JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **39**: 953–965 (2018)
Published online 9 September 2018 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12426

SPECIAL ISSUE

ON THE COMPARISON OF INTERVAL FORECASTS

ROSS ASKANAZI,[a] FRANCIS X. DIEBOLD,[b] FRANK SCHORFHEIDE[b] AND MINCHUL SHIN[c]

[a] *Cornerstone Research, Washington, DC, USA*
[b] *University of Pennsylvania, Philadelphia, PA, USA*
[c] *University of Illinois, Champaign, IL, USA*

We explore interval forecast comparison when the nominal confidence level is specified, but the quantiles on which intervals are based are not specified. It turns out that the problem is difficult, and perhaps unsolvable. We first consider a situation where intervals meet the Christoffersen conditions (in particular, where they are correctly calibrated), in which case the common prescription, which we rationalize and explore, is to prefer the interval of shortest length. We then allow for mis-calibrated intervals, in which case there is a calibration-length tradeoff. We propose two natural conditions that interval forecast loss functions should meet in such environments, and we show that a variety of popular approaches to interval forecast comparison fail them. Our negative results strengthen the case for abandoning interval forecasts in favor of density forecasts: Density forecasts not only provide richer information, but also can be readily compared using known proper scoring rules like the log predictive score, whereas interval forecasts cannot.

Received 12 January 2018; Accepted 02 August 2018

Keywords: Forecast accuracy; forecast evaluation; prediction.
JEL. C53.
MOS subject classification: 62M10; 62M20; 62F25.

1. INTRODUCTION

We consider the following situation: A researcher, R, is presented with two long sequences of univariate interval forecasts and the corresponding realizations, where each interval forecast has target (nominal) coverage $(1-\alpha)\times 100\%$. By 'long', we mean that we need not distinguish between consistent estimators and their probability limits, effectively working in population. Knowing nothing else, R must decide which interval forecast he prefers; that is, he must rank the two. Perhaps surprisingly, a full solution to this highly-practical and simply-stated problem remains elusive. In this article we address it.

We emphasize – and this is the source of difficulty – that R knows only that an interval's target coverage is $(1-\alpha)$, not the quantiles on which it is based. A target 90% interval, for example, *might* be equal-tailed (target probability 5% in each tail), *or it might not*, in which case there are uncountably infinitely many possibilities. R sees only the two sets of intervals, each with known target coverage $(1-\alpha)$, but with each based on unknown and potentially different quantiles.

Related foundational literature dates at least to Aitchison and Dunsmore (1968) and includes Winkler (1972) and Casella *et al.* (1993). Additional literature includes, among others, Granger *et al.* (1989) who note the tradeoff between an interval's expected length and its coverage, Giacomini and Komunjer (2005) who explore the relative assessment of conditional quantile estimates, and Gneiting and Raftery (2007) who propose proper scoring rules for interval predictions based on identical pairs of quantiles; see also Schervish (1996).

* Correspondence to: Francis X. Diebold, University of Pennsylvania, 133 South 36th Street, Philadelphia, PA 19104-6297, USA. E-mail: fdiebold@sas.upenn.edu

The comparison of interval forecasts is closely connected to the assessment of coverage intervals for unknown parameters. Optimality results for frequentist confidence intervals and Bayesian coverage intervals can be found in many statistics textbooks, for example, Lehmann (1986), Robert (1994), and Schervish (1996). An important difference between the theory of interval estimation and the comparison of interval forecasts is that the former typically restricts the comparison to interval estimates that satisfy a lower bound constraint on the coverage probability. In our forecasting context, on the other hand, there is no compelling reason to rule out forecasts that cover the actual observations with a frequency that is (slightly) smaller than the nominal coverage probability, in particular in settings in which R wants to make minimal assumptions about the data generating process (DGP). Thus, a key issue is the trade-off between interval length and coverage.

We proceed as follows. In Section 2 we treat the case where both interval forecasts meet the optimality conditions of Christoffersen (1998) ('the Christoffersen conditions', CC); that is, when the 0–1 hit sequences associated with the interval forecast sequences are iid Bernoulli with the correct hit probability. In that case the prescriptive optimum, which we characterize and explore in detail, is to prefer the interval of shortest length. In Section 3 we move to the richer case where one or both interval forecasts fail the CC due to mis-calibration. We consider a natural pair of conditions that interval forecast loss functions should meet, and we show that a variety of popular approaches to interval forecast comparison fail them. We conclude in Section 4.

2. THE FRAMEWORK, A PRESCRIPTIVE APPROACH, AND THE CC

2.1. The Basic Framework

Consider a univariate time series $\{y_t\}$ and let $y_{t+h|t}$ be an h-step-ahead point forecast of y_{t+h} made at time t. The corresponding h-step-ahead forecast error is $e_{t+h|t} = y_{t+h} - y_{t+h|t}$. Similarly, denote an h-step-ahead interval forecast ('prediction interval') by $d_{t+h|t}(1-\alpha) = [d^l_{t+h|t}(1-\alpha), d^u_{t+h|t}(1-\alpha)]$, where the '$1-\alpha$' denotes target $(1-\alpha)\times 100\%$ coverage, $0 \leq \alpha \leq 1$. Let $|d_{t+h|t}(1-\alpha)|$ be the length of the interval. The corresponding h-step 'hit sequence' is $1\{y_{t+h} \in d_{t+h|t}\}$, where $1\{\cdot\}$ is the indicator function. Finally, denote an h-step-ahead density forecast by $f_{t+h}(y|\Omega_t)$, where Ω_t is the information set on which the forecast is based. The corresponding probability integral transform (PIT) is $z_{t+h|t} = F_{t+h}(y_{t+h}|\Omega_t)$, where $F_{t+h}(\cdot|\Omega_t)$ is the c.d.f. associated with the density $f_{t+h}(\cdot|\Omega_t)$. We assume that the researcher R is presented with two forecasts and wants to rank them.

