Forecast Scoring and Calibration

Advanced Topics in Statistical Learning, Spring 2023

Ryan Tibshirani

1 Introduction

In this lecture, we cover a topic that sits adjacent to conformal prediction in a sense, which is calibration. We will cover the perspective(s) on calibration from the forecasting literature, where it is arguably the most developed. Instead of focusing solely on calibration, we'll cover some of the broader theory of forecasting, pertaining to scoring rules, calibration, and ensembles.

There is a rich literature on forecast scoring and calibration. This literature has roots in statistics, but over the years it has been expanded and driven by applied forecasting disciplines, primarily weather forecasting. It is still quite an active area in terms of development: new results that seem foundational in nature are still being discovered year to year.

As per our usual comment, what we cover in this lecture is only a sample of what is known in the theory of forecasting. Unlike other topics, we are not aware of a book that gives a modern and comprehensive treatment of forecast scoring rules, calibration, and so on. (Perhaps this would be premature, as these topics are still in active development, and such a book will come later.) One of the most active and influential researchers in forecasting theory today is Tilmann Gneiting, and much of what we cover will be drawing from his work with collaborators, such as Gneiting and Raftery (2007); Gneiting et al. (2007); Ranjan and Gneiting (2010); Gneiting and Ranjan (2013).

1.1 Probabilistic forecasting

Our setup in this lecture will be that of *probabilistic* (as opposed to *point*) forecasting. In this setting, a forecaster outputs a predicted distribution P of a given target Y of interest. We will generally focus on the case in which Y is real-valued. In probabilistic forecasting, both P and Y are random variables. To help give context, we can think of an underlying sequential problem indexed by time $t = 1, 2, 3, \ldots$, where:

> a forecaster predicts P_t for the distribution of Y_t, based on information available at times $s < t$.

This sequential problem will be "lurking in the background" throughout, but often we will hide the time dependence because it will be more convenient (and just as general) to talk about a single random forecaster P and random target Y.

1.2 Probabilities or quantiles?

It is most common to represent a probabilistic forecast P by its cumulative distribution function (CDF) F. Of course, instead of representing P by its CDF F, we can choose to represent it by its quantile function F^{-1}. To be clear, the quantile function F^{-1} is always well-defined, even when F is not continuous, or injective, and it is defined precisely as

$$F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\}.$$

When F is not continuous, this shares some but not all properties of the usual inverse; recall the properties you derived on the homework.

An important note: typically these objects, the CDF F or quantile function F^{-1}, will be discretized in practice, and *converting from one to the other will not always be possible/practical, particularly in the*

tails. Thus whether we choose at the outset to represent a predicted distribution by a CDF or quantile function can have nontrivial implications on what tools we have at our disposal for downstream tasks, like evaluation and aggregation.

Later, in Section 4, we will discuss the differences between aggregating probabilities (CDFs) and aggregating quantiles, in detail. Here, we provide some higher-level practical considerations to keep in mind when choosing to represent forecasts via probabilities versus quantiles.

- A quantile representation for forecasts alleviates concerns about choosing the "right" bins for probability forecasts. Choosing bins in practice can be highly nonobvious, both in terms of the "right" scale, and the "right" range.
- Quantiles can be directly fit a straightforward way, via quantile regression, and this can be carried out jointly over multiple probability levels. Constructing flexible (conditional) probability distributions can often be a more complex modeling task.
- For a discrete response that is concentrated around low values (such as counts), a quantile representation may be undesirable since "exact" quantiles only exist at a coarse set of probability levels (e.g., a reported level 0.05 quantile may actually be a level 0.09 quantile).

2 Scoring rules

We now turn to discussing scoring rules, for evaluating probabilistic forecasts. A *scoring rule* (or simply a score) is a function S that maps a predicted distribution P and target value y to a real or infinite number $S(P, y) \in \overline{\mathbb{R}}$ (where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ denotes the extended reals). We will, by convention, always take the score to be negatively-oriented: a lower value means a better forecast. Thus to be clear, in practice (in the sequential setting): we would compare two forecasters $P_t, \tilde{P}_t$ over times $t = 1, 2, 3, \ldots$ by comparing

$$\frac{1}{T}\sum_{t=1}^{T} S(P_t, Y_t) \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} S(\tilde{P}_t, Y_t),$$

with a lower average score suggesting a better forecaster.

In general, we use the notation $S(P, Q)$ to denote the expectation of the score over draws $Y \sim Q$,

$$S(P, Q) = \mathbb{E}_{Y\sim Q}[S(P, Y)] = \int S(P, y)\, dQ(y).$$

2.1 Proper scores

A scoring rule S is said to be *proper* if

$$S(P, Q) \geq S(Q, Q), \quad \text{for all } P, Q, \tag{1}$$

and *strictly proper* if strict inequality holds in (1) for $P \neq Q$. To quote Gneiting and Raftery (2007), when Q denotes the distribution of Y and the score S is proper: "the forecaster has no incentive to predict any $P \neq Q$ and is encouraged to report his or her true belief".

Several examples of proper scores are given below. (An example of an intuitively appealing but improper score will also be given.) Throughout, we will flexibly write the scores as taking densities, CDFs, or quantile functions as inputs (instead of distributions), as convenient.

Log score. For a forecast that has density or probability mass function p, *log score* is defined by

$$\mathrm{LogS}(p, y) = -\log p(y). \tag{2}$$

This is a strictly proper score. To see this, we compute:

$$\mathrm{LogS}(p, q) - \mathrm{LogS}(q, q) = \int \log \frac{q(y)}{p(y)} q(y)\, dy, \tag{3}$$

which recall is the Kullback-Leibler (KL) divergence between q and p, often denoted $\mathrm{KL}(q, p)$ (note carefully the order of arguments here). This is known to be nonnegative, and positive for $p \neq q$.

Log score (2) very sharply penalizes forecasts that place insufficiently low probability on events that materialize: for small $p(y)$, the score $-\log p(y)$ is very large (and asymptotes at $p(y) = 0$).

Interlude: connection to Bregman divergence. KL divergence is a special instance of a *Bregman divergence*, which is of the form (note carefully the order of arguments)

$$d_\phi(q, p) = \phi(q) - \phi(p) - \langle D\phi(p), q - p \rangle, \tag{4}$$

for a convex differentiable function ϕ, with derivative $D\phi$, and suitable inner product $\langle \cdot, \cdot \rangle$. In particular, KL divergence (3) is a Bregman divergence (4) when $\phi(p) = \int p(y) \log p(y)\, dy$, and $\langle p, q \rangle = \int p(y) q(y)\, dy$.

The connection between log score and Bregman divergence is no coincidence, and the connection between proper scores and Bregman divergences runs much deeper. If ϕ is strictly convex, then by the first-order characterization for convexity, we have $d_\phi(q, p) \geq 0$, with strict inequality for $p \neq q$. Due to this fact, if S is any score for which we have the Bregman representation (for strictly convex ϕ)

$$S(p, q) - S(q, q) = d_\phi(q, p), \tag{5}$$

then S is strictly proper. In essence, the converse is true as well. The formal details will be covered in Section 2.2.

Quadratic score. For a forecast that has density or probability mass function p, *quadratic* or *Brier score* is defined by

$$\mathrm{QuadS}(p, y) = -2p(y) + \|p\|_2^2, \tag{6}$$

where $\|p\|_2^2 = \int p(y)^2\, dy$. This is a strictly proper score. To see this, we compute:

$$S(p, q) - S(q, q) = \|p\|_2^2 - 2 \int p(y) q(y)\, dy + \|q\|_2^2 = \|p - q\|_2^2,$$

which is the L^2 distance between p and q. It is nonnegative, and positive for $p \neq q$.

Quadratic score (6) is more robust than log score (2), in the sense that it is gentler in penalizing forecasts that place low probability on events that materialize.

