its corresponding Fréchet mean, which is also known as the $q-$Wasserstein barycenter. Specifically, the barycenter is defined as the minimizer of the following optimization problem,

$$V_q(\{P_i\}_{i=1}^N) := \min_{P \in \mathcal{P}} \frac{1}{N} \sum_{i=1}^{N} W_q(P_i, P)^q, \tag{3}$$

which we will denote by $\overline{P}$. The Wasserstein barycenter with $q \in 1, 2$ has previously been considered in the forecasting literature; Examples with $d = 1$ include Irpino and Verde (2006), Verde and Irpino (2007), Arroyo and Maté (2009), Arroyo et al. (2011), González-Rivera and Arroyo (2012), Lichtendahl et al. (2013), and Busetti (2017) while Cumings-Menon and Shin (2020) consider the case with $d \geq 1$. Our proposed measure of dispersion is $V_q(\{P_i\}_{i=1}^N)$, the value of the objective function at $\overline{P}$, which is also known as the Fréchet variance.

Note that $V_q(\cdot)$ can be defined in a particularly simple manner in the one dimensional case. Specifically, in this case we have,

$$V_q(\{P_i\}_{i=1}^N) = \min_{P \in \mathcal{P}} \frac{1}{N} \sum_{i=1}^{N} \int_0^1 \|P_i^{-1}(\tau) - P^{-1}(\tau)\|_q^q \mathrm{d}\tau, \tag{4}$$

where $P_i^{-1}(\cdot)$ and $P^{-1}(\cdot)$ are the quantile function of agent i and of the combination method, respectively. Section 2.2 also describes another case in which $V_q(\cdot)$ can be found in closed form, which is distributions with a location-scale parameterization. Outside of these two cases, $V_q(\cdot)$ can also be estimated by solving the convex problem (3) after discretizing.

2.1 Properties of $V_q(\{P_i\}_{i=1}^N)$

In this section we will outline four advantageous properties of $V_q(\cdot)$. The first two properties are trivial to prove, as they follow from the fact that $W_q(\cdot)$ is always well defined in the present setting, and is also a distance metric.

Property 1: Existence The measure of dispersion $V_q(\cdot)$ can be used to measure dispersion of a set of probability density functions, a set of probability mass functions, or a set containing both probability density functions and probability mass functions.

Property 2: Non-negativity For any set of input distributions $\{P_i\}_{i=1}^N$, we have $V_q(\{P_i\}_{i=1}^N) \geq 0$ and $V_q(\{P_i\}_{i=1}^N) = 0$ if and only if $P_i = P_j$ for all $i, j \in \{1, \dots N\}$.

Property 3: Monotonicity Let $\overline{P}_1$ and $\overline{P}_2$ denote the Wasserstein barycenters when the input densities are $\{P_i\}_{i=1}^N$ or $\{P_i\}_{i=1}^{N-1} \cup P_{N+1}$ respectively. If $W_q(\overline{P}_2, P_{N+1}) > W_q(\overline{P}_1, P_N)$, then $V_q(\{P_i\}_{i=1}^{N-1} \cup P_{N+1}) >$

$V_q(\{P_i\}_{i=1}^N)$.

Proof: Since $W_q(\overline{P}_2, P_{N+1}) > W_q(\overline{P}_1, P_N)$, we have

$$V_q(\{P_i\}_{i=1}^{N-1} \cup P_{N+1}) = \frac{1}{N}\left(W_q(P_{N+1}, \overline{P}_2)^q + \sum_{i=1}^{N-1} W_q(P_i, \overline{P}_2)^q\right) >$$
$$\frac{1}{N}\left(W_q(P_N, \overline{P}_1)^q + \sum_{i=1}^{N-1} W_q(P_i, \overline{P}_2)^q\right) \geq \frac{1}{N}\sum_{i=1}^{N} W_q(P_i, \overline{P}_1)^q = V_q(\{P_i\}_{i=1}^N).$$

□

In some sense Property 2 and Property 3 can be viewed as generalizations of properties that are shared by many common measures of dispersion for random variables. For example, the sample variance of $\{x_i\}_{i=1}^N$ is a positive measure of dispersion that would increase if any observation $x_i \in \{x_i\}_{i=1}^N$ were to be replaced by $\tilde{x}$ such that $\|x_i - \bar{x}\|_2 \leq \|\tilde{x} - \bar{x}\|_2$, where $\bar{x}$ denotes the sample mean.

Before moving onto the fourth property, it is also worth noting a more rigorous link between the 2–Wasserstein metric and cross-sectional variance of point forecasts. For example, given the sample of point forecasts $\{x_i\}_{i=1}^N$, we can define a corresponding set of distributions $\{P_i\}_{i=1}^N$, such that P_i is the distribution defined as a point mass at x_i. In this case, $V_2(\{P_i\}_{i=1}^N)$ is equal to the cross-sectional variance of $\{x_i\}_{i=1}^N$. In a similar way, but when $q = 1$, our proposed measure corresponds to the to average absolute error from median.

Property 4: Lower bound For any univariate input distributions $\{P_i\}_{i=1}^N$, the following inequality holds,

$$V_2(\{P_i\}_i) \geq \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \overline{\mu})^2 + \frac{1}{N}\sum_{i=1}^{N}(\sigma_i - \overline{\sigma})^2.$$

Proof: See Alvarez-Esteban et al. (2018). □

This inequality provides a useful lower bound based on the cross-sectional variance of first two moments of the input densities. Moreover, it is evident from this inequality that the proposed measure naturally extends a disagreement measure of point forecasts to probability and density forecasts by adding terms related to the cross-sectional variation of predictive distributions beyond the mean. The bound holds with equality when all input densities are Gaussian, as we shall see from the next section.

