Supplementary Material for "Evaluating infectious disease forecasts with allocation scoring rules"

Aaron Gerding, Nicholas G. Reich, Benjamin Rogers, Evan L. Ray

December 21, 2023

Contents

1 Introduction

We address some technical and methodological points from the main text. We begin in section 2 by defining proper scoring rules and showing that the allocation score is proper. In section 3, we formalize the concept of a *shortage* of resources, give some key results about expected resource shortages under a distribution characterizing uncertainty about (future) levels of resource need, and illustrate how quantiles arise as the Bayes act in a decision making problem about the quantity of a resource to purchase. Section 4 gives a justification for the result that the Bayes act for the allocation problem is given by a vector of quantiles of the forecast distributions in each location at a shared probability level, which was stated in section **??** of the main text. We describe the algorithm that we use to compute allocations given a forecast distribution in each location in section 5. In section 6, we describe the methods for approximating a full distribution from a set of quantiles, implemented in the R package `distfromq`, that we used to support computation of allocation scores from quantile forecasts that were submitted to the COVID-19 Forecast Hub for the application in the main text. Finally, in section 7 we examine implications for propriety of an analysis that uses summaries of forecasts (such as predictive quantiles) rather than the full forecast distributions for computation of allocation scores.

2 Proper scoring rules

In decision theory, a loss function l is used to formalize a decision problem by assigning numerical value $l(x, y)$ to the *result* of taking an *action* x in preparation for an *outcome* y. A *scoring rule* S is a loss function for a decision problem where the action is a probabilistic forecast F of the outcome y (or the statement of F by a forecaster). We refer to the realized loss $S(F, y)$ as the *score* of F at y.

Probabilistic forecasts can be seen as a unique kind of action in that they can be used to generate their own (simulated) outcome data, against which they can be scored using S. A probabilistic forecast F is thus committed to the "self-assessment" $\mathbb{E}_F[S(F,Y)] := \mathbb{E}[S(F,Y^F)]$, where $Y^F \sim F$ is the random variable defined by sampling from F, as well to an assessment $\mathbb{E}_F[S(G,Y)]$ of any alternative forecast G.

A natural consistency criterion for S is that, for observations assumed to be drawn from F, it will not assess any other forecast G as being better than F itself, that is, that

$$\mathbb{E}_F[S(F,Y)] \leq \mathbb{E}_F[S(G,Y)] \tag{1}$$

for any F, G. A scoring rule meeting this criterion is called *proper*. If S were improper, then from the perspective of a forecaster focussed (solely) on expected loss minimization, the decision to state a forecast G other than the forecast F which they believe describes Y could be superior to the decision to state F. S is *strictly proper* when (1) is a strict inequality, in which case the *only* optimal decision for a forecaster seeking to minimize their expected loss is to state the forecast they believe to be true.

2.1 The allocation score is proper

Our primary decision theoretical procedure, outlined in section **??** of the main text, uses a decision problem with loss function $s(x,y)$ to define a scoring rule

$$S(F,y) := s(x^F, y) \tag{2}$$

where $x^F := \mathrm{argmin}_x \mathbb{E}_F[s(x,Y)]$ is the Bayes act for F with respect to s. Such scoring rules, which we call *Bayes scoring rules*, are proper by construction since

$$\begin{aligned} \mathbb{E}_F[S(F,Y)] &= \mathbb{E}_F[s(x^F,Y)] \\ &= \min_x \mathbb{E}_F[s(x,Y)] \quad \text{(by definition of } x^F) \tag{3} \\ &\leq \mathbb{E}_F[s(x^G,Y)] \tag{4} \\ &= \mathbb{E}_F[S(G,Y)]. \end{aligned}$$

The allocation scoring rule is Bayes and therefore proper.