Linear score. For a forecast that has density or probability mass function p, *linear score* is defined by

$$\mathrm{LinS}(p, y) = -p(y).$$

We mention this score because it may seem intuitively appealing, but is *not proper* (and is basically never used in practice). To see this, let q be the standard normal density, and let p be the uniform density on the interval $(-\epsilon, \epsilon)$, for some $\epsilon > 0$ to be specified later. Then

$$\begin{aligned}
\mathrm{LinS}(p, q) - \mathrm{LinS}(q, q) &= \int q(y)^2\, dy - \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} q(y)\, dy \\
&= \frac{1}{2\pi} \int e^{-y^2}\, dy - \frac{1}{\sqrt{2\pi}} \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} e^{-y^2/2}\, dy \\
&= \frac{1}{\sqrt{2\pi}} \left(1/\sqrt{2} - \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} e^{-y^2/2}\, dy \right).
\end{aligned}$$

The last display will be negative for small enough $\epsilon > 0$: in particular, for $\epsilon < \sqrt{\log 2}$, the integrand in the above satisfies $e^{-y^2/2} > 1/\sqrt{2}$, so

$$\mathrm{LinS}(p, q) - \mathrm{LinS}(q, q) < \frac{1}{\sqrt{2\pi}} \left(1/\sqrt{2} - \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} 1/\sqrt{2}\, dy \right) < 0.$$

Continuous ranked probability score. Not all probabilistic forecasts can be expressed in terms of a density or probability mass function. For example, some forecast distributions might place a point mass at zero, and would be otherwise continuous. The *continuous ranked probability score* (CRPS) accommodates this and, in general, any forecast expressed in terms of a predicted CDF F. It is defined by

$$\mathrm{CRPS}(F, y) = \int (F(x) - 1\{y \leq x\})^2 \, dx. \tag{7}$$

This is a strictly proper score. To see this, we compute:

$$\begin{aligned}
\mathrm{CRPS}(F, G) - \mathrm{CRPS}(G, G) &= \int \Big(F(x)^2 - G(x)^2 - 2(F(x) - G(x)) \mathbb{E}_{Y \sim G}[1\{Y \leq x\}] \Big) \, dx \\
&= \int \Big(F(x)^2 - G(x)^2 - 2(F(x) - G(x)) G(x) \Big) \, dx \\
&= \int (F(x) - G(x))^2 \, dx.
\end{aligned}$$

This is the Cramér-von Mises distance between F and G. It is nonnegative, and positive for $p \neq q$.

CRPS (7) has become recently quite popular in various applied forecasting communities (e.g., in meterological sciences), in part because it is considered more robust than log score, and as already mentioned, it is more broadly applicable, beyond log (2) and quadratic (6) scores. However, unlike the latter two, CRPS is nontrivial to compute. Some authors have pursued quadrature-based approximations to the integral defining CRPS. A simpler approximation is given by noting its connection to weighted interval score/quantile loss, which will be covered just a bit later on.

Interval score. For an interval forecast expressed as $[\ell_\alpha, u_\alpha]$, where the lower and upper endpoints $\ell_\alpha = F^{-1}(\alpha/2)$ and $u_\alpha = F^{-1}(1 - \alpha/2)$ are the predicted $\alpha/2$ and $1 - \alpha/2$ quantiles of the forecast distribution, *interval score* is defined by

$$\mathrm{IS}_\alpha([\ell_\alpha, u_\alpha], y) = (u_\alpha - \ell_\alpha) + \frac{2}{\alpha} \cdot \mathrm{dist}(y, [\ell_\alpha, u_\alpha])$$

where $\mathrm{dist}(y, S) = \inf_{x \in S} |x - y|$ for a set S. We can see that this combines a reward for sharpness (first term) and a penalty for miscoverage (second term). For a collection of interval forecasts $[\ell_\alpha, u_\alpha]$, $\alpha \in \mathcal{A}$, *weighted interval score* (WIS) is defined by

$$\begin{aligned}
\mathrm{WIS}_{\mathcal{A}}\big(\{[\ell_\alpha, u_\alpha]\}_{\alpha \in \mathcal{A}}, \, y\big) &= \sum_{\alpha \in \mathcal{A}} \alpha \cdot \mathrm{IS}_\alpha([\ell_\alpha, u_\alpha], y) \\
&= \sum_{\alpha \in \mathcal{A}} \Big(\alpha(u_\alpha - \ell_\alpha) + 2 \cdot \mathrm{dist}(y, [\ell_\alpha, u_\alpha]) \Big).
\end{aligned} \tag{8}$$

Weighted interval score (and thus interval score, as effectively a special case) is a proper score for predicting the quantiles at levels $\alpha/2$ and $1 - \alpha/2$, over $\alpha \in \mathcal{A}$. To be clear, here we say (by analogy to (1)) that a score S defined in terms of a discrete set of predicted quantiles $\{q_\tau\}_{\tau \in \mathcal{T}}$ is proper if it holds that

$$S\big(\{q_\tau\}_{\tau \in \mathcal{T}}, \, F\big) \geq S\big(\{F^{-1}(\tau)\}_{\tau \in \mathcal{T}}, \, F\big), \quad \text{for all } \{q_\tau\}_{\tau \in \mathcal{T}}, \text{ and all } F, \tag{9}$$

where as usual $S(\{q_\tau\}_{\tau \in \mathcal{T}}, F) = \mathbb{E}_{Y \sim F}[S(\{q_\tau\}_{\tau \in \mathcal{T}}, Y)]$. Propriety of weighted interval score can be verified by drawing a connection to quantile loss, covered shortly.

Quantile score. For a forecast expressed as a collection of predicted quantiles q_τ, $\tau \in \mathcal{T}$, *quantile score* (QS) is defined by

$$\mathrm{QS}_{\mathcal{T}}\big(\{q_\tau\}_{\tau \in \mathcal{T}}, \, y\big) = \sum_{\tau \in \mathcal{T}} \rho_\tau(y - q_\tau), \tag{10}$$

where $\rho_\tau(u) = \tau |u|$ for $u \geq 0$ and $(1 - \tau)|u|$ for $u < 0$, often called the tilted ℓ_1 loss at the level τ. This is nothing else than the loss that is the standard choice in *quantile regression*. That quantile score is proper, in the sense of (9), follows from the standard fact that the true level τ quantile is a minimizer of the expected tilted ℓ_1 loss at the level τ.

Connection between WIS and QS. The following is an connection between WIS (8) and QS (10): for any set $\mathcal{A}$ of error levels, if we define a corresponding set $\mathcal{T} = \cup_{\alpha \in \mathcal{T}} \{\alpha/2, 1-\alpha/2\}$ of probability levels, then it holds that

$$\mathrm{WIS}_{\mathcal{A}}\big(\{[\ell_\alpha, u_\alpha]\}_{\alpha \in \mathcal{A}},\, y\big) = 2 \cdot \mathrm{QS}_{\mathcal{T}}\big(\{q_\tau\}_{\tau \in \mathcal{T}},\, y\big). \tag{11}$$

The equivalence (11) between WIS and QS can be verified by straightforward algebra. Apart from being interesting, and lending a new interpretation to how quantile score (10) combines a sharpness reward with a miscoverage penalty (recall the discussion following (8)), it is also a useful observation as it provides an indirect but simple proof that WIS (8) is proper, in the sense of (9).

Connection between CRPS and QS. The connections continue: interestingly, CRPS (7) also has an equivalent formulation in terms of QS: if F differentiable, with density f, and admits an expectation, $\int y f(x)\, dy < \infty$, then

$$\int (F(x) - 1\{y \le x\})^2\, dx = 2 \int \rho_\tau(y - F^{-1}(\tau))\, d\tau. \tag{12}$$

The right-hand side is an integral of QS all probability levels $\tau \in [0,1]$. This is quite interesting because these two error metrics—CRPS and QS—are motivated from very different perspectives, not to mention different parametrizations (CDF space versus quantile space).