2.2 Example: Normal distributions

Agueh and Carlier (2011) provide a closed form solution for $V_2(\{P_{\mathcal{N},i}\}_{i=1}^N)$, where each $P_{\mathcal{N},i} \in \{P_{\mathcal{N},i}\}_i$ is a Gaussian distribution with mean $\mu_i \in \mathbb{R}^d$ and variance matrix $S_i \in \mathbb{R}^{d\times d}$. More generally, this solution for the 2–Wasserstein metric also holds for other distributions with a location-scale parameterization; see also (Knott and Smith, 1994; Panaretos and Zemel, 2019). Specifically, in these cases $V_2(\{P_{\mathcal{N},i}\}_i)$ is given

by

$$V_2(\{P_{\mathcal{N},i}\}_i) = \frac{1}{N}\sum_{i=1}^{N}\|\mu_i - \overline{\mu}\|_2^2 + \text{Trace}\left(S_i + S - 2\left(S_i^{1/2} S S_i^{1/2}\right)^{1/2}\right), \tag{5}$$

where S is defined by the fixed point of

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(S^{1/2} S_i S^{1/2}\right)^{1/2}. \tag{6}$$

When input densities are univariate Gaussian, this expression can be simplified further. For example, in this case we have $W_2(P_{i,\mathcal{N}}, P_{\mathcal{N}})^2 = (\mu_i - \mu)^2 + (\sigma_i - \sigma)^2$, so the Wasserstein barycenter can be defined as $\overline{P}_{\mathcal{N}} =_d N(\overline{\mu}, \overline{\sigma}^2)$, where $\overline{\mu} = \frac{1}{N}\sum_i \mu_i$ and $\overline{\sigma} = \frac{1}{N}\sum_i \sigma_i$. Thus, in this case, our proposed disagreement measure is equal to,

$$V_2(\{P_i\}_i) = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \overline{\mu})^2 + \frac{1}{N}\sum_{i=1}^{N}(\sigma_i - \overline{\sigma})^2,$$

or the sum of the cross-sectional variances of means and the cross-sectional variances of standard deviations.

As discussed previously, the literature on disagreement measures of forecasts focuses primarily on dispersion of point forecasts. In settings in which forecasters' distribution predictions are in the form of both a mean and a standard deviation of a normal distribution, this equality provides a link between the dispersion measure $V_2(\{P_i\}_i)$ and this prior work. Specifically, it is given by the cross-sectional variance of mean forecasts plus the cross-sectional variance of standard deviation forecasts.

3. Alternative dispersion measures I: Robust dispersion measure

One advantageous feature of $V_q(\cdot)$ that was not discussed in Section 2.1 is that this measure also encompasses robust dispersion measures, which, relative to our previous example, correspond to $q \in [1, 2)$. In this subsection we will describe alternative dispersion measures that also share this property. Although, these measures of dispersion are not monotonic and they may be zero even when some of the input distributions are not equal to one another.

MAD-type dispersion measure One can also generalize $V_q(\cdot)$ by considering nonlinear aggregations of the values in $\{W_q(P_i, \overline{P})^q\}_{i=1}^N$. For example, one possibility would be the following alternative measure of dispersion for $q = 1$.

$$D_1(\{P_i\}_{i=1}^N) = \text{median}_i\ \{W_1(P_i, \overline{P})\}_{i=1}^N$$

which, in terms of samples of random variables, is analogous to the median absolute deviation (MAD).

Dispersion without location measure Rousseeuw and Croux (1993) propose two related cross-sectional measures of dispersion for data points that do not depend on a notion of center. These can be viewed as

analogous to the interquartile range, which, like MAD, is robust to outliers, as these two measures have identical expectations for symmetrically distributed data, but, unlike MAD, the interquartile range is also robust to cases in which the variates are skewed.

The natural generalizations of the measures proposed by Rousseeuw and Croux (1993) to our setting are,

$$Q_q(\{P_i\}_{i=1}^N) = k^{\text{th}} \text{ order statistic of } \{W_q(P_i, P_j)\}_{i<j}, \tag{7}$$

where $k = (N \text{ choose } 2)/4$, and,

$$S_q(\{P_i\}_{i=1}^N) = \text{median}_{i\in\{1,2,\ldots,N\}} \text{ median}_{j\in\{1,2,\ldots,N\}}\{W_q(P_i, P_j)\}_{i,j=1}^N. \tag{8}$$

These dispersion measures have the advantage of being applicable for distributions of random variables that are discrete as well as continuous (Property 1). However, it is easy to construct cases in which Properties 2 and 3 do not hold. For example, for either of these two measures of dispersion, there exists a sufficiently large value of N such that $S_q(\{P_i\}_{i=1}^N) = Q_q(\{P_i\}_{i=1}^N) = 0$ when all elements of $\{P_i\}_{i=1}^{N-1}$ are identical, regardless of the value of P_N.

Recently, Rich and Tracy (2020) propose a measure of the *individual* disagreement (average absolute density disagreement) defined as

$$_iAADD_t = \frac{1}{N-1}\sum_{j\neq i} W_1(P_j, P_i),$$

to measure how the individual probability distribution P_i is different from others. In their descriptive analysis, they present and discuss time-series plot of $median_{i\in\{1,2,\ldots,N\}}(_iAADD_t)$, which is closely related to $S_1(\{P_i\}_{i=1}^N)$ introduced in Eqn (8): They coincide if we replace the second median operator in Eqn (8) with the sample average operator.