Necessary conditions for forecast optimality with respect to an information set exist and are widely-used for all of point, interval, and density forecasts. The basic idea is just the well-known 'orthogonality condition:' the relevant notion of 'forecast error' should be unforecastable using information available when the forecast was made. In the canonical one-step-ahead case this is commonly assessed as: (i) point forecast errors $e_{t+1|t}$ are independently and identically distributed (iid) with mean 0; (ii) interval forecast hits $1\{y_{t+h} \in d_{t+h|t}\}$ are iid with mean $1-\alpha$ as in Christoffersen (1998); and (iii) density forecast PIT's $z_{t+1|t}$ are iid $U(0,1)$ as in Diebold *et al.* (1998).[1]

However, quite apart from 'absolute evaluation,' that is, evaluation of whether a forecast satisfies conditions necessary for efficiency with respect to an information set, there is also the issue of 'relative evaluation' (comparison) – simply quantifying a forecast's performance and comparing it to that of competitors, regardless of whether any forecast under consideration is 'optimal' in any sense. Two forecasts, for example, may each be efficient (optimal), but with respect to different information sets, or both may simply be inefficient. Relative evaluation is typically done for point forecasts by invoking a loss function (e.g., quadratic loss) and examining the corresponding realized average loss (e.g., mean-squared error (MSE)). In precise parallel, for density forecasts one typically invokes a loss function, for example, the log predictive score (LPS)

$$LPS = \log\left(f_{t+h}(y|\Omega_t)\right),$$

[1] In the general h-step-ahead case we simply replace 'iid' with 'at most $(h-1)$-dependent'.

DOI: 10.1111/jtsa.12426

which is the log of the predictive density evaluated at the corresponding realization, and examines the corresponding realized average LPS, as in Amisano and Geweke (2017) and many of the references therein. However, it is much less clear what to do – and there is no 'typical' procedure – for interval forecasts, to which we now turn.

2.2. The Prescriptive Optimal Interval Forecast

We begin by providing a decision-theoretic framework to derive an optimal interval forecast.[2] Denote an interval forecast by $d = [d^l, d^u]$, with length $|d| = d^u - d^l$. That is, to reduce notational clutter, we drop the '$t + h|t$' and '$(1 - \alpha)$' notation, as meaning will be clear from the context. Consider a game between a forecaster F and an adversary A, where F chooses d and A chooses a scalar $\delta \in [-\infty, 0]$. Suppose that F's loss function is

$$\mathcal{L}_F(y, d, \delta; \alpha) = |d| + \delta\big(1\{y \in d\} - (1 - \alpha)\big). \tag{1}$$

This just says that, other things the same, F prefers short intervals (small $|d|$ promotes small $\mathcal{L}$) and intervals that contain the realization ($y \in d$ implies that $[1\{y \in d\} - (1-\alpha)]$ is positive, which promotes small $\mathcal{L}$). Note that F's loss function reflects a tradeoff: the wider the interval, the more likely it is to contain the truth, which decreases loss as per the second term in the loss function, but increases loss as per the first term in the loss function.

Suppose also that A's loss function is

$$\mathcal{L}_A(y, d, \delta; \alpha) = -\delta\big(1\{y \in d\} - (1 - \alpha)\big). \tag{2}$$

A's loss function is very simple: he incurs negative loss (i.e., positive utility) of $\delta(1 - \alpha)$ when F 'gets it wrong' (i.e., when the realization is outside the interval), and positive loss of $-\delta\alpha$ when F 'gets it right' (the realization is inside the interval). Hence A simply wants F to 'get it wrong', which makes explicit the sense in which A is an adversary.

Notice that d and δ appear in the loss functions of both F and A. However, F chooses d and treats δ as fixed (it is set externally by A), whereas A chooses δ and treats d as fixed (it is set externally by F). Finally, assume that both players know both loss functions and the associated strategic considerations, and that they use the same posterior to calculate posterior expected loss (risk). The equilibrium of the game is easily obtained by considering F's three choices:

1. Choose a correctly-calibrated interval, that is, an interval such that $\mathbb{P}(y \in d) = (1 - \alpha)$, where $\mathbb{P}$ denotes posterior probability. Risk to F associated with the second term of his loss function then vanishes, regardless of the adversary's action, so his risk is governed entirely by posterior expected interval length.
2. Choose an interval such that $\mathbb{P}(y \in d) > (1 - \alpha)$. That improves risk to F associated with the second term of his loss function if $\delta < 0$, so the adversary will choose $\delta = 0$, which F knows. Hence F has no incentive to choose, and will not choose, intervals with $\mathbb{P}(y \in d) > (1 - \alpha)$.
3. Choose an interval such that $\mathbb{P}(y \in d) < (1 - \alpha)$. That worsens risk to F associated with the second term of his loss function if $\delta < 0$, so the adversary will choose $\delta = -\infty$, which F knows. Hence F has no incentive to choose, and will not choose, intervals with $\mathbb{P}(y \in d) < (1 - \alpha)$.

Clearly, then, F always prefers the first option, a correctly-calibrated interval, in which case his expected loss (risk) collapses to[3]

$$\mathbb{E}\big[\mathcal{L}_F^*(y, d; \alpha)\big] = \mathbb{E}[|d|], \tag{3}$$

so he prefers the interval with the shortest length. Hence the decision-theoretic prescription optimizes the length-coverage tradeoff in a very special, lexicographic, way: restrict attention to correctly-calibrated intervals, and then pick the shortest (on average).

[2] The subsequent exposition builds on Herbst and Schorfheide (2015).
[3] The $*$ superscript indicates that the loss is conditional on the best response of the adversary A.

DOI: 10.1111/jtsa.12426

2.3. Characterization of the 'Shortest Well-Calibrated Interval'

It will prove useful to characterize the shortest well-calibrated interval with respect to the predictive distribution of y. This interval minimizes the expected length subject to a coverage constraint:

$$\min_{d^l, d^u} (d^u - d^l) \quad \text{s.t.} \quad \mathbb{P}(y \in d) = 1 - \alpha. \tag{4}$$

Because we restrict our attention to connected intervals, the constraint can be written as

$$\mathbb{P}(y < d^u) - \mathbb{P}(y < d^l) = 1 - \alpha. \tag{5}$$

The optimal interval satisfies the condition

$$f(d_o^l(1-\alpha)) = f(d_o^u(1-\alpha)), \tag{6}$$

where $f(\cdot)$ is the predictive density and the subscript 'o' denotes the optimal interval.[4] That is, the optimal interval equalizes the heights of the predictive density $f(y)$ at d_o^l and d_o^u. If the predictive density is unimodal, then the solution to the optimization problem yields the highest-probability-density interval – meaning that it includes all values y such that $f(y) \geq f(d_o^l(1-\alpha))$ – which is prominently featured in Bayesian analysis; see, for instance, the textbook of Robert (1994).