The equivalence in (12) appears to have been discovered by Laio and Tamea (2007); their argument uses integration by parts, but it ignores a few subtleties, and so here we reproduce the proof from Fakoor et al. (2021), which basically just fills in a few missing details. Starting with the right-hand side in (12), we can substitute $x = F^{-1}(\tau)$ to rewrite the integral as

$$2 \int (1\{y \le F^{-1}(\tau)\} - \tau)(F^{-1}(\tau) - y)\, d\tau = 2 \int (1\{y \le x\} - F(x))(x - y) f(x)\, dx.$$

Let $u'(x) = 2(1\{y \le x\} - F(x))f(x)$ and $v(x) = x - y$. The idea is now to use integration by parts, but there are two subtleties. First, one has to be careful about framing the antiderivative u of u', since $x \mapsto 1\{y \le x\}$ is not classically differentiable. Note that we can actually redefine u' to be

$$u'(x) = 2(1\{y \le x\} - F(x))(f(x) - \delta_y(x)),$$

where δ_y is the Dirac delta function centered at y, because the "extra" piece integrates to zero:

$$\int 2(1\{y \le x\} - F(x))(x - y)\delta_y(x)\, dx = 2(1\{y \le x\} - F(x))(x - y)\Big|_{x=y} = 0.$$

With this new definition of u', its antiderivative is rigorously

$$u(x) = -(1\{y \le x\} - F(x))^2,$$

because, in the distributional sense, the derivative of the heavyside function $x \mapsto 1\{y \le x\}$ is indeed the delta function δ_y. Thus we have

$$\begin{aligned}
\int u'(x)v(x)\, dx &= u(x)v(x)\Big|_{-\infty}^{\infty} - \int u(x)v'(x)\, dx \\
&= -(1\{y \le x\} - F(x))^2(x - y)\Big|_{-\infty}^{\infty} + \int (1\{y \le x\} - F(x))^2\, dx.
\end{aligned}$$

The second subtlety is to show that the first term above is indeed zero. This is really a question of how fast the tails of F decay. As F yields an expectation, note that we must have $1 - F(x) \lesssim x^{-1}$ (by Markov's inequality). Hence

$$(1 - F(x))^2 x \lesssim x^{-1} \to 0 \quad \text{as } x \to \infty,$$

and the other limit, as $x \to -\infty$, is similar.

Quick recap. Here is a quick recap of what we learned on our tour of scores. Log score (2) and quadratic score (6) are classical choices, each strictly proper, with the former being more aggressive at penalizing underconfidence assigned to rare events. CRPS (7) is a more modern alternative, also strictly proper, which has become popular due to its robustness and generality, but is nontrivial to compute.

For forecasts specified in terms of quantiles at discrete probability levels, WIS (8) and QS (10) are each proper, simple, and popular choices, motivated from different perspectives; and in fact they turn out to be equivalent (11). Moreover, CRPS turns out to be equivalent to an integrated quantile score (12) over all probability levels. Thus a simple approximation to CRPS in practice would be to discretize at any desired set of probability levels and use this to define WIS/QS; propriety is always maintained, in the sense of (9), for any choice of discretization.

Figures 1 and 2 visualize some of these scores in action.

2.2 Bregman representation

Below we state an important result from Gneiting and Raftery (2007) that shows that the Bregman representation (5) which we saw was possible for log score (and quadratic score, and CRPS) is "no accident", and in a precise sense, characterizes *all* proper scores.

Before stating the result, we must introduce some terminology and concepts. Recall that by convention, we take all scores to be negatively-oriented. First, we say a score S is *regular* if $S(P, P)$ is real-valued for all P (i.e., we allow $S(P, Q) = \infty$ for $P \neq Q$). Regularity is essentially needed in order for subderivatives (defined shortly) to make sense but we omit the details.

Next, we refine the definition of propriety: for a class $\mathcal{P}$ of probability distributions, we say a score S is *proper relative to* $\mathcal{P}$ if (1) holds for $P, Q \in \mathcal{P}$, with *strictly proper relative to* $\mathcal{P}$ again meaning that the inequality is strict for $P \neq Q$.

Lastly, for a function ϕ acting over a set of distributions $\mathcal{P}$, with each $P \in \mathcal{P}$ having a sample space $\mathcal{Y}$, we say that $D\phi(P, \cdot)$ is a *subderivative* of ϕ at $P \in \mathcal{P}$ provided that

$$\phi(Q) \geq \phi(P) + \int D\phi(P, y)\, d(Q - P)(y), \quad \text{for all } Q \in \mathcal{P}. \tag{13}$$

We are now ready to state the result.

Theorem 1. *Let ϕ be a convex function acting over a set of distributions $\mathcal{P}$. Define for $P \in \mathcal{P}$ and $y \in \mathcal{Y}$ the regular score*

$$S(P, y) = -\phi(P) - D\phi(P, y) + \int D\phi(P, z)\, dP(z), \tag{14}$$

where $D\phi(P, \cdot)$ is a subderivative of ϕ at P. Then for any $P, Q \in \mathcal{P}$,

$$S(P, Q) - S(Q, Q) = \underbrace{\phi(Q) - \phi(P) - \int D\phi(P, y)\, d(Q - P)(y)}_{d_\phi(Q, P)}, \tag{15}$$

which is the Bregman divergence with respect to ϕ, and the score S is proper relative to $\mathcal{P}$.

Conversely, if S is a regular score that is proper relative to $\mathcal{P}$, then S can be written in the form (14) with respect to the convex function $\phi(P) = -S(P, P)$, and the Bregman representation (15) holds.

Finally, the above equivalence also holds when the terms convex and proper are replaced by strictly convex and strictly proper, respectively.

Proof. If ϕ is convex and we define S according to (14), then direct calculation leads to $S(P, Q) - S(Q, Q) = -\phi(P) - \int D\phi(P, y)\, dQ(y) + \int D\phi(P, y)\, dP(y) + \phi(Q) + \int D\phi(Q, y)\, dQ(y) - \int D\phi(Q, y)\, dQ(y)$ and the last two terms cancel, which establishes (15). Meanwhile, $d_\phi(Q, P) \geq 0$ for any $P, Q \in \mathcal{P}$ by definition of the subderivative (13), which proves that S is proper relative to $\mathcal{P}$.