4. Alternative dispersion measures II: Dispersion measures using other metrics

Clearly one could use a similar approach as the one taken here using an alternative metric to the Wasserstein distance, such as total variation, Hellinger, L_2, Kullback-Leibler divergence, etc. For example, the Hellinger distance is one particularly popular fidelity criterion in statistical theory; see for example, (Beran, 1977; Kitamura et al., 2013). This distance metric is defined as,

$$H(p_1, p_2)^2 = \frac{1}{2}\int \left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2 \mathrm{d}x,$$

and its corresponding Fréchet variance can be used to measure disagreement among probability/density forecasts,

$$V_H(\{P_i\}_{i=1}^N) = \min_{P\in\mathcal{P}} \frac{1}{N}\sum_{i=1}^N H(p_i, p)^2.$$

A similar disagreement measure is possible for statistical divergences such as Kullback-Leibler divergence even though it is not a metric. For example, Shoja and Soofi (2017) and Lahiri and Wang (2019) employ the following averaged divergence to measure disagreement among probability distributions of professional forecasters

$$V_{KL}(\{P_i\}_i^N) = \frac{1}{N}\sum_{i=1}^{N} KL(P_i, P_*),$$

where $KL(P_i, P_*)$ is the Kullback-Leibler divergence between P_* to P_i and P_* is a consensus forecast.

Next we will provide two simple examples that illustrate the difference between our proposal and other measures of disagreement between distribution and density forecasts. While any choice of metric is inherently subjective, our reasoning behind choosing the Wasserstein metric is that it continues to provide meaningful information when the support of the input distributions do not overlap, in the sense that it is not simply equal to a constant in all such cases, which is demonstrated in the next example.

Example 1. Consider two uniform distributions, $P_1 = U(0,1)$ and $P_2 = U(x, x+1)$. In the case of the $q-$Wasserstein metric, the barycenter between these distributions is given by $U(x/2, x/2+1)$, and thus, $V_q(\{U(0,1), U(x,x+1)\}) = W_q(U(x/2, x/2+1), U(0,1)) = W_q(U(x/2, x/2+1), U(x,x+1)) = (x/2)^q$. In contrast, these alternative measures would not depend on x whenever $x > 1$. For example, $V_{KL}(\{P_i\}_i^N) = \log 2$.

We will call the corresponding property of the measure of dispersion *non-invariance to bijective transformations of the domain*. Admittedly, the desirability of this property is subjective; for example, Zanardo (2017) includes *invariance* to bijective transformations of the domain in a list of desiderata for a measure of dispersion of probability mass functions. However, this property has the advantage of not precluding the measure of dispersion being informative when the input distributions have supports that do not overlap. Next we provide an additional example of the effect of certain mappings of the input distributions' domains on the Wasserstein metric.

Example 2. Suppose $P_{1,\mathcal{N}} =_d N(\mu_1, \sigma_1^2)$ and $P_{2,\mathcal{N}} =_d N(\mu_2, \sigma_2^2)$, and let $P'_{1,\mathcal{N}}$ and $P'_{2,\mathcal{N}}$ be defined as $P_{1,\mathcal{N}}$ and $P_{2,\mathcal{N}}$ after rescaling the domain. Specifically, let $P'_{1,\mathcal{N}} =_d N(2\mu_1, 4\sigma_1^2)$ and $P'_{2,\mathcal{N}} =_d N(2\mu_2, 4\sigma_2^2)$. In this case we have $H(p_{1,\mathcal{N}}, p_{2,\mathcal{N}}) = H(p'_{1,\mathcal{N}}, p'_{2,\mathcal{N}})$ and $KL(p_{1,\mathcal{N}}, p_{2,\mathcal{N}}) = KL(p'_{1,\mathcal{N}}, p'_{2,\mathcal{N}})$ while the $2-$Wasserstein metric satisfies $W_2(P_{1,\mathcal{N}}, P_{2,\mathcal{N}}) = \frac{1}{2}W_2(P'_{1,\mathcal{N}}, P'_{2,\mathcal{N}})$.

This example can also be generalized in a straightforward manner. For example, suppose that $P_1(x), P_2(y), P'_1(x), P'_2(y) \in \mathcal{P}$ are defined so that $P'_1(x) = P_1(x/2)$ and $P'_2(x) = P_2(x/2)$. After rescaling the z_i, z_j in Eqn (1), we have $W_2(P_1, P_2) = \frac{1}{2}W_2(P'_1, P'_2)$.

Thus, the impact on disagreement measures based on the 2–Wasserstein metric from rescaling the domain is analogous to the impact on the cross-sectional variance from rescaling the datapoints, while other disagreement measures based on Hellinger distance or KL divergence would be scale invariant.

5. Empirical Applications

This section will provide two applications. The first application illustrates how our proposed measure of dispersion enhances analyses of SPF data that use traditional disagreement measures. The second application illustrates how our proposed measure can be used to gauge disagreement among *multivariate* input densities.