2.4. Interval Forecast Comparison Under the CC

Provided that two interval forecasts satisfy the CC either 'exactly' with a hit probability equal to $1-\alpha$ or 'weakly' with a hit probability that is no less than $1-\alpha$, the conventional prescription is to rank the intervals based on their average length, preferring shorter to longer, as intervals shorter in expectation presumably condition on more valuable information sets. This prescription is a consequence of the decision-theoretic framework presented in Section 2.2. We could simply replace the forecaster 'F' by the researcher 'R' to formally justify this prescription.

Unfortunately, adopting the framework of Section 2.2 has some rather undesirable implications and does not solve the interval-forecast-comparison problem in a satisfactory manner. First, forecasters receive no penalty for reporting a somewhat longer interval that exceeds the nominal coverage probability. Second, and more importantly, the penalty for undercoverage is unreasonably harsh. The risk score associated with an interval that does not satisfy the coverage probability constraint is $\mathbb{E}\left[\mathcal{L}_R^*(y, d; \alpha)\right] = -\infty$. Thus, any forecast that does not achieve the nominal coverage should be disregarded and the approach produces no useful ranking of two interval forecasts whose actual coverage probabilities fall short of the nominal coverage probability. In the remainder of this article we therefore explore alternative evaluation approaches.

3. COMPARING POTENTIALLY MIS-CALIBRATED INTERVAL FORECASTS

A major practical issue is what to do when we acknowledge that, in general, the CC *fail*.[5] The 'shortest well-calibrated interval' result provides a prescriptive recommendation for *constructing* interval forecasts, not a descriptive recommendation for *evaluating* interval forecasts, precisely because in general the CC fail (because one or both intervals may be mis-calibrated), in which case there is a calibration-length tradeoff. In this section we study aspects of that tradeoff.[6]

[4] We obtain (6) from the first-order conditions of the constrained optimization problem.
[5] We assume the *iid* part of the CC throughout; hence when we speak of failure of the CC we mean $\mathbb{E}[1\{y_{t+h} \in d_{t+h|t}\}] \neq 1-\alpha$.
[6] We assume throughout that all intervals are connected, and closely related, that the predictive distribution of y, $f(y)$, has a single mode.

DOI: 10.1111/jtsa.12426

3.1. Loss-Function Desiderata

We consider loss functions of the form $\mathcal{L}(d, y; \lambda)$, where $d = [d^l, d^u]$, y is a realization of Y with density function is $f(y)$ and distribution function $F(y)$, and λ is a finite-dimensional parameter vector that includes α.[7]

There are two key desirable traits of such interval-forecast loss functions in our environment, in which the researcher R knows nothing about the intervals apart from target coverage.

D1. *$\mathcal{L}$ should (i) reflect an inverse tradeoff between length and coverage; (ii) avoid the Casella paradox; and (iii) have corresponding risk minimized by the shortest well-calibrated interval.*

D2. *$\mathcal{L}$ should be evaluable with no knowledge of (i) the DGP, or (ii) the specific quantiles used by the forecasters to form the intervals.*

Several explanatory remarks related to D1 and D2 are in order. First, as regards D1(i), both short expected length and correct coverage are desirable. Hence a loss function's 'indifference curves' between average length and (absolute) coverage distortion should be negatively sloped. One should be willing, for example, to accept an increase in expected length in exchange for a 'large enough' improvement in coverage.

Second, as regards D1(ii), the Casella paradox refers to situations where the optimal interval is dominated in expected loss by arbitrarily short mis-calibrated intervals. It emerges when loss functions place 'too much' weight on length as opposed to coverage. In particular, loss functions that embed a linear length-coverage tradeoff typically fall victim to the Casella paradox. We provide additional discussion in Appendix A.

Third, as regards D1(iii), imagine that one of the forecasters is endowed with knowledge of the DGP and hence the 'true' conditional distribution, denoted by $f_0(y_{t+h}|\mathcal{F}_t)$. Moreover, suppose that this forecaster reports the shortest well-calibrated interval described in Section 2.3. Then, given the nominal coverage probability $1 - \alpha$, and despite the length-coverage trade-off encoded in the loss function, no other forecast should yield a lower risk.

Finally, as regards D2, at some level it is self-evident. *Of course* evaluation of $\mathcal{L}$ should not require knowledge of the DGP (no one knows the DGP), and *of course* it should not require knowledge of the specific left and right quantiles used (the forecasters are simply asked to provide $(1 - \alpha)$ % intervals). More formally, what we mean is that the loss function parameter λ should not depend on the probability measure $\mathbb{P}$ associated with the DGP, which is used to compute expected loss $\mathbb{E}[\mathcal{L}(d, y; \lambda)]$.

In what follows we will restrict attention to prominent loss functions that satisfy D1(i) and D1(ii). However, we will show that in general it is difficult to reconcile D1(iii) and D2.

3.2. Candidate Approaches

In view of D1 and D2, we now examine several interval-forecast evaluation approaches. We begin with a modification of the framework in Section 2.2 and then proceed to alternative loss/risk functions that have been proposed in the literature. A key problem for many of the loss functions is the following. Had the forecasters known that they will be evaluated based on such a loss function, they would not have reported the shortest $1 - \alpha$ coverage interval under their subjective beliefs about y_{t+h} conditional on their Ω_t information sets. That is, those 'scoring rules' are not 'proper'.

3.2.1. The Two-Player Game Revisited

The decision-theoretic framework outlined in Section 2.2 evidently fails D1(i), because it does not generate a (reasonable) trade-off between average length and coverage probability. However, a small modification of the setup can generate some improvements. Suppose we restrict the choice set of the adversary A to $\delta \in [-\underline{\delta}, 0]$, and define $\lambda = [\underline{\delta}, \alpha]'$, which does not depend on the DGP and therefore satisfies D2. As earlier, the forecaster does not receive a penalty for providing an interval with a coverage probability that exceeds $1 - \alpha$, but there is now a

[7] Alternatively, we could of course cast the discussion in terms of predictive score functions. The score function $S(d, y; \lambda)$ is simply $-\mathcal{L}(d, y; \lambda)$.