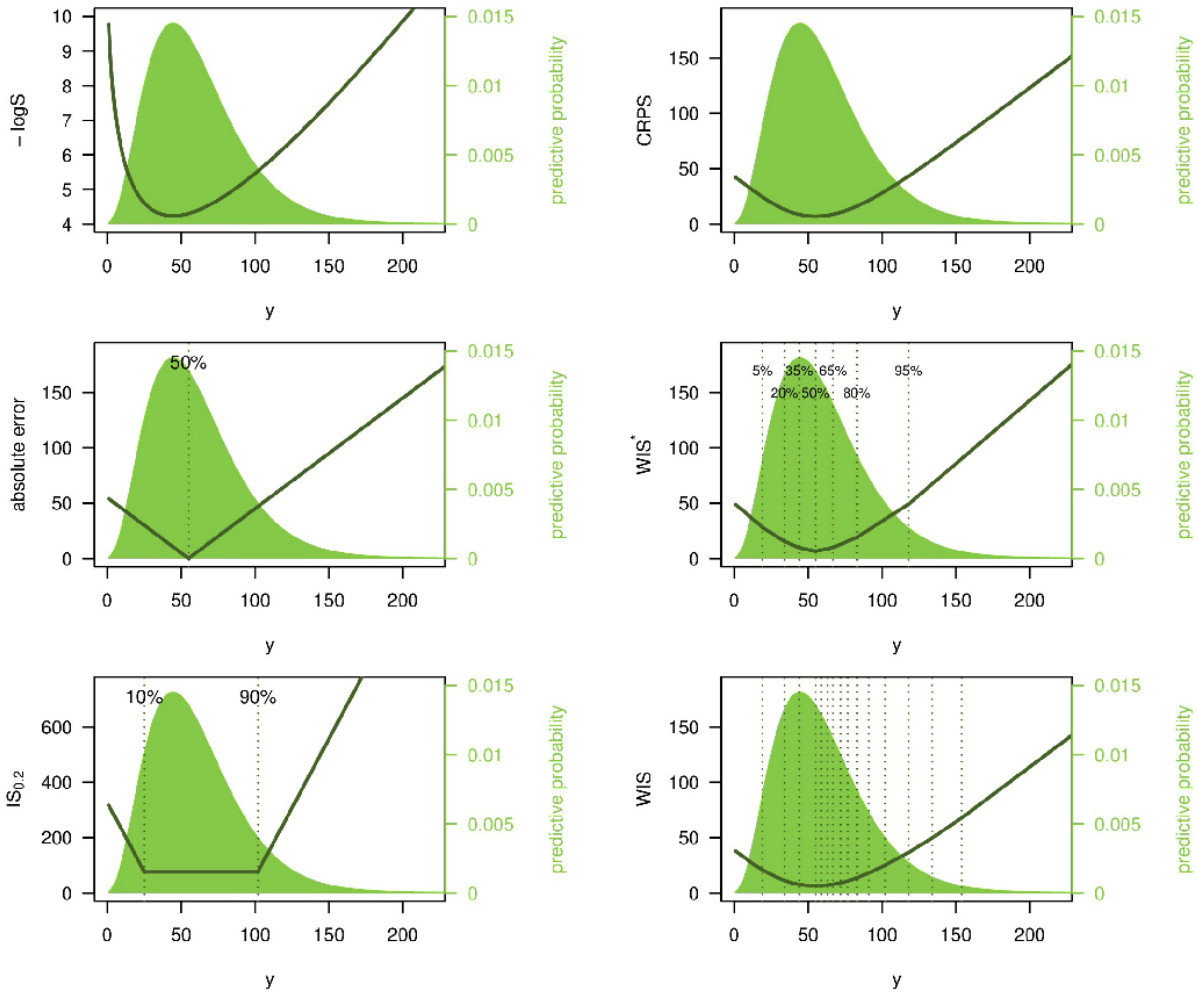

Figure 1: *Various scores visualized as functions of y, based on the predicted distribution plotted in green. Here WIS* and WIS denote two versions of weighted interval score at a coarser and finer set of probability levels, respectively. Credit: Bracher et al. (2021).*

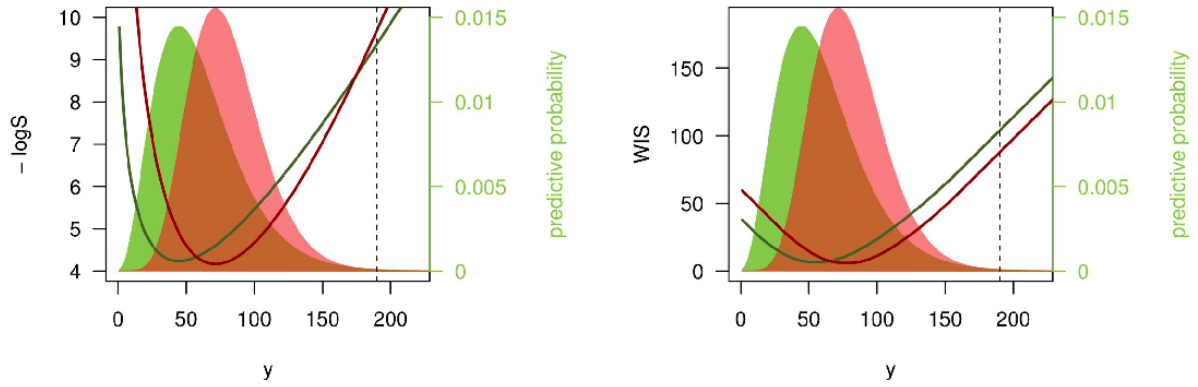

Figure 2: *Comparison between log score and WIS for two predicted distributions: F in green and G in red. Note that G has a higher expectation, while F is more dispersed. If we consider an event y that is marked as a dashed vertical line, then we see log score prefers F to G, while WIS prefers G to F. Credit: Bracher et al. (2021).*

For the other direction, if S is regular and proper relative to $\mathcal{P}$, then letting $\phi(Q) = -S(Q,Q)$ is equivalent to letting

$$\phi(Q) = -\inf_{P \in \mathcal{P}} S(P,Q) = \sup_{P \in \mathcal{P}} -S(P,Q).$$

For fixed P, the function $Q \mapsto -S(P,Q)$ is linear, and hence convex, and thus by the above ϕ is a pointwise supremum of convex functions and hence itself convex. Furthermore, we can see that $D\phi(P,\cdot) = -S(P,\cdot)$ is a valid subderivative of ϕ, since (13) becomes $-S(Q,Q) \geq -S(P,Q)$ for all $Q \in \mathcal{P}$, which is satisfied because S is proper relative to $\mathcal{P}$. Under the choice $D\phi(P,\cdot) = -S(P,\cdot)$, the score representation (14) is a tautology. The Bregman representation (15) again follows by simple algebra.

This completes the proof of the claimed equivalence, and the arguments for the strict version of the result follow similarly. □

Even though our focus has been (and will continue to be) real-valued forecasts, it is worth emphasizing that the previous theorem treats the sample space $\mathcal{Y}$ as arbitrary. In the case of categorical forecasts, where $\mathcal{Y} = \{1,\dots,k\}$, we get the following corollary, which is originally due to Savage (1971). We denote the standard k-dimensional probability simplex by $\Delta^k = \{p \in \mathbb{R}^k : p \geq 0,\ \sum_{i=1}^k p_i = 1\}$.

Corollary 1. *Let $\mathcal{Y} = \{1,\dots,k\}$. Then a regular score S, which we treat as acting on a vector of probabilities $p \in \Delta^k$, is proper with respect to the set of all distributions on $\{1,\dots,k\}$ if and only if*

$$S(p,i) = -\phi(p) - D_i\phi(p) + \langle D\phi(p), p\rangle, \quad \text{for } i = 1,\dots,k,$$

for a convex function ϕ on Δ^k, where $D\phi(p)$ denotes the subgradient of ϕ at p (with components $D_i\phi(p)$, $i = 1,\dots,k$). The same equivalence also holds when the terms convex and proper are replaced by strictly convex and strictly proper, respectively.

3 Modes of calibration

Next we turn to calibration. There are in fact many modes or "flavors" of calibration, and they are related but not all equivalent. We will walk through three such definitions, but first, we review some important preliminary concepts, which will help motivate the definitions. (Side note: translating calibration precisely to the sequential setting is a bit more tricky than it was for scoring rules, since it requires more than just writing expectations as averages; see Gneiting et al. (2007).)

3.1 Preliminary concepts

For any CDF F (fixed or random) and target random variable Y, we define

$$F^*(Y) = V \cdot F(Y) + (1-V) \cdot F(Y^-), \tag{16}$$

where $V \sim \text{Unif}(0,1)$ is independent of F and Y, and $F(y^-) = \lim_{x \to y^-} F(x)$. This is called the *probability integral transform* (PIT) associated with F and Y. If F is continuous (i.e., if F is a continuous function, and hence invertible) almost surely, then the PIT reduces to $F^*(Y) = F(Y)$.

Recall that if we take F to be *fixed* and equal to the CDF of Y, then $F^*(Y)$ (or $F(Y)$ in the continuous case) is itself standard uniform. You proved this fact on the homework fact. We can write this simply as

$$F^*(Y) \overset{d}{=} U,$$

where here and throughout the remainder of these lecture notes section, we denote U to denote a standard uniform random variable, $U \sim \text{Unif}(0,1)$, independent of everything else that is random. Recall also that if we take F^{-1} to be *fixed* and equal to the quantile function of Y, then

$$F^{-1}(U) \overset{d}{=} Y,$$

which you also proved on the homework. The definitions of probabilistic and marginal calibration, covered over the next two subsections, are based on requiring the above two properties to hold, respectively, in the case when F is also random.

3.2 Probabilistic calibration

A forecaster with predicted CDF F is said to be *probabilistically calibrated* for a target Y if

$$F^*(Y) \stackrel{d}{=} U \tag{17}$$

where recall $F^*(Y)$ is defined as in (16), and $U \sim \text{Unif}(0,1)$. To be explicit, here both F and Y are *random*. Another name for probabilistic calibration is *PIT calibration*, and another way of writing (17) is

$$\mathbb{P}\Big(V \cdot F(Y) + (1-V) \cdot F(Y^-) \leq \tau\Big) = \tau, \quad \text{for all } \tau \in [0,1],$$

for $V \sim \text{Unif}(0,1)$, independent of F and Y, which is typically how you will see it defined.