5.1 Application 1: Term structure of disagreement among professional forecasters

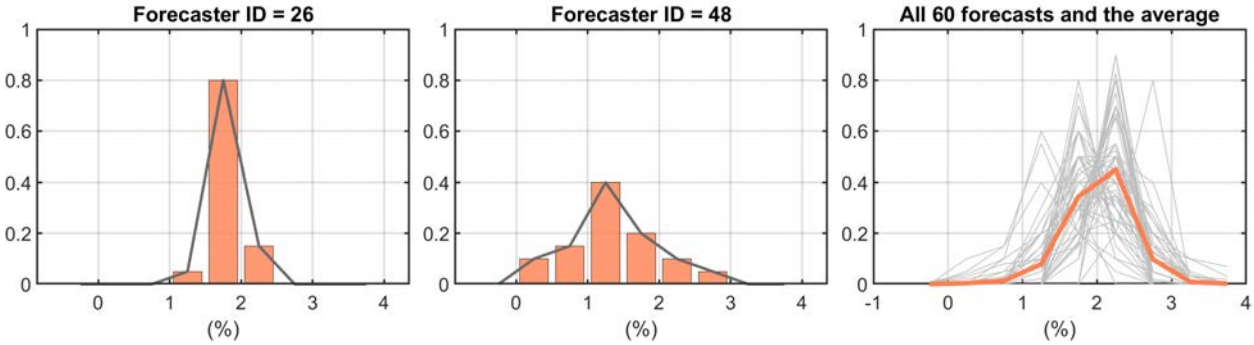
Researchers have reported that there is a large degree of disagreement among professional forecasters about various economic outcomes (See, for example, Sill, 2014). The disagreement among professional forecasters exhibit systematic patterns over time as well as over the forecasting horizon. The latter is sometimes referred to as the term structure of disagreement, and it has been a useful source of understanding the professional forecasters' behavior (see, for example, Lahiri and Sheng, 2008, 2010a; Patton and Timmermann, 2010; Clements, 2014; Andrade et al., 2016).

Data and methodology. In this section, we use our proposed measure of disagreement to quantify disagreement among forecasters across forecasting horizon using the Survey of Professional Forecasters conducted by the European Central Bank (ECB). This survey asks professionals about their point and probability forecasts about various economic outcomes. There are several questions in each survey in terms of forecasting target and forecasting horizons. In this application, we focus on two target variables: inflation rate and real GDP growth rate for Euro area. For each economic outcome, we consider three survey questions regarding forecasting horizons: (1) year-over-year growth rate at the end of the current calendar year; (2) year-over-year growth rate at the end of the next calendar year; (3) year-over-year growth rate at the end of five years ahead. We consider surveys conducted during 2001Q1–2019Q4 (76 quarters) and the average number of survey respondents was approximately 46–51 in each year.

It is important to note that depending on the timing of the survey, survey answers to the same question could imply a forecast with a different forecasting horizon. For example, the inflation rate estimate for the current calendar year is approximately a 3-quarter-ahead prediction if the survey was conducted in 2001Q1, while it is about 0-quarter-ahead prediction (nowcasting) if the survey was conducted in 2001Q4. Therefore, the forecasting horizon for the current calendar year estimate varies from $h = 0$ to $h = 3$. A similar logic applies to answers for the next calendar year and five years ahead, and the corresponding forecasting horizons range from $h = 4$ to $h = 7$ and from $h = 18$ to $h = 21$, respectively.

Professional forecasters indeed disagree with one another on mean, variance, and shape of the predictive distribution. In Figure 1 (left two panels), we present actual probability forecasts submitted by two forecasters from the survey conducted in the first quarter of 2001. The third panel of the same figure presents probability forecasts made by all 60 forecasters in the same survey. A thick line represents the average of those 60 forecasts.

Figure 1: Probability forecasts for inflation rates in 2001Q1 (one-year-ahead prediction)



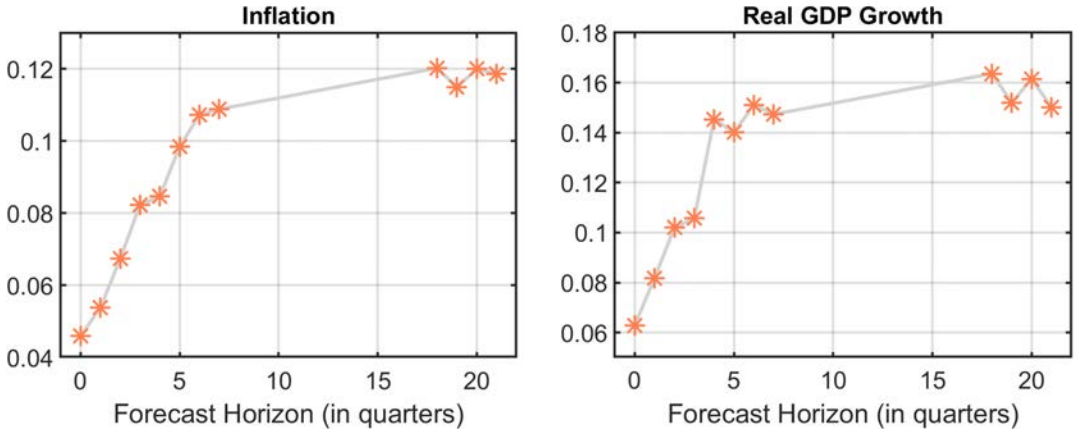
In what follows, we compute and present forecaster disagreement based on our proposed measure, $V_2(\{P_i\}_{i=1}^N)$, over the forecasting horizon. Although computation of Wasserstein metric and its related quantities such as barycenter and our disagreement measure can become complicated for general input densities, it is relatively simple to compute them when all input densities are in the form of a histogram (e.g., Arroyo and Maté, 2009). End bins of those histogram forecasts in the survey are open, and therefore they are unbounded. To facilitate computation, we assume that end bins are bounded and their length is the same as the other bins.