DOI: 10.1111/jtsa.12426

trade-off between length and coverage for intervals whose actual coverage is less than nominal coverage:

$$\mathbb{E}[\mathcal{L}_R^*(y,d;\lambda)] = \mathbb{E}[|d|] + \begin{cases} 0 & \text{if } \mathbb{E}[1\{y \in d\}] \geq 1-\alpha \\ \underline{\delta}\big((1-\alpha) - \mathbb{E}[1\{y \in d\}]\big) & \text{otherwise.} \end{cases} \tag{7}$$

As long as $\underline{\delta}$ is not too small, the Casella paradox can be avoided. Moreover, sufficiently large values of $\underline{\delta}$ may not distort the forecasters' incentive to truthfully reveal $1-\alpha$ coverage sets based on their predictive distributions. A practical difficulty is the choice of the parameter $\underline{\delta}$. To facilitate this choice in an *ex post* evaluation, it may be convenient to replace $|d|$ by $\log|d|$ and focus on percentage differences between average interval lengths across forecasts.[8] In this case, setting $\underline{\delta} = 1$ implies the following: compared to the shortest interval with nominal coverage probability, an interval that has the same length and undercovers by one percentage point receives the same penalty as an interval that attains the nominal coverage probability but is 1% longer than the baseline interval.

A key question is whether one can choose $\underline{\delta}$ sufficiently small to keep the tradeoff between expected length and coverage probability interesting and simultaneously satisfy D1(iii) and D2. Knowledge about some features of the DGP is required to ensure that reducing the length of the optimal $1-\alpha$ interval under the DGP does not lead to a reduction in the expected loss.

Suppose, for example, that under the DGP the predictive distribution is $N(\mu_{t+h|t}, \sigma^2_{t+h|t})$. Then the log length of the predictive interval is $|d| = \ln 2\sigma_{t+h|t} + \ln\left(\Phi_N^{-1}(1-\alpha/2)\right)$, where $\Phi_N^{-1}(\cdot)$ is the inverse cumulative density function of a standard normal random variable. Suppose that $\alpha = 0.05$. Increasing the critical value from 0.05 to 0.06 (by 1 percentage point) reduces the log length of the interval by $\ln(1.8808/1.96) \approx -0.04$, meaning the length shrinks by about 4%. This means that at $\alpha = 0.05$ if we set $\underline{\delta} \geq 5$, say, then intervals that undercover yield a larger loss.

3.2.2. Winkler Loss and its Relatives

The most commonly-used interval-forecast loss function is due to Winkler (1972); see also Schervish (1996):

$$\mathcal{L}_{eq}(y,d;\lambda) = |d| + \lambda_l(d^l - y)1\{y < d^l\} + \lambda_u(y - d^u)1\{y > d^u\}, \tag{8}$$

where $\lambda = [\lambda_l, \lambda_u]'$ with $\lambda_l > 0$ and $\lambda_u > 0$. The first term penalizes the length of the interval, and the second and third terms penalize misses from the left and right tails respectively. If the interval misses the realization from the left ($y < d^l$), the second term ($\lambda_l(d^l - y)$) becomes positive and the third term becomes zero. If the interval misses from the right ($y > d^u$), the second term becomes zero and the third term ($\lambda_u(y - d^u)$) becomes positive.

It can be shown Gneiting and Raftery (2007) that the optimal interval forecast under Winkler loss has left and right endpoints given by the $1/\lambda_l$ and $1-1/\lambda_u$ quantiles, so that for the evaluation of intervals with $(1-\alpha)$ coverage, one should set $1/\lambda_l + 1/\lambda_u = \alpha$. A popular choice for λ_l and λ_u in practice is $\lambda_l = \lambda_u = 2/\alpha$, in which case the optimal interval is equal-tailed Gneiting and Raftery (2007),

$$\mathcal{L}(d,y;\alpha) = |d| + \frac{2}{\alpha}\inf_{x \in d}|y - x|.$$

Notice, crucially, that *good interval forecasting under Winkler loss amounts to good quantile forecasting, for known, stated, quantiles*.

Some other well-known loss functions are similar to Winkler's. For example, Schlag and van der Weele (2015) discuss Schmalensee (1976)'s loss function

$$\mathcal{L}_S(y,d,\lambda_l,\lambda_u) = |d| + \frac{2}{\alpha}(d^l - y)1\{y < d^l\} + \frac{2}{\alpha}(y - d^u)1\{y > d^u\} + \frac{2}{\alpha}\left|y - \frac{d^u + d^l}{2}\right|. \tag{9}$$

[8] Of course, from an *ex ante* perspective, this would create an incentive to report intervals that have length zero.

DOI: 10.1111/jtsa.12426

The first three terms are the same as Winkler's. The difference is an additional penalty for the divergence between the realization and the midpoint of the interval. Nevertheless, good interval forecasting under this loss function amounts to good quantile forecasting.

The key insight is that all Winkler-type loss functions are quantile-type loss functions – they involve properties of estimates of known quantiles that define the left and right interval endpoints; see also Gneiting (2011b). Quantile-type loss functions appear reasonable if one knows (or is willing to assume) that the intervals to be compared are equal-tailed, or, more generally, are based on known left and right quantiles. However, that's not our case – we are simply provided with two sequences of intervals, and we must decide which we prefer.

Only in very special cases can quantile-type loss functions resolve the tension between D1(iii) and D2. Consider the following example based on Winkler loss. To generate a $1-\alpha$ interval forecast, it must be the case that $1/\lambda_l + 1/\lambda_u = \alpha$. Moreover, to satisfy D2, the parameters λ_l and λ_u cannot depend on features of the DGP. If the DGP is of the form $y_{t+h}|\mathcal{F}_t \sim N(\mu_{t+h|t}, \sigma^2_{t+h|t})$, then the optimal $1-\alpha$ interval forecast in any period corresponds to the $\alpha/2$ and $1-\alpha/2$ conditional quantiles, which also minimize expected Winkler loss for $\lambda_l = \lambda_u = 2/\alpha$. Thus, no further information about the DGP is required to satisfy D1(iii) and D2.