Quantile reformulation. When F is continuous almost surely, then we can reinterpret probabilistic calibration as follows: F is calibrated for Y if

$$\mathbb{P}(Y \leq F^{-1}(\tau)) = \tau, \quad \text{for all } \tau \in [0,1].$$

This is highly intuitive; for example, when we inspect the forecaster's predicted quantile at the level 0.9, the target Y should lie below this 90% of the time, and so on, for all probability levels.

Dispersion: over and under. If F places too little mass in the tails, then the PIT $F^*(Y)$ will be U-shaped, and its variance will be large compared to that of a uniform distribution. Conversely, if F places too much mass in the tails, then the PIT $F^*(Y)$ will be upside-down U-shaped, and its variance will be comparably small.

Recalling that we use $U \sim \text{Unif}(0,1)$ to denote a standard uniform random variable, this leads us to the following definitions, in the context of PIT calibration:

- F is said to be *overdispersed* for Y if $\text{Var}[F^*(Y)] < \text{Var}[U]$;
- F is said to be *underdispersed* for Y if $\text{Var}[F^*(Y)] > \text{Var}[U]$.

Figures 3 and 4 provide examples.

3.3 Marginal calibration

A forecaster with predicted CDF F is said to be *marginally calibrated* for a target Y if

$$F^{-1}(U) \stackrel{d}{=} Y, \tag{18}$$

where recall $U \sim \text{Unif}(0,1)$. Another way of expressing this is

$$\mathbb{E}[F(y)] = \mathbb{P}(Y \leq y), \quad \text{for all } y,$$

which is typically how you will see it defined. This also explains its name. Interestingly, marginal calibration is *not* the same as probabilistic calibration, and it is neither more general nor less general. There are nontrivial forecasters that satisfy (17) but not (18), and vice versa, to be discussed shortly.

Dispersion: over and under. When F is random and $U \sim \text{Unif}(0,1)$ is independent of F, note that we can then interpret $F^{-1}(U)$ as follows: first draw F, then draw a random variable according to F. Thus if F places too much mass on the tails, then the variance of $F^{-1}(U)$ will be comparably large, and if F places insufficient mass in the tails, then the variance of $F^{-1}(U)$ will be comparably small.

This leads to the following definitions, in the context of marginal calibration:

- F is said to be *overdispersed* for Y if $\text{Var}[F^{-1}(U)] > \text{Var}[Y]$;
- F is said to be *underdispersed* for Y if $\text{Var}[F^{-1}(U)] < \text{Var}[Y]$.

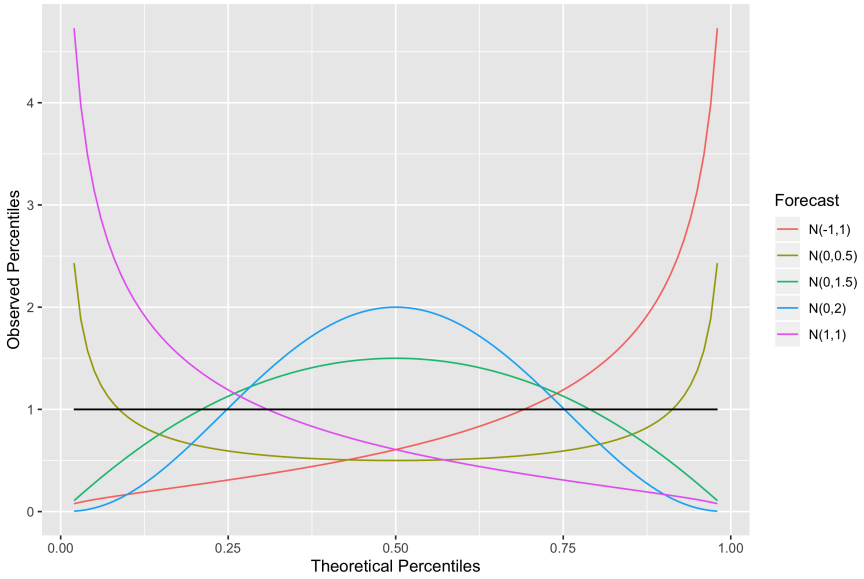

Figure 3: *Densities of PIT distributions for several simple normal forecasters, when the true target distribution is* $N(0,1)$. *Credit: Rumack et al. (2022).*

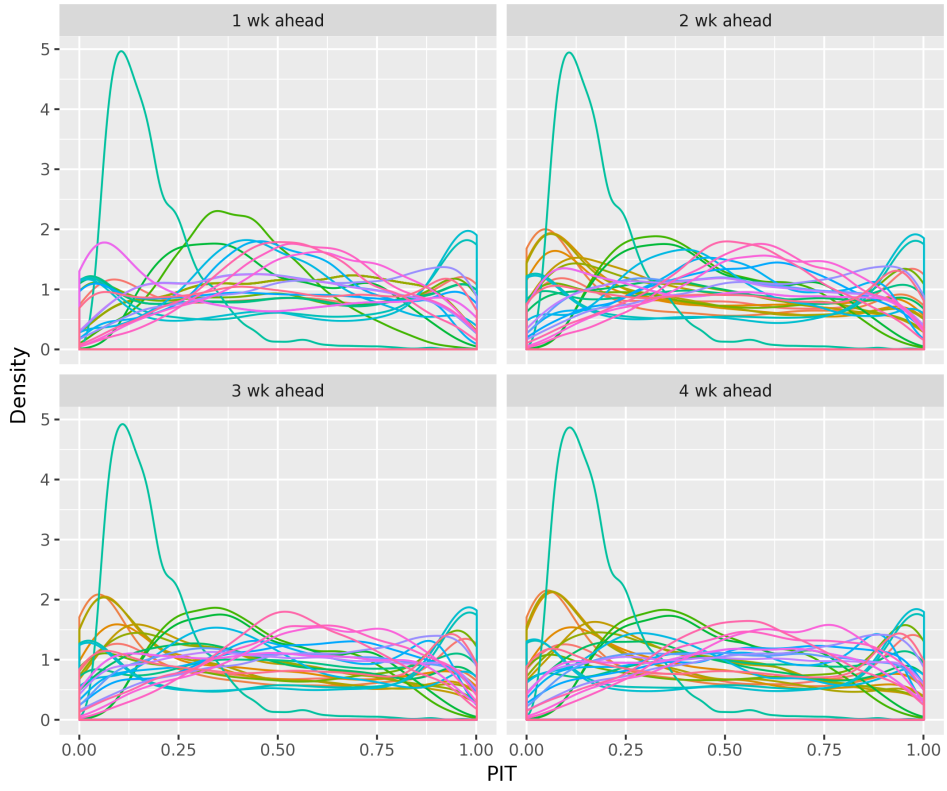

Figure 4: *Densities of PIT distributions from 27 forecasters submitted to the annual FluSight (seasonal influenza forecasting) challenges held by CDC, over 9 seasons (2010-11 to 2018-19). The PIT densities fall mostly into one of two categories: overdispersed with a peak around 0.5, and underdispersed with peaks at 0 and 1. (The outlier with a peak at 0.1 is the PIT density of a simple baseline forecaster.) Credit: Rumack et al. (2022).*

PIT versus marginal calibration. PIT calibration is a statement about the *joint* distribution of the forecaster F and target Y. Marginal calibration is not; it is only a statement, as its name suggests, about the *marginal* distributions of F and Y. In a sense, it can hence be much simpler to understand and reason about than PIT calibration.

To give an example, when we think about systematic miscalibration—a forecaster being overdispersed or underdispersed—we might often think about this in terms of the forecast distribution being too spread out or too peaked, respectively. But this intuition is really only justified for marginal calibration, and it is an incomplete way to think about PIT calibration. This is because the latter is about how F and Y behave *jointly*. Recall, overdispersion means that the PIT is too peaked, and underdispersion means that the PIT is too spread out. This can be entirely due to the dependence between F and Y, i.e., it can happen even when the forecast distribution F is marginally calibrated—it has "just enough spread".