Results In the first row of Figure 2, we present our proposed disagreement measure, $V_2(\{P_i\}_{i=1}^N)$, over the forecasting horizons from $h = 0$ to $h = 21$ for inflation rate (left) and real GDP growth rate (right). Each asterisk represents the time-series average of the disagreements among forecasters' predictive distributions for the same target variable and the same forecasting horizon. The overall shape of the disagreement curves is quite similar for both target variables. Professional forecasters disagree more about the distant future. Interestingly, both disagreement curves become flatter as the forecasting horizon becomes longer.

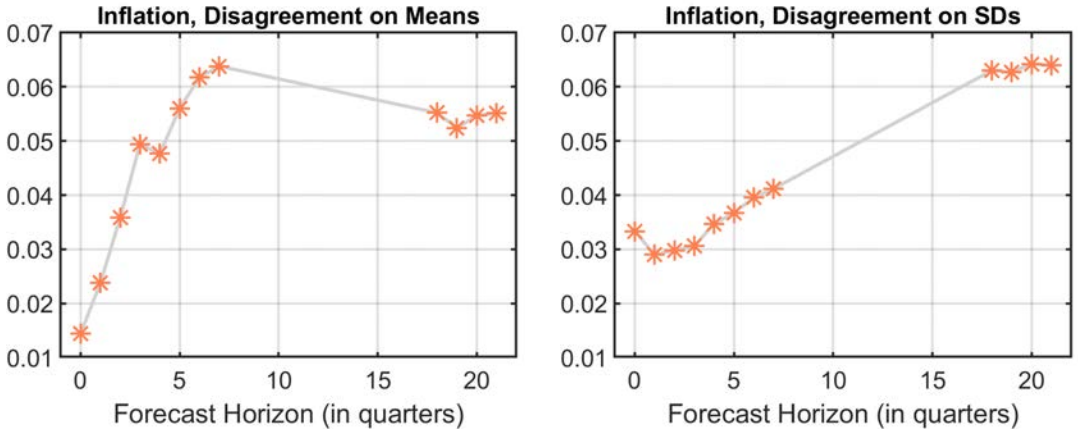
Recall that Property 4 in Section 2 shows that our disagreement measure can be decomposed into three non-negative terms. The first two terms are the cross-sectional variance of means and standard deviations of input densities. The third term then can be viewed as the disagreement about all remaining moments. In the middle row of Figure 2, we present cross-sectional variance of means (left) and standard deviations (right) of individual histogram forecasts for inflation rate. This decomposition reveals several important features. First, these two cross-sectional variances almost add up to $V_2(\{P_i\}_{i=1}^N)$. This implies that most of the variation in our proposed disagreement measure can be explained by the cross-sectional variance of the first two moments, and higher moments beyond mean and variance do not contribute much. Second, disagreement on means for inflation rate has an inverted U-shape relationship with forecasting horizon while disagreement on standard deviations is roughly an increasing function of forecast horizon from $h = 1$. Interestingly, professional forecasters agree about the long-run inflation forecast more than that for 5 or 6-quarter-ahead. This may be explained by the fact that the ECB governing Council goal of keeping the annual inflation rate below, but close to, 2% over the medium-term acts as a focal point for the longer-run

Figure 2: The term structure of disagreement among ECB-SPF forecasters

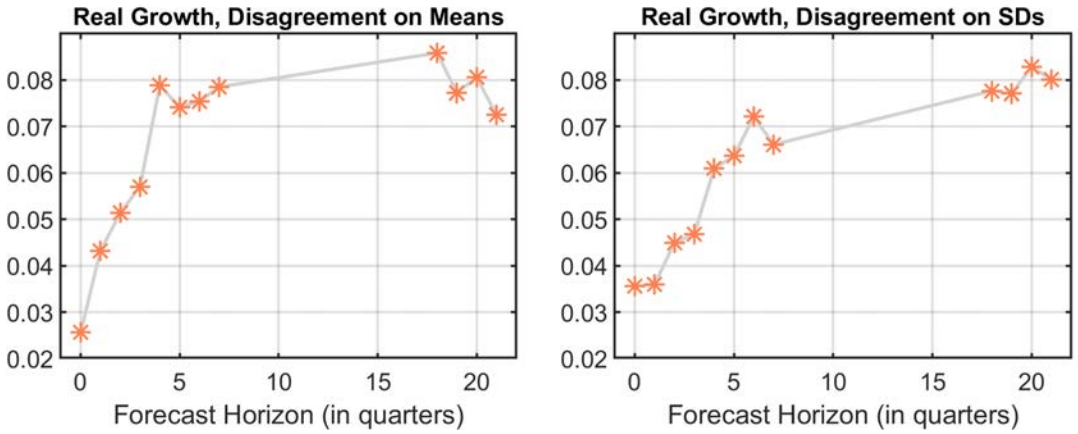
(a) Wasserstein metric based disagreement measure, $V_2(\{P_i\}_{i=1}^N)$



(b) Decomposition of $V_2(\{P_i\}_{i=1}^N)$ for inflation rate



(c) Decomposition of $V_2(\{P_i\}_{i=1}^N)$ for real GDP growth



inflation forecast. Unlike disagreement on means, forecasters exhibit a greater disagreement on standard deviations (i.e., uncertainty about their forecast) in the long run. This means that forecasters have a different view about how likely ECB governing Council's aim will be achieved.

In the last row of Figure 2, we present a similar decomposition for the real GDP growth rate. Again, our measure of disagreement can be mostly explained by the cross-sectional variance of first two moments. However, unlike the term structure of inflation rate forecast, both cross-sectional variance of means and standard deviations are an increasing function of forecasting horizon. This makes sense as the ECB governing Council explicitly aims to stabilize prices but not the real GDP growth rate. The disagreement among professional forecasters about how these price stabilization policies affect the real GDP growth rate may be the cause of the relatively high disagreement about the real GDP growth rate.