We note that in the probabilistic forecasting literature (see e.g., **?**, Theorem 3) what we have termed Bayes scoring rules typically appear via (2) where x^F is some given functional of F which can be shown to be *elicitable*, that is, to be the Bayes act for some loss function s. Such a loss function is said to be a *consistent loss (or scoring) function* for the functional $F \mapsto x^F$, and many important recent results in the literature (e.g., **?**) address whether there *exists* any loss function that is consistent for x^F. Our orientation is different from this insofar as we *begin* by specifying a decision problem and a loss function of subject matter relevance and use the Bayes act only as a bridge to a proper scoring rule. Consistency is never in doubt.

3 Expected shortages

A key feature of loss functions used in the decision theoretic definition of quantiles and related scoring rules such as the CRPS and the WIS (see ...), as well as the allocation loss function presented in this work, is the presence of a *shortage*: the amount $\max\{0, y-x\}$ by which a resource demand y exceeds a supply decision variable x, which, for convenience, we write as $(y-x)_+$. In particular, a quantile at probability level α of a distribution F on $\mathbb{R}^1$ (which we assume to have a well-defined density $f(x)$) is a Bayes act for the loss function

$$l(x,y) = Cx + L(y-x)_+$$

where $\alpha = 1 - C/L$ and C and L can be interpreted as the cost per unit of a resource (such as medicine) and the loss incurred when a unit of demand (such as illness) cannot be met due to the

shortage $(y - x)_+$. This follows because a Bayes act, as a minimizer of $\mathbb{E}_F[l(x, Y)]$, must also be a vanishing point of the derivative

$$
\begin{aligned}
\frac{d}{dx}\mathbb{E}_F\left[l(x, Y)\right] &= \mathbb{E}_F\left[\frac{d}{dx}l(x, Y)\right] \\
&= C + L\,\mathbb{E}_F\left[\frac{d}{dx}(Y - x)_+\right] \\
&= C - L\,\mathbb{E}_F\left[\mathbf{1}\{Y > x\}\right] \\
&= C + L(F(x) - 1),
\end{aligned}
\tag{5}
$$

so that $1 - C/L = F(x)$. The formula $\frac{d}{dx}\mathbb{E}_F\left[(Y - x)_+\right] = F(x) - 1$ for the derivative of the shortage, used above in (5), can be obtained from an application of the "Leibniz Rule":

$$
\begin{aligned}
\frac{d}{dx}\mathbb{E}_F[(Y - x)_+] &= \frac{d}{dx}\int_x^\infty (y - x)f_Y(y)dy \\
&= \int_x^\infty \frac{d}{dx}(y - x)f_Y(y)dy - (x - x)f_Y(x) = -\int_x^\infty f_Y(y)dy = F(x) - 1.
\end{aligned}
\tag{6}
$$

Note that more care is required when F does not have a density. We will also use this result below in deriving the Bayes act for the allocation loss.

4   Allocation Bayes acts as vectors of marginal quantiles.

Here we study the form of the Bayes act for the allocation problem (AP) (equation (??) in section ??) of the text:

$$
\underset{0 \leq x}{\text{minimize}} \ \mathbb{E}_F[s_A(x, Y)] = \sum_{i=1}^N L \cdot \mathbb{E}_{F_i}[(Y_i - x_i)_+] \ \text{subject to} \ \sum_{i=1}^N x_i = K,
\tag{7}
$$

where the marginal forecasts F_i for $i = 1, \ldots, N$ represent forecasts for N distinct locations. We show that the Bayes act $x^{F,K} = (x_1^{F,K}, \ldots, x_N^{F,K})$ for a forecast F and resource constraint level K is a vector of quantiles of the marginal forecast distributions F_i at a single probability level $\tau^{F,K}$, that is, $x_i^{F,K} = q_{F_i, \tau^{F,K}}$. An immediate consequence used in the examples in Section ?? in the main text is that if $F_i = \text{Exp}(1/\sigma_i)$ for all i, then the Bayes act is proportional to $(\sigma_1, \ldots, \sigma_N)$, since $q_{\text{Exp}(1/\sigma), \tau} = -\sigma \log(1 - \tau)$.