Examples. Following Gneiting et al. (2007), suppose μ is drawn from $N(0,1)$, and Y is drawn (independently of μ) from $N(\mu,1)$. Consider (where we use F to denote both a distribution and a CDF):

- the *ideal* forecaster: $F = N(\mu, 1)$;
- the *climatological* forecaster: $F = N(0, 2)$;
- the *flipped* forecaster: $F = N(-\mu, 1)$;
- the *unfocused* forecaster: $F = \frac{1}{2}[N(\mu,1) + N(\mu+\xi,1)]$, for $\xi = \pm 1$ with equal probability, independent of Y, μ.

To gain intuition for the setup, you can think of an underlying sequential problem, where at time t, we draw $\mu_t \sim N(0,1)$, then draw $Y_t \sim N(\mu_t, 1)$ independently, and the forecasters are $F_t = N(\mu_t, 1)$ (ideal), $F_t = N(0,2)$ (climatological), and so on.

Direct calculations, as given in Section 2 of Gneiting et al. (2007), reveal the following:

- the ideal forecaster is both probabilistically and marginally calibrated;
- the climatological forecaster is both probalistically and marginally calibrated;
- the flipped forecaster is marginally but not probabilistically calibrated;
- the unfocused forecaster is probabilistically but not marginally calibrated.

3.4 Conditional calibration

When Y is binary, we can think of a forecaster as providing a predicted probability p of the event $Y = 1$. In this setting, a forecaster p is said to be *conditionally calibrated* if

$$\mathbb{E}[Y|p] = p, \quad \text{almost surely.} \tag{19}$$

This is highly intuitive; for example, when the forecaster outputs a predicted probability of 0.2, the event $Y = 1$ should materialize 20% of the time, and so on, for all probability levels.

Interestingly, and perhaps surprisingly, this is the same as probabilistic calibration in the binary seting. This is due to Gneiting and Ranjan (2013).

Theorem 2. *Let Y be binary valued, and suppose a forecaster outputs a predicted probability p of the event $Y = 1$, with associated predicted CDF*

$$F(y) = (1-p) \cdot 1\{y \geq 0\} + p \cdot 1\{y \geq 1\}.$$

Then F is probabilistically calibrated as in (17) if and only if p is conditionally calibrated as in (19).

Proof. Supposing p is probabilistically calibrated, we can write, for $V \sim \text{Unif}(0,1)$, independent of everything else,

$$F^*(Y) = V(1-p) \cdot (1-Y) + (1-p+Vp) \cdot Y.$$

Conditional calibration (19) says $Y \,|\, p \sim \text{Bern}(p)$, and fixing any $\tau \in [0,1]$, we can compute

$$\begin{aligned}\mathbb{P}(F^*(Y) \le \tau \,|\, p) &= \mathbb{P}\big(V(1-p)(1-Y) + (1-p+Vp)Y \le \tau \,\big|\, p\big) \\ &= \mathbb{P}\big(V(1-p) \le \tau \,\big|\, p\big) \cdot (1-p) + \mathbb{P}\big(Vp \le \tau - (1-p) \,\big|\, p\big) \cdot p.\end{aligned}$$

Denoting $a = \mathbb{P}(V(1-p) \le \tau \,|\, p)$ and $b = \mathbb{P}(Vp \le \tau - (1-p) \,|\, p)$, observe that

$$a = \frac{\tau}{1-p} \wedge 1, \quad b = \frac{\tau - (1-p)}{p} \vee 0.$$

Thus when $\tau \le 1-p$, we get $\mathbb{P}(F^*(Y) \le \tau \,|\, p) = a(1-p) = \tau$, and when $\tau > 1-p$, we get $\mathbb{P}(F^*(Y) \le \tau \,|\, p) = (1-p) + bp = \tau$. Marginalizing over p proves the probabilistic calibration property (17).

The other direction, that probabilistic calibration implies conditional calibration, is less elementary and we refer to the proof of Theorem 2.11 in Gneiting and Ranjan (2013) for details. □

It is also interesting to remark that this equivalence does *not* extend beyond binary outcomes. For three or more distinct levels of a discrete outcome, it is no longer true that PIT calibration implies conditional calibration. Gneiting and Resin (2022) provide a counterexample.

More broadly, it is worth remarking that the generalization of (19) beyond the binary setting is sometimes called *auto-calibration*, which requires the predicted CDF F to satisfy

$$Y \,|\, F \overset{d}{=} F^{-1}(U) \,|\, F, \quad \text{almost surely.} \tag{20}$$

A draw from $F^{-1}(U) \,|\, F$ is simply a random variable distributed according to F. Thus, in other words, the property (20) says that almost surely, conditional law of $Y \,|\, F$ should indeed be F. You'll sometimes see this written as $\mathcal{F}[Y|F] = F$, almost surely.

Auto-calibration (20) is stronger than all notions of calibration we have seen thus far, and it implies both probabilistic calibration (17) and marginal calibration (18). However, in general, it is not really possible to check whether auto-calibration holds in practice. See Gneiting and Resin (2022) for a discussion of this and related topics.

4 Probability versus quantile aggregation

Model aggregation is a rich and important topic in it of itself, and could easily be the topic of its own lecture. To give a broader context, model aggregation methods—also called ensemble methods—occupy a central place in machine learning, both in theory and in practice. Seminal work on this topic arose in the 1990s, with the development of Bayesian model averaging, bagging, boosting, and stacking.

The machine learning literature has mostly focused on ensembling point predictions, while ensembling distributions has an even longer tradition in probabilistic forecasting, dating back to the 1960s. Over the next two sections, we'll focus on a simple class of aggregation methods, linear ones, from the perspective of probabilistic forecasting. The current section is adapted from Section 3 of Fakoor et al. (2021).

For convenience, we'll use the term *average* to refer to a weighted linear combination where the weights are nonnegative and sum to 1. For each $j = 1, \ldots, p$, let F_j be a cumulative distribution function (CDF), assumed continuous, and let $f_j = F_j'$ be its probability density function; let $Q_j = F_j^{-1}$ denote the corresponding quantile function, also assumed continuous, and let $q_j = Q_j'$ the quantile density function. A standard fact that relates these objects:

$$q_j(t) = \frac{1}{f_j(Q_j(t))} \quad \text{and} \quad f_j(x) = \frac{1}{q_j(F_j(x))}. \tag{21}$$

The first fact can be checked by differentiating $Q_j(F_j(x)) = x$, applying the chain rule, and reparametrizing via $t = F_j(x)$. The second follows similarly via $F_j(Q_j(t)) = t$.

4.1 Two ways of averaging

We compare and contrast two ways of averaging distributions. The first way is in probability space, where we define for weights $w_j \geq 0$, $j = 1, \ldots, p$ such that $\sum_{j=1}^p w_j = 1$,

$$F = \sum_{j=1}^p w_j F_j.$$

The associated density is simply $f = \sum_{j=1}^p w_j f_j$ since differentiation is a linear operator. The second way to average is in quantile space, defining

$$\bar{Q} = \sum_{j=1}^p w_j Q_j,$$

where now $\bar{q} = \sum_{j=1}^p w_j q_j$ is the associated quantile density, again by linearity of differentiation. Denote the CDF and probability density associated with the quantile average by $\bar{F} = \bar{Q}^{-1}$, and $\bar{f} = \bar{F}'$. Note that from (21), we can reason that $\bar{f}$ is highly *nonlinear* as a function of f_j, $j = 1, \ldots, p$.

A simple example can go a long way to illustrate the differences between the distributions resulting from probability and quantile averaging. In Figure 5, we compare these two ways of averaging on a pair of normal distributions with different means and variances. Here we see that probability averaging produces the familiar mixture of normals, which is bimodal. The result of quantile averaging is very different: it is always unimodal, and instead of interpolating between the tail behaviors of f_1 and f_2 (as f does), it appears that *both* tails of $\bar{f}$ are generally thinner than those of f_1 and f_2.