5.2 Empirical illustration 2: Multivariate density prediction

In this empirical illustration we show how one can use our dispersion measure for the multivariate predictive distribution. To this end, we consider the following 3-variable vector autoregression (VAR) that includes GDP growth rate, inflation rate, and federal funds rate for the US data.

We consider 21 hypothetical forecasters who produce their own 1-step-ahead joint predictive distributions for GDP growth rate and inflation rate at each point in time over the forecast evaluation sample. Their time t predictive distributions are then 2 dimensional multivariate normal distribution with mean $\mu_{i,t}$ and variance-covariance matrix $\Sigma_{i,t}$. We assume that forecasters estimate $\mu_{i,t}$ and $\Sigma_{i,t}$ using their own information set.

We further assume that information set differs only by the number of most recent observations when they construct a predictive distribution. For example, forecaster i estimates $\mu_{i,t}$ and $\Sigma_{i,t}$ using

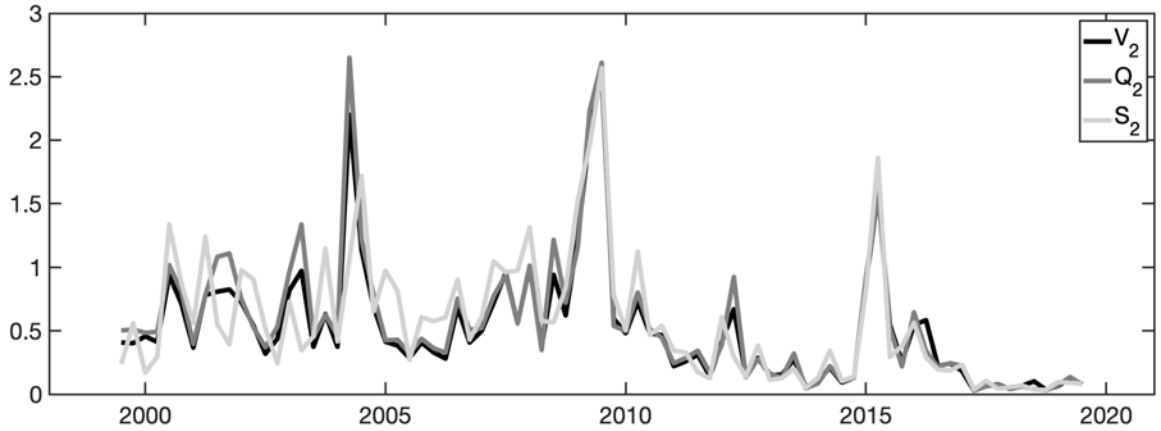
$$\Omega_{i,t-1} = \{Y_s : s = t - R_i, t - R_i + 1, ..., t - 1\}$$

where $\Omega_{i,t}$ is the information set of forecaster i at time t to produce a joint predictive density of Y_{t+1}. R_i is the number of most recent observations in the information set, and we consider 21 different choices $R_i = \{50, 55, ..., 150\}$ (hence, 21 different forecasters). Differential choice of R_i can be explained by either the theory of rational inattention or differential beliefs about the stability of the system.

Data and model We consider the variables: the GDP growth rate, inflation rate, and the federal funds rate. They are all annualized (in %). Our dataset begins in 1959Q1 and ends in 2019Q3. Forecasters generate predictions from 1998Q2 until 2019Q3. All forecasters use the same empirical model, a VAR with four lags.

$$Y_t' = \Phi_0 + \sum_{p=1}^{4} Y_{t-p}' \Phi_p + u_t'$$

Figure 3: Evolution of dispersion in joint predictive distributions over time



and their predictive distribution is $Y_{t+1|t} \sim \mathcal{N}(\mu_{i,t}, \Sigma_{i,t})$ where

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

and $(\widehat{\Phi}_{i,t,0}, \widehat{\Phi}_{i,t,1}, \widehat{\Phi}_{i,t,2}, \widehat{\Phi}_{i,t,3}, \widehat{\Phi}_{i,t,4}, \widehat{\Sigma}_{i,t})$ is the posterior mean of $p(\Phi_0, \Phi_1, \Phi_2, \Phi_3, \Phi_4, \Sigma | Y_{t-R_i:t-1})$ with a flat prior.