For asymmetric and time-varying distributions, however, it is generally not true that one can find λ's such that the optimal $1-\alpha$ interval forecast also minimizes Winkler loss. In Appendix B we provide a specific example in which a quantile-type loss function prefers a mis-calibrated interval of greater length to the shortest-length correctly calibrated interval, thereby violating D1(iii).

3.2.3. Length/Hit Loss

Loss functions that directly respect a length / hit tradeoff are of the form,

$$\mathcal{L}_G(d, y; \lambda) = M(|d|, 1\{y \in d\}; \lambda), \tag{10}$$

where $M(\cdot,\cdot)$ is a potentially nonlinear function parametrized by $\lambda \in \Lambda \subseteq \mathcal{R}^n$. The first argument is the interval length and the second argument is the hit indicator.[9] $M(\cdot,\cdot)$ is increasing in length, and $M(x, 1; \lambda) < M(x, 0; \lambda)$ for all x and $\lambda \in \Lambda$. Note that y enters M only through the hit indicator. We assume that M's risk minimizer is unique, and that $M(\cdot, 1; \lambda)$ and $M(\cdot, 0; \lambda)$ are smooth functions (twice differentiable).[10]

One prominent example proposed by Casella *et al.* (1993) is,

$$\mathcal{L}_{CHR}(d, y; \lambda) = \frac{|d|}{|d| + \lambda} - 1\{y \in d\} \text{ for } \lambda \in (0, \infty), \tag{11}$$

and another is the 'most likely interval' loss of Schlag and van der Weele (2015),

$$\mathcal{L}_{MLI}(d, y; \lambda) = -\left(1 - \frac{|d|}{b-a}\right)^{\lambda} 1\{y \in d\} \text{ for } \lambda \in (0, \infty), \tag{12}$$

where the support of y is restricted to a bounded interval, $[a, b]$.

The bounds of the risk-minimizing interval d_* satisfy the following for any $\lambda \in \Lambda$:

$$f(d^l_*(\lambda)) = f(d^u_*(\lambda)). \tag{13}$$

To see this, write $|d| = d^u - d^l$ and note that the expected loss is

$$\mathbb{E}[\mathcal{L}_G(d, y; \lambda)] = M(|d|, 1; \lambda)[F(d^u) - F(d^l)] + M(|d|, 0; \lambda)[1 - F(d^u) + F(d^l)], \tag{14}$$

[9] The function $\mathcal{L}_F(y, d, \delta; \alpha)$ in Section 2.2 is a function of the length and the hit indicator, but it also depends on the decision δ of the adversary.
[10] Casella *et al.* (1993) provide a set of conditions for M that avoid Casella's paradox.

DOI: 10.1111/jtsa.12426

so that the first order conditions with respect to d^l and d^u are

$$\begin{aligned} \partial d^l \;:\; & -M'(|d|,1;\lambda)[F(d^u)-F(d^l)] - M(|d|,1;\lambda)f(d^l) \\ & - M'(|d|,0;\lambda)[1-F(d^u)+F(d^l)] + M(|d|,0;\lambda)f(d^l) = 0 \\ \partial d^u \;:\; & M'(|d|,1;\lambda)[F(d^u)-F(d^l)] + M(|d|,1;\lambda)f(d^u) \\ & + M'(|d|,0;\lambda)[1-F(d^u)+F(d^l)] - M(|d|,0;\lambda)f(d^u) = 0. \end{aligned} \tag{15}$$

After rearranging terms, we find that $f(d^l) = f(d^u)$. This implies that the optimal interval prediction under loss function (10) satisfies one of the requirements for the shortest well-calibrated interval, as stated in equation (6).

Unfortunately, however, the coverage rate of this interval depends on λ. That is, λ needs to be set in a way that generates $(1-\alpha)$ coverage, which requires knowledge of the true predictive distribution (i.e., the DGP, which is specified as a conditional distribution). In time-series settings, moreover, the shape of the predictive distribution can change over time, which implies that the λ that induces $(1-\alpha)$ coverage can also vary over time. In any event, it is clear that D1(iii) and D2 cannot be satisfied simultaneously.

3.2.4. Direct Risk-Based Comparisons

In practice, researchers and practitioners often separately report and examine interval forecasts' average length and empirical coverage, as in Granger *et al.* (1989). In this way, people can presumably rank interval forecasts according to their own preferences (i.e., relative importance of length vs. calibration).

An interesting question is therefore whether there exists a loss function whose expected loss (risk) (i) can be written as a function of expected length and coverage, and (ii) is minimized by the shortest well-calibrated interval. Requirement (i) allows us to compute the expected loss easily by just replacing the expected length and coverage rate by their sample counterparts.

This motivates the direct study of risk functions, where risk depends only on expected length ($L = \mathbb{E}[|d|]$) and the difference between the target coverage rate and the actual coverage rate ($C = (1-\alpha) - \mathbb{E}[1\{y \in d\}]$, the 'coverage distortion'). The risk function (7) that resulted from the two player game takes this particular form. More generally, consider

$$\mathbb{E}[\mathcal{L}_{G'}(d,y;\lambda)] = M(L,C;\lambda), \tag{16}$$

where the expectation is taken with respect to the true predictive distribution, the function M is smooth (piecewise differentiable) for $\lambda \in \Lambda \subseteq \mathcal{R}^n$, and

$$\frac{\partial M(L,C;\lambda)}{\partial L} > 0 \quad \text{and} \quad \frac{\partial M(L,C;\lambda)}{\partial C} > 0, \tag{17}$$

whenever the derivatives exist.[11] An immediate example is the linear loss function,

$$\mathbb{E}[\mathcal{L}_{G'}(d,y;\lambda)] = L + \lambda C, \tag{18}$$

where $\lambda \in (0,\infty)$. Due to the linearity, this risk function corresponds to a loss function of the form (10).

Direct measures (16) have some good properties. In particular, under our earlier assumptions on $M(\cdot,\cdot)$, for any given λ the risk minimizer satisfies:

$$f(d^l_*(\lambda)) = f(d^u_*(\lambda)), \tag{19}$$

[11] As discussed in Section 3.2.1, in principle we would also need to impose restrictions on $M(\cdot,\cdot)$ sufficient to ensure avoidance of Casella's paradox, but we do not pursue that here.