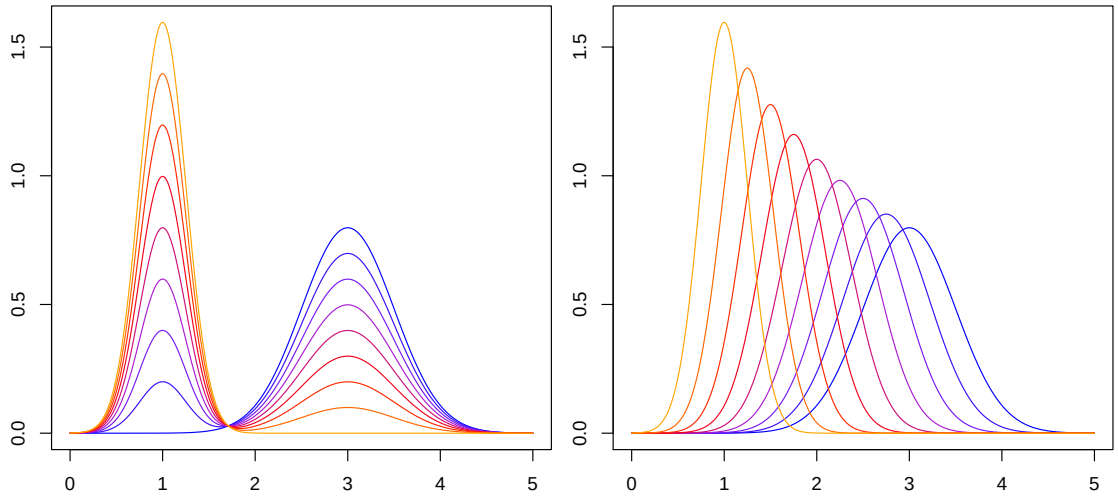

Figure 5: *Densities that result from a probability average (left) or quantile average (right) of two normal distributions* $N(1, 0.25^2)$ *and* $N(3, 0.5^2)$*, as the weight on the first density varies from 1 (drawn in orange) to 0 (drawn in blue). Credit: Fakoor et al. (2021).*

It seems that quantile averaging is doing something that is both like translation and scaling in probability density space. Next we explain this phenomenon precisely by recalling a classic result.

4.2 Shape preservation

An aggregation procedure H is said to be *shape-preserving* if, for any location-scale family $\mathcal{P}$ (such as Normal, t, Laplace, or Cauchy) whose elements differ only by scale and location parameters, we have

$$F_j \in \mathcal{P},\ j = 1, \ldots, p \implies F = H(F_1, \ldots, F_p) \in \mathcal{P}.$$

Probability averaging is clearly not shape-preserving, however, interestingly, quantile averaging is: if each F_j a member of the same location-scale family with a base CDF L, then we can write $F_j(x) = L((x -$

$\theta_j)/\sigma_j)$, and thus $Q_j(t) = \theta_j + \sigma_j L^{-1}(t)$, so $\bar{Q}$ is still of the form $\theta + \sigma L^{-1}$ and $\bar{F}$ is also in the location-scale family. The next proposition collects this and related results from the literature, due to Thomas and Ross (1980); Genest (1992).

Proposition 1.

(i) Quantile averaging is shape-preserving.

(ii) Location-scale families $\mathcal{P}$ are the only ones with respect to which quantile averaging is a closed operation (meaning $F_j \in \mathcal{P}$, $j = 1, \ldots, p$ implies $\bar{F} \in \mathcal{P}$).

(iii) Quantile averaging is the only aggregation method H, of those satisfying (for h not depending on t):

$$H(F_1, \ldots, F_p)^{-1}(t) = h(Q_1(t), \ldots, Q_p(t)),$$

that is shape-preserving.

The parts of Proposition 1, taken together, suggest that quantile averaging is somehow "tailor-made" for shape preservation in a location-scale family—which can be seen as either a pro or a con, depending on the application one has in mind. To elaborate, suppose that in a quantile regression ensembling application, each base model outputs a normal distribution for its predicted distribution at each x (with different means and variances). If the normal assumption is warranted (i.e., it actually describes the data generating distribution) then we would want our ensemble to retain normality, and quantile averaging would do exactly this. But if the normal assumption is used only as a working model, and we are looking to combine base predictions as a way to construct some flexible and robust model, then the shape-preserving property of quantile averaging would be problematic. In general, to model arbitrary distributions without imposing strong assumptions, we are therefore driven to use linear combinations of quantiles that allow *different aggregation weights to be used for different quantile levels*, of the form $\bar{Q}(t) = \sum_{j=1}^p w_j(t) Q_j(t)$.

4.3 Moments and sharpness

Next, we consider moments of the distributions returned by probability and quantile averages, recalling a result due to Lichtendahl et al. (2013). For a distribution G, we denote its uncentered moment of order $k \geq 1$ by $m_k(G) = \mathbb{E}_{X \sim G}[X^k]$.

Proposition 2.

(i) A probability and quantile average always have equal means: $m_1(F) = m_1(\bar{F})$.

(ii) A quantile average is always sharper than a probability average: $m_k(\bar{F}) \leq m_k(F)$ for any even $k \geq 2$.

Note that sharpness is only a desirable property if it does not come at the expense of calibration. With this in mind, the above result cannot be understood as a pro or con of quantile averaging without any context on calibration—this will be studied in Section 5. That said, the relative sharpness of quantile averages to probability averages is an important general phenomenon to be aware of.

4.4 Tail behavior

Lastly, we study the action of quantile averaging on the tails of the subsequent probability density. Simply starting from $\bar{Q} = \sum_{j=1}^p w_j Q_j$, differentiating, and using (21), we get

$$\frac{1}{\bar{f}(\bar{Q}(t))} = \sum_{j=1}^p \frac{w_j}{f_j(Q_j(t))}.$$

That is, the probability density $\bar{f}$ at the level t quantile is a (weighted) *harmonic mean* of the densities f_j at their respective level t quantiles. Since harmonic means are generally (much) smaller than arithmetic means, we would thus expect $\bar{f}$ to have thinner tails than f.

The next result formalizes this. We use $g(x) = o(h(x))$ to mean $g(x)/h(x) \to 0$ as $x \to \infty$, and $g(x) \asymp h(x)$ to mean $g(x)/h(x) \to c \in (0, \infty)$ as $x \to \infty$.

Proposition 3. *Assume that $p = 2$, $f_2(x) = o(f_1(x))$, and the weights w_1, w_2 are nontrivial (they lie strictly between 0 and 1). Then the density from probability averaging satisfies $f(v) \asymp f_1(v)$. Assuming further that f_1 is log-concave, the density from quantile averaging satisfies $\bar{f}(v) = o(f_1(v))$.*

This result is due to Fakoor et al. (2021). The assumption that f_1 is log-concave for the quantile averaging result is stronger than it needs to be (as is the restriction to $p = 2$), but is used to simplify the proof.

Proposition 3 reiterates the importance of allowing for level-dependent weights in a linear combination of quantiles. For applications in which there is considerable uncertainty about extreme events (especially ones in which there is disagreement in the degree of uncertainty between individual base models), we would not want an ensemble to de facto inherit a particular tail behavior—whether thin or thick—but want to endow the aggregation procedure with the ability to adapt its tail behavior as needed.

5 Aggregation and (mis)calibration

In this last section, we cover results on the way linear aggregation rules affect calibration. The setup is as follows. We are given forecasters in the form of predicted CDFs or predicted quantile functions. We assume that each forecaster is calibrated, by some definition. We then take an average of CDFs or quantile functions, and ask whether the result is calibrated, or not.

The results are summarized in Table 1. In a sense, they are surprising—linear aggregation rules are quite common and popular in practice, and so the fact that they can destroy PIT calibration in the case of a probability average, or marginal calibration in the case of a quantile average, is fairly disturbing. On the other hand, their proofs are very simple, as we will see in the following subsections, and arguably these results are not too unexpected, at least in hindsight.