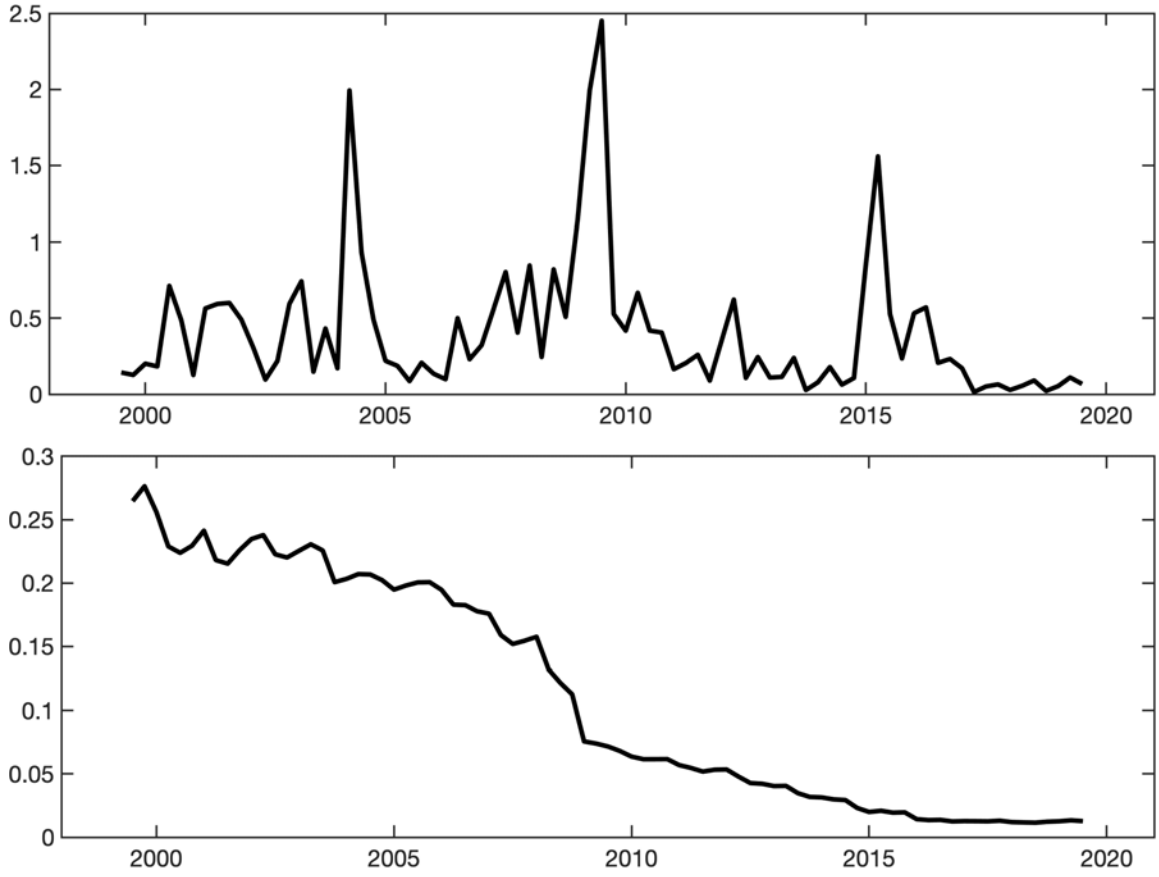
Results We generated individual predictive distributions for output growth rate and inflation rate and compute our disagreement measures starting from 1998Q2 through 2019Q3 (81 observations). We measured disagreement of these 21 forecasters using three measures D_2, S_2, and Q_2.

Figure 3 shows evolution of disagreement in joint predictive distributions over time. All three measures move quite closely to one another. Pairwise correlations of (V_2, Q_2), (V_2, S_2), and (Q_2, S_2) are 0.99, 0.86, and 0.80, respectively. Figure 4 presents decomposition of V_2. Specifically, since all predictive distributions are Gaussian, we decompose V_2 into a mean component and a variance component, as described in Section 2 in the case of the univariate Gaussian distribution. The upper figure presents the evolution of the mean component. The lower panel shows the evolution of the variance component.

Mean component can be viewed as a dispersion of mean of forecasters predictive distributions (i.e., point forecasts). It is serially correlated over time with the first autocorrelation value of approximately 0.5. There are three distinct peaks around 2004, 2009, and 2015. The observed time-variation in mean-disagreement can be explained by the fact that there is a difference in how the new information (shock) is weighted across different models with heterogenous memory capacity.

Interestingly, the mean component of our disagreement measure for GDP growth and inflation rate is positively correlated with Philadelphia Fed's forecast dispersion index (interquartile range of individual point forecasts) of GDP growth rate and inflation rate. Their correlation is about 0.45 and 0.49, respectively.

Figure 4: Decomposition of V_2: Mean (upper panel) and Variance (lower panel) component



This implies that the dispersion of professional forecasters can be partly explained by their memory capacity or concern about the structural break.

The variance component of our disagreement measure is downward trending over our sample. This is because of the Great Moderation effect. At the beginning of our estimation sample (1998Q2), there are forecasters using data observations from a time period with high volatility (i.e., pre-1985). However, the proportion of forecasters doing so decreases over time. In turn, the variance covariance matrix of forecasters' predictive distribution becomes similar to each other.

References

ABEL, J., R. RICH, J. SONG, AND J. TRACY (2016): "The Measurement and Behavior of Uncertainty: Evidence from the ECB Survey of Professional Forecasters," *Journal of Applied Econometrics*, 31, 533–550.

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

ALVAREZ-ESTEBAN, P. C., E. DEL BARRIO, J. A. CUESTA-ALBERTOS, C. MATRÁN, ET AL. (2018): "Wide consensus aggregation in the Wasserstein space. Application to location-scatter families," *Bernoulli*, 24, 3147–3179.

ANDRADE, P., R. K. CRUMP, S. EUSEPI, AND E. MOENCH (2016): "Fundamental Disagreement," *Journal of Monetary Economics*, 83, 106–128.

ARROYO, J., G. GONZÁLEZ-RIVERA, C. MATÉ, AND A. MUNOZ SAN ROQUE (2011): "Smoothing Methods for Histogram-Valued Time Series: An Application to Value-at-Risk," *Statistical Analysis and Data Mining*, 4, 216–228.

ARROYO, J. AND C. MATÉ (2009): "Forecasting histogram time series with k-nearest neighbours methods," *International Journal of Forecasting*, 25, 192–207.

BAJGIRAN, A. H., M. MARDIKORAEM, AND E. S. SOOFI (2020): "Maximum Entropy Distributions with Quantile Information," *European Journal of Operational Research*, In Press.