DOI: 10.1111/jtsa.12426

which follows trivially from the first order conditions,

$$\begin{aligned} \partial d^l &: -\frac{\partial M(L,C)}{\partial L} + \frac{\partial M(L,C)}{\partial C} f(d^l) = 0 \\ \partial d^u &: \frac{\partial M(L,C)}{\partial L} - \frac{\partial M(L,C)}{\partial C} f(d^u) = 0, \end{aligned} \tag{20}$$

rearrangement of which yields $f(d^l) = f(d^u)$. This implies that the optimal interval prediction under risk function (16) satisfies one of requirements for the shortest well-calibrated interval (equation (6)).

Unfortunately, however, direct risk measures (16) also have the earlier-discussed bad properties that plagued other approaches. In particular, the coverage rate of the optimal predictive interval depends on λ. That is, for this risk functions of the form (16) to be proper for the shortest connected interval with $(1-\alpha)$ coverage, we need to select λ accordingly, but λ depends on the DGP, which is unknown. Of course more can be said if the DGP is known. For example consider again the linear loss function (16) with $\lambda \in (0, \infty)$, in which case $1/\lambda = f(d^l) = f(d^u)$. If f is a standard normal density, then $\lambda = 1/f(q_{\alpha/2})$ uniquely produces the shortest connected interval with $(1-\alpha)$ coverage, where $q_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution.

3.2.5. On Medians, Means, and Modes

Note that the use of equal-tailed interval forecasts naturally parallels the use of conditional-median point forecasts. To see this, note that the equal-tailed interval with ϵ coverage probability approaches the conditional median as $\epsilon \to 0$. The associated absolute-error loss function is well studied in the literature, which facilitates finding good scoring rules for equal-tailed intervals.

As an aside we note that, interestingly, little attention has been given to interval forecasts that parallel conditional-*mean* point forecasts. We are aware of just one article, Rice *et al.* (2008), which considers the loss function

$$\mathcal{L}(d, y; \lambda) = \lambda \frac{|d|}{2} + \frac{\left(y - |d^u + d^l|/2\right)^2}{|d|/2}, \tag{21}$$

which leads to the interval forecast $\left[m \pm s/\sqrt{\lambda}\right]$, where m and s are the mean and standard deviation of the predictive distribution of y. The coverage rate is implicitly defined through λ and the predictive distribution of y. As the coverage rate approaches zero, the interval approaches the mean of the predictive distribution.[12]

Now, finally, consider finding good scoring rules for the shortest connected interval. Shortest connected interval forecasts naturally parallel conditional-mode point forecasts, because the shortest connected interval with ϵ coverage probability approaches the conditional mode as $\epsilon \to 0$.[13] The associated 0-1 loss function, however, is mathematically more clumsy than absolute-error loss, which impedes finding good scoring rules for shortest connected intervals. Gneiting (2011a), for example, discusses a scoring rule for the mode, but he does not find a scoring rule that selects the mode as the global optimum.

3.3. Additional Remarks

First, the overarching point is that, to use any of the leading loss or risk functions discussed above for evaluating interval forecasts, in general one needs to know at least some features of the DGP. This creates a tension between D1(iii) and D2, which often can only be jointly satisfied for a limited class of DGP's.

Second, shortest connected intervals have drawbacks even beyond the difficulty of finding proper scoring rules. One example: The shortest interval may not be invariant to nonlinear transformations $g(\cdot)$; hence, for example

[12] Note that setting the target coverage rate requires knowledge of the predictive distribution of y.
[13] Here we are implicitly assuming unimodality of the predictive distribution, as the shortest connected interval is the high probability density region when the predictive interval is unimodal.

DOI: 10.1111/jtsa.12426

if $[a, b]$ is the shortest connected interval for y with support $y > 0$, then $[a^2, b^2]$ is not necessarily the shortest connected interval for y^2. Another example: If the predictive density is multi-modal, the shortest interval is not necessarily the shortest connected interval; rather, it can be a union of disjoint intervals.

Third, various methods have been proposed to construct the shortest predictive interval (the shortest connected interval for a class of unimodal distributions) based on parametric models Hyndman (1995, 1996), nonparametric models (Polonik and Yao, 2000), and semi-parametric models (Wu, 2012). However, quantile-based predictive intervals remain dominant, probably because proper scoring rules are readily available for quantile-based intervals. There seems to be willingness to accept some extra length in exchange for availability of a proper scoring rule.

Fourth, we note, however, that under some additional assumptions it may be possible to infer the predictive distribution from the interval forecast, which would let us move from interval forecast comparisons to the well-understood world of density forecast comparisons using well-known proper scoring rules. The conditions are of course strong, but perhaps not absurdly strong. Suppose that each interval prediction is a shortest connected interval (based on predictive distribution $\tilde{P}$ and density $\tilde{f}$, which do not necessarily match the true predictive distribution P and density f). Then the interval forecast $d = [d^l, d^u]$ contains the following information: (1) $\tilde{P}(y \in d) = 1 - \alpha$, and (2) $\tilde{f}(d^l) = \tilde{f}(d^u)$. If we knew $\tilde{P}$, then the log predictive score $\log \tilde{P}(y)$ could serve as a proper scoring rule, but we know only d, not $\tilde{P}$. However, if we know the true predictive distribution P then we can infer $\tilde{P}$ by solving

$$\min_q \int P \log\left(\frac{P}{\tilde{P}}\right) \mathrm{d}\mu \tag{22}$$

$$\text{s.t.} \quad \tilde{P}(y \in d) = 1 - \alpha \quad \text{and} \quad \tilde{f}(d^l) = \tilde{f}(d^u),$$

and then compute the log predictive score for the interval forecast by

$$\mathcal{L}_{KL}(y, d) = \log \tilde{P}(y; d). \tag{23}$$

This scoring rule is proper, so long as the true predictive distribution is unimodal and we consider a connected interval. To see this, suppose that a forecaster's predictive distribution is $\tilde{P} = P$. Then the solution to the minimization problem is any $\hat{\tilde{P}}$ such that $\hat{\tilde{P}} = P$ almost everywhere, which yields a value of 0 for the objective function (22). Achieving a value of 0 implies that the log predictive score is maximized with probability 1.[14]

Fifth, in general, making operational interval comparisons with desirable properties requires (some) knowledge of the 'true' predictive distribution, whereas it is in fact unknown. However, one could perhaps approximate it flexibly using, for example, ARMA conditional mean dynamics and GARCH conditional variance dynamics, together with a nonparametrically specified conditional density. This approximation could then be used to calibrate the loss and risk functions that are used to evaluate the interval forecasts.