	PIT calibration	Marginal calibration
Probability average	Overdispersed	Calibrated
Quantile average	?	Underdispersed

Table 1: *Summary of known results on aggregation and calibration. In the first column, each forecaster is assumed to be PIT calibrated, and in the second column, each is assumed to be marginally calibrated. The result in the bottom left entry, on the general behavior of a quantile average of PIT calibrated forecasters, is currently unknown.*

5.1 PIT calibration, probability average

We start with the assumption that each forecaster is PIT calibrated, in the sense of (17), and inspect their probability average. The next result is essentially due to Ranjan and Gneiting (2010).

Theorem 3. *Let F_j, $j = 1, \ldots, p$ be CDF forecasts for some target random variable Y, and for fixed weights $w_j \geq 0$, $j = 1, \ldots, p$, such that $\sum_{j=1}^p w_j = 1$, consider*

$$F = \sum_{j=1}^p w_j F_j.$$

Assume that for some $i \neq j$, we have $F_i \neq F_j$ with positive probability, and $w_i w_j > 0$.

a. For any convex function φ, it holds that

$$\mathbb{E}[\varphi(F^*(Y))] \leq \sum_{j=1}^p w_j \mathbb{E}[\varphi(F_j^*(Y))],$$

with strict inequality when φ is strictly convex.

b. For any $k \geq 1$, the k^{th} moment of $F^*(Y)$ is thus strictly smaller than the maximum of k^{th} moments of $F_j^*(Y)$, $j = 1, \ldots, p$. This also holds for the central moments if each $F_j^*(Y)$ has the same mean.

c. *In particular, if F_j, $j = 1, \ldots, p$ are each probabilistically calibrated, as in (17), then F is overdispersed, meaning that* $\mathrm{Var}(F^*(Y)) < \mathrm{Var}(U)$, *where* $U \sim \mathrm{Unif}(0,1)$.

Proof. A key realization to begin is that

$$F^*(Y) = \sum_{j=1}^p w_j F_j^*(Y),$$

which we can see by noting that F inherits the left-discontinuities of F_j, $j = 1, \ldots, p$. For part a, we simply calculate for any convex φ,

$$\mathbb{E}\bigg[\varphi\bigg(\sum_{j=1}^p w_j F_j^*(Y)\bigg)\bigg] \leq \sum_{j=1}^p w_j \mathbb{E}[\varphi(F_j^*(Y))],$$

by Jensen's inequality, with strict inequality when φ is strictly convex (by our assumption about the existence of a pair F_i, F_j that are not equal with positive probability, and $w_i w_j > 0$).

The first statement in part b is simply proved by considering $\varphi(x) = x^k$ for any positive integer k. In the case that $F_j^*(Y)$, $j = 1, \ldots, p$ all have the same mean, note that by taking $\varphi(x) = \pm x$, we learn that $F^*(Y)$ must also have the same mean, which proves the second statement in part b about central moments.

Lastly, part c is just the special case for the variance, i.e., second central moment, when each $F_j^*(Y) \sim \mathrm{Unif}(0,1)$, $j = 1, \ldots, p$. □

The fact that any linear combination of probabilistically calibrated CDF forecasters must be miscalibrated and systematically overdispersed (part c) is simple but striking. In the binary case, where $Y \in \{0,1\}$, the same is true when we replace the notion of probabilistic calibration with conditional calibration (recalling the equivalence between the two). This reproduces an earlier result from Ranjan and Gneiting (2010).

5.2 Marginal calibration, probability average

Now we move to the case where each forecaster is marginally calibrated, and still consider a probability average. The next result is very simple but noting, and once again emphasizes the differences between marginal and probabilistic calibration.

Theorem 4. *Let F_j, $j = 1, \ldots, p$ be CDF forecasts for some target random variable Y, and for fixed weights $w_j \geq 0$, $j = 1, \ldots, p$, such that $\sum_{j=1}^p w_j = 1$, consider*

$$F = \sum_{j=1}^p w_j F_j.$$

If F_j, $j = 1, \ldots, p$ are each marginally calibrated, as in (18), then F is also marginally calibrated.

Proof. We simply compute $\mathbb{E}[\sum_{j=1}^p w_j F_j(y)] = \sum_{j=1}^p w_j \mathbb{P}(Y \leq y) = \mathbb{P}(Y \leq y)$. □

5.3 Marginal calibration, quantile average

Keeping marginal calibration in mind, but moving to a quantile average (hence moving clockwise through the results in Table 1), we have the following result.

Theorem 5. *Let F_j^{-1}, $j = 1, \ldots, p$ be quantile forecasts for some target random variable Y, and for fixed weights $w_j \geq 0$, $j = 1, \ldots, p$, such that $\sum_{j=1}^p w_j = 1$, consider*

$$F^{-1} = \sum_{j=1}^p w_j F_j^{-1}.$$

Assume that for some $i \neq j$, we have $F_i^{-1} \neq F_j^{-1}$ with positive probability, and $w_i w_j > 0$.

a. *For any convex function* φ*, and* $U \sim \text{Unif}(0,1)$ *it holds that*

$$\mathbb{E}[\varphi(F^{-1}(U))] \leq \sum_{j=1}^{p} w_j \mathbb{E}[\varphi(F_j^{-1}(U))],$$

with strict inequality when φ *is strictly convex.*

b. *For any* $k \geq 1$*, the* k^{th} *moment of* $F^{-1}(U)$ *is thus strictly smaller than the maximum of* k^{th} *moments of* $F_j^{-1}(U)$*,* $j = 1, \ldots, p$*. This also holds for the central moments if each* $F_j^{-1}(U)$ *has the same mean.*

c. *In particular, if* F_j*,* $j = 1, \ldots, p$ *are each marginally calibrated, as in* (18)*, then* F *is underdispersed, meaning that* $\text{Var}(F^{-1}(U)) < \text{Var}(Y)$*.*

Proof. The proof is similar to that of Theorem 3. For part a, we calculate for any convex φ,

$$\mathbb{E}[\varphi(F^{-1}(U))] \leq \sum_{j=1}^{p} \mathbb{E}[\varphi(F_j^{-1}(U))],$$

by Jensen's inequality, with strict inequality when φ is strictly convex (again by our assumption about the existence of a pair F_i^{-1}, F_j^{-1} that are not equal with positive probability, and $w_i w_j > 0$). The rest of the proof is identical to that of Theorem 3. □

Again, the fact that any linear combination of marginally calibrated quantile forecasters must be miscalibrated and systematically underdispersed (part c) is simple but striking.

5.4 PIT calibration, quantile average

This case seems to elude simple analysis (at least for now). If the world was just, then one might hope (by symmetry in Table 1) that a quantile average of PIT calibrated forecasters would itself be calibrated. However, this is not the case, because—sadly—we know of at least one case where it can be shown analytically that the quantile average here is overdispersed.

Given what we know about the relationship between probability and quantile averages though (recall, e.g., Proposition 2), one might still hope that a quantile average is *less* overdispersed than the corresponding probability average, when the constituent forecasters are all PIT calibrated.

References

Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):1–15, 2021.

Rasool Fakoor, Taesup Kim, Jonas Mueller, Alexander Smola, and Ryan J. Tibshirani. Flexible model aggregation for quantile regression. arXiv: 2103.00083, 2021.

Christian Genest. Vincentization revisited. *Annals of Statistics*, 20(2):1137–1142, 1992.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013.

Tilmann Gneiting and Johannes Resin. Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. arXiv: 2108.03210, 2022.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007.

Francesco Laio and Stefania Tamea. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11:1267–1277, 2007.

Kenneth C. Lichtendahl, Yael Grushka-Cockayne, and Robert L. Winkler. Is it better to average probabilities or quantiles? *Management Science*, 59(7):1594–1611, 2013.

Roopesh Ranjan and Tilmann Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B*, 72(1):71–91, 2010.

Aaron Rumack, Ryan J. Tibshirani, and Roni Rosenfeld. Recalibrating probabilistic forecasts of epidemics. *PLOS Computational Biology*, 18(12):e1010771, 2022.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Ewart A. C. Thomas and Brian H. Ross. On appropriate procedures for combining probability within the same family. *Journal of Mathematical Psychology*, 21:136–152, 1980.