BERAN, R. (1977): "Minimum Hellinger distance estimates for parametric models," *The Annals of Statistics*, 5, 445–463.

BOERO, G., J. SMITH, AND K. F. WALLIS (2008): "Uncertainty and Disagreement in Economic Prediction: The Bank of England Survey of External Forecasters," *The Economic Journal*, 118, 1107–1127.

——— (2014): "The Measurement and Characteristics of Professional Forecasters' Uncertainty," *Journal of Applied Econometrics*, 30, 1029–1046.

BOMBERGER, W. (1996): "Disagreement as a Measure of Uncertainty," *Journal of Money, Credit and Banking*, 28, 381–392.

BRUINE DE BRUIN, W., C. F. MANSKI, G. TOPA, AND W. VAN DER KLAAUW (2011): "Measuring Consumer Uncertainty about Future Inflation," *Journal of Applied Econometrics*, 26, 454–478.

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

CLEMENTS, M. P. (2014): "Forecast Uncertainty – Ex Ante and Ex Post: U.S. Inflation and Output Growth," *Journal of Business & Economic Statistics*, 32, 206–216.

COIBION, O. AND Y. GORODNICHENKO (2012): "What Can Survey Forecasts Tell Us About Informational Rigidities?" *Journal of Political Economy*, 120, 116–159.

CUMINGS-MENON, R. AND M. SHIN (2020): "Probability Forecast Combination via Entropy Regularized Wasserstein Distance," *Entropy*, 22, 929.

D'AMICO, S. AND A. ORPHANIDES (2008): "Uncertainty and Disagreement in Economic Forecasting," .

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

GIORDANI, P. AND P. SÖDERLIND (2003): "Inflation forecast uncertainty," *European Economic Review*, 47, 1037–1059.

GLAS, A. (2020): "Five Dimensions of the Uncertainty-Disagreement Linkage," *International Journal of Forecasting*, 36, 607–627.

GONZÁLEZ-RIVERA, G. AND J. ARROYO (2012): "Time Series Modeling of Histogram-Valued Data: The Daily Histogram Time Series of S&P500 Intradaily Returns," *International Journal of Forecasting*, 28, 20–33.

IRPINO, A. AND R. VERDE (2006): "A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data," in *Data science and classification*, Springer, 185–192.

KITAMURA, Y., T. OTSU, AND K. EVDOKIMOV (2013): "Robustness, infinitesimal neighborhoods, and moment restrictions," *Econometrica*, 81, 1185–1201.

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.

LAHIRI, K. AND X. SHENG (2008): "Evolution of Forecast Disagreement in a Bayesian Learning Model," *Journal of Econometrics*, 144, 325–340.

——— (2010a): "Learning and Heterogeneity in GDP and Inflation Forecasts," *International Journal of Forecasting*, 26, 265–292.

——— (2010b): "Measuring forecast uncertainty by disagreement: The missing link," *Journal of Applied Econometrics*, 25, 514–538.

LAHIRI, K., C. TEIGLAND, AND M. ZAPOROWSKI (1988): "Interest Rates and the Subjective Probability Distribution of Inflation Forecasts," *Journal of Money, Credit, and Banking*, 20, 233–248.

LAHIRI, K. AND W. WANG (2019): "Estimating macroeconomic uncertainty and discord using info-metrics," .

LI, Y. AND A. S. TAY (2017): "The role of macroeconomic, policy, and forecaster uncertainty in forecast dispersion," .

LICHTENDAHL, K. C., Y. GRUSHKA-COCKAYNE, AND R. WINKLER (2013): "Is it better to average probabilities or quantiles," *Management Science*, 59, 1594–1611.

LIU, F. AND K. LAHIRI (2004): "Determinants of Multi-period Forecast Uncertainty Using a Panel of Density Forecasts," in *Econometric Society 2004 Australasian Meetings*, Econometric Society.

MANKIW, N. G., R. REIS, AND J. WOLFERS (2003): "Disagreement About Inflation Expectations," in *NBER Macroeconomics Annual 2003*, National Bureau of Economic Research, 209–270.

PANARETOS, V. M. AND Y. ZEMEL (2019): "Statistical aspects of Wasserstein distances," *Annual review of statistics and its application*, 6, 405–431.

PATTON, A. J. AND A. TIMMERMANN (2010): "Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion," *Journal of Monetary Economics*, 57, 803–820.

RICH, R. AND J. TRACY (2004): "Uncertainty and labor contract durations," *Review of Economics and Statistics*, 86, 270–287.

——— (2020): "A Closer Look at the Behavior of Uncertainty and Disagreement: Micro Evidence from the Euro Area," *Journal of Money, Credit and Banking*, In press.

ROUSSEEUW, P. J. AND C. CROUX (1993): "Alternatives to the median absolute deviation," *Journal of the American Statistical association*, 88, 1273–1283.

SHOJA, M. AND E. S. SOOFI (2017): "Uncertainty, information, and disagreement of economic forecasters," *Econometric Reviews*, 36, 796–817.

SILL, K. (2014): "Forecast Disagreement In the Survey of Professional Forecasters," *Business Review*, Q2, 15–24.

VERDE, R. AND A. IRPINO (2007): "Dynamic clustering of histogram data: using the right metric," in *Selected contributions in data analysis and classification*, Springer, 123–134.

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

ZANARDO, E. (2017): "How to Measure Disagreement," Ph.D. thesis, Columbia University.

ZARNOWITZ, V. AND L. A. LAMBROS (1987): "Consensus and uncertainty in economic prediction," *Journal of Political economy*, 95, 591–621.