4. CONCLUDING REMARKS

We have explored aspects of comparing two interval forecasts when the nominal confidence level is specified, but the quantiles on which intervals are based are not specified. The problem seems simple, and the lack of attention to it in the literature seems odd. It turns out that the problem is generally quite difficult, and perhaps unsolvable.

We first considered a situation where both intervals meet the CC (in particular, where both are correctly calibrated), in which case a common prescription is to prefer the interval of shortest length, and we explored and rationalized that prescription. We then allowed for mis-calibrated intervals, which cannot happen under the CC.

[14] This scoring rule will generally not be strictly proper, however, in that many predictive distributions can have the same shortest predictive interval as the true distribution and will thus achieve 0 loss.

DOI: 10.1111/jtsa.12426

We proposed two natural desiderata for interval forecast loss functions in that case, and we showed that a variety of popular approaches fail to meet them.

A positive suggestion that springs from our negative results is that interval forecasts might be beneficially abandoned, with focus instead placed on making and comparing complete density forecasts. This has already been happening in recent decades, because density forecasts are richer than interval forecasts in terms of the information conveyed, yet easily computed by simulation in both frequentist and Bayesian frameworks. Our results come from a different perspective yet add still more weight to the case for abandoning interval forecasts in favor of density forecasts: density forecasts can be readily compared using known proper scoring rules like the log predictive score, whereas interval forecasts cannot.

As for future research, it will be interesting to explore whether our results could be made less negative if we were to make stronger assumptions on the underlying distributions. Of course the answer is yes, which we have highlighted in various places, where we specialized to situations like Gaussian DGP's. However, those are basically just one-off examples, and it would be very interesting to attempt a more systematic characterization where the DGP is (say) a mixture of normals, determining what can be said depending on the mixture weights, number of components, etc.

It may also be interesting to explore whether our framework and results might extend to sets of interval forecasts of different series, as opposed to sets of interval forecasts of the same series, where the various series are potentially measured on different scales.[15] In calibration / length loss functions, for example, comparison of calibration is straightforward as it is dimensionless (measured in percent), but comparison of length is more challenging, as it is influenced by scale. Of course standardization may help, but the details remain to be explored.

ACKNOWLEDGEMENTS

For helpful discussion we thank the Editors (S. Leybourne and R. Taylor) and two referees, as well as Ulrich Mueller, Adrian Raftery, Glenn Rudebusch, and Jonathan Wright. For research support we thank the National Science Foundation and the Real-Time Data Research Center at the Federal Reserve Bank of Philadelphia. The usual disclaimer applies.

REFERENCES

Aitchison J, Dunsmore I. 1968. Linear-loss interval estimation of location and scale parameters. *Biometrika* **55**: 141–148.
Amisano G, Geweke J. 2017. Prediction using several macroeconomic models. *Review of Economics and Statistics* **99**: 912–925.
Casella G, Hwang J, Robert C. 1993. A paradox in decision-theoretic interval estimation. *Statistica Sinica* **3**: 141–155.
Christoffersen P. 1998. Evaluating interval forecasts. *International Economic Review* **39**: 841–62.
Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts, with applications to financial risk management. *International Economic Review* **39**: 863–883.
Giacomini R, Komunjer I. 2005. Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics* **23**: 416–431.
Gneiting T. 2011a. Making and evaluating point forecasts. *Journal of American Statistical Association* **106**: 746–762.
Gneiting T. 2011b. *Quantiles as optimal point forecasts*, Vol. 27.
Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**: 359–378.
Granger CWJ, Kamstra M, White H. 1989. Interval forecasting: an analysis based upon ARCH-quantile estimators. *Journal of Econometrics* **40**: 87–96.
Herbst E, Schorfheide F. 2015. *Bayesian Estimation of DSGE Models*. Princeton, NJ: Princeton University Press.
Hyndman RJ. 1995. Highest-density forecast regions for non-linear and non-normal time series models. *Journal of Forecasting* **14**: 431–441.
Hyndman RJ. 1996. Computing and graphing highest density regions. *The American Statistician* **50**: 120–126.
Hyndman RJ, Koehler AB. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**: 679–688.

[15] Such issues have been considered in depth for point forecasts (e.g., Hyndman and Koehler, 2006) but not interval forecasts.

DOI: 10.1111/jtsa.12426

Lehmann E. 1986. *Testing Statistical Hypotheses*. New York: Springer.
Polonik W, Yao Q. 2000. Conditional minimum volume predictive regions for stochastic processes. *Journal of the American Statistical Association* **95**: 509–519.
Rice KM, Lumley T, Szpiro AA. 2008. *Trading Bias for Precision: Decision Theory for Intervals and Sets. Biostatistics Working Paper*. Washington, DC: University of Washington.
Robert CP. 1994. *The Bayesian Choice*. New York: Springer-Verlag.
Schervish MJ. 1996. *Theory of Statistics*. New York: Springer.
Schlag KH, van der Weele JJ. 2015. A method to elicit beliefs as most likely intervals. *Judgment and Decision Making* **10**: 456–468.
Schmalensee R. 1976. An experimental study of expectation formation. *Econometrica* **44**: 17–41.
Winkler RL. 1972. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association* **67**: 187–191.
Wu JJ. 2012. Semiparametric forecast intervals. *Journal of Forecasting* **31**: 189–228.

APPENDIX A: CASELLA'S PARADOX

Let us illustrate the paradox by means of an example. Consider a loss function that trades off length and calibration,

$$\mathcal{L}(d, y; \lambda) = \lambda|d| - 1\{y \in d\},$$

where λ is a scale parameter chosen by a researcher R, and consider forecasting a Gaussian $y \sim N(0, \sigma^2)$. Suppose forecaster 1 issues a degenerate interval that is just the set $d_1 = \{0\}$, while forecaster 2 issues the correct equal-tailed interval $d_2 = [-z_{\alpha/2}\sigma, z_{\alpha/2}\sigma]$, with $z_{\alpha/2}$ denoting the α critical value. Then the expected losses are

$$\mathbb{E}[\mathcal{L}(d_1, y; \lambda)] = 0$$

$$\mathbb{E}[\mathcal{L}(d_2, y; \lambda)] = 2\lambda t_\alpha \sigma - (1 - \alpha).$$

Hence, to avoid selecting the degenerate interval forecast d_1, R must set $\lambda < \dfrac{1 - \alpha}{2z_{\alpha/2}\sigma}$.

That the parameter λ depends on the possibly unknown σ, and that for a fixed λ increasing σ will inevitably yield selection of the degenerate interval instead of an increasingly wide interval, is known as 'Casella's paradox'. The Casella *et al.* (1993) solution is to augment the loss function with a size function $S(\cdot)$, so that

$$\mathcal{L}\big(d, y; S(\cdot)\big) = S(|d|) - 1\{y \in d\}.$$

It is simple to derive conditions on $S(\cdot)$ that rule out the selection of *empty* intervals over those with correct coverage; however, parametric forms of $S(\cdot)$ that allow a researcher to consistently describe a desired tradeoff between coverage and length remain elusive.

APPENDIX B: AN EXAMPLE OF COMMON LOSS FUNCTIONS FAILING: EQUAL-TAILED INTERVALS AND ASYMMETRIC DISTRIBUTIONS

We provide an example in which a quantile-type loss function prefers a mis-calibrated interval of greater length to the shortest-length correctly calibrated interval. Consider the Gneiting and Raftery (2007) loss function,

$$\mathcal{L}(d, y; \alpha) = |d| + \frac{2}{\alpha} \inf_{x \in d} |y - x|.$$

Under this loss function, the optimal interval is equal tailed and given by the $\alpha/2$ and $1 - \alpha/2$ quantiles.

DOI: 10.1111/jtsa.12426

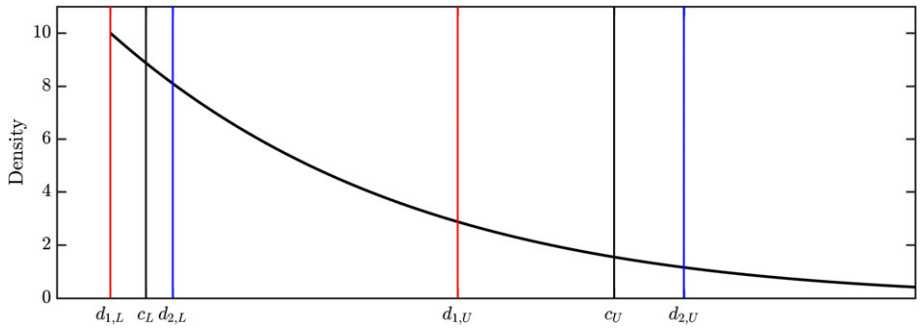

Figure B1. Motivating example. The predictive density is beta with parameters $a = 1$, $b = 10$. Interval $c = [c_1, c_2]$ is the equal-tailed interval with 75% coverage. Interval d_1 (red) is constructed as the shortest interval with 75% coverage. Interval d_2 (blue) targets the equal-tailed interval, but the lower endpoint is incorrectly shifted up by ϵ_L and the upper endpoint is shifted up by ϵ_U ($\epsilon_L = 0.010$ and $\epsilon_U = 0.026$). Interval d_2 will be chosen over d_1 by the Gneiting–Raftery loss function despite being longer and having incorrect coverage [Color figure can be viewed at wileyonlinelibrary.com]

Suppose that data is generated from a Beta distribution with shape parameters $a = 1$, $b = 10$. We will compare two intervals. The first interval, d_1, is constructed as the shortest interval with 75% coverage. The second interval, d_2, is constructed by shifting the lower and upper endpoints of the 75% equal-tailed interval to the right by some ϵ_1 and $\epsilon_2 \neq \epsilon_1$ respectively. The density of the beta distribution and the end points for the two intervals are displayed in Figure B1. Lengths and coverage probabilities are given by:

$$\begin{aligned} d_1 &: |d_1| = .13, \quad \mathbb{P}(y \in d_1) = .75 = 1 - \alpha \\ d_2 &: |d_2| = .19, \quad \mathbb{P}(y \in d_2) = .70 \neq 1 - \alpha. \end{aligned}$$

Let $p_B(y)$ denote the pdf of a Beta distribution. The expected Gneiting–Raftery loss for interval d_1 can be calculated as follows:

$$\mathbb{E}[\mathcal{L}(d_1, y; \alpha)] = |d_1| + \frac{2}{\alpha} \int_{d_{1,U}}^{1} (y - d_{1,U}) p_B(y) \mathrm{d}y \approx 0.29.$$

A similar calculation can be used to determine $\mathbb{E}[\mathcal{L}(d_2, y\ \alpha)]$:

$$\mathbb{E}[\mathcal{L}(d_2, y; \alpha)] = |d_2| + \frac{2}{\alpha} \int_{0}^{d_{2,L}} (d_{2,L} - y) p_B(y) \mathrm{d}y + \frac{2}{\alpha} \int_{d_{2,U}}^{1} (y - d_{2,U}) p_B(y) \mathrm{d}y \approx 0.26.$$

Hence in this example

$$\mathbb{E}[\mathcal{L}(d_2, y; \alpha)] < \mathbb{E}[\mathcal{L}(d_1, y; \alpha)].$$

That is, under the Gneiting–Raftery loss the longer interval d_2, with coverage probability less than $1-\alpha$, is preferred to the shorter interval with the desired coverage probability. The crux of the issue is that the forecasters and the researcher R do not have the same target. In using this loss function, R is effectively stating that he prefers intervals closer to the given quantiles over intervals with correct coverage that are far from these quantiles. Forecaster who place priority only on length and coverage will be routinely discarded, despite doing their best for their given objective.

DOI: 10.1111/jtsa.12426