In order for $x^\star \in \mathbb{R}_+^N$ to solve the AP it must be true that reallocating $\delta > 0$ units of the resource from location i to location j will lead to a net increase in expected shortage — in other words, the reallocation increases the expected shortage in location i is at least as much as it decreases the expected shortage in location j:

$$
\begin{aligned}
\mathbb{E}_{F_i}[(Y_i - (x_i^\star - \delta))_+] &- \mathbb{E}_{F_i}[(Y_i - x_i^\star)_+] \ \text{(increase in } i) \\
&\geq \mathbb{E}_{F_j}[(Y_j - x_j^\star)_+] - \mathbb{E}_{F_j}[(Y_j - (x_j^\star + \delta))_+] \ \text{(decrease in } j) \ .
\end{aligned}
$$

Dividing by δ and letting $\delta \searrow 0$, this implies from (6) that

$$
\begin{aligned}
1 - F_i(x_i^\star) &= -\frac{d}{dx_i}\mathbb{E}_{F_i}[(Y_i - x_i^\star)_+] \\
&= \lim_{\delta \searrow 0}\frac{1}{\delta}\left\{\mathbb{E}_{F_i}[(Y_i - (x_i^\star - \delta))_+] - \mathbb{E}_{F_i}[(Y_i - x_i^\star)_+]\right\} \ \text{(increase in } i) \\
&\geq \lim_{\delta \searrow 0}\frac{1}{\delta}\left\{\mathbb{E}_{F_j}[(Y_j - x_j^\star)_+] - \mathbb{E}_{F_j}[(Y_j - (x_j^\star + \delta))_+]\right\} \ \text{(decrease in } j) \\
&= -\frac{d}{dx_j}\mathbb{E}_{F_j}[(Y_j - x_j^\star)_+] = 1 - F_j(x_j^\star)
\end{aligned}
\tag{8}
$$

Note that negative derivatives appear because our optimality condition addresses how a *decrease* in resources will *increase* the expected shortage in i and vice versa in j. Since (8) also holds with i and j

reversed, a number λ (a *Lagrange multiplier*) exists such that $L(1 - F_k(x_k^\star)) = \lambda$ for all $k \in 1, \ldots, N$. (We scale by L to facilitate possible future interpretations of λ in terms of the partial derivatives of $\mathbb{E}_F[s_A(x, Y)]$.) That is, $x_k^\star$ is a quantile q_{τ, F_k} for $\tau = 1 - \lambda/L$. The value of τ is then determined by the constraint equation

$$\sum_{i=1}^{N} q_{\tau, F_i} = K. \tag{9}$$

It is important to note that τ depends on F and K and is *not* a fixed parameter of the allocation scoring rule.

5 Numerical computation of allocation Bayes acts

To compute an allocation score $S_A(F, y; K) := s_A(x^{F,K}, y)$, we require numerical values for a Bayes act solving the AP (7) — that is, we must find the specific resource allocations for each location that are determined by the forecast F under the resource constraint K. Assuming we have reliable means of calculating quantiles q_{α, F_i} of the marginal forecasts F_i, these allocations are given by $q_{\tau^\star, F_i}$ where $\tau^\star$ solves the equation (9). However, this equation is not analytically tractable and we must resort to a numerical method for finding an approximation $\tilde{\tau}$ of $\tau^\star$.

We have implemented an iterative bisection method that makes use of the fact that $\sum_{i=1}^{N} q_{\tau, F_i}$ is an increasing function of τ. The algorithm begins with an initial search interval $[\tau_{L,1}, \tau_{U,1}]$ (such as $[0, \max_i F_i(K)]$) that clearly contains the solution $\tau^\star$. At each step j of the algorithm, we evaluate the total allocation $\sum_{i=1}^{N} q_{\tau_{M,j}, F_i}$ at the midpoint of the search interval, $\tau_{M,j} = \frac{1}{2}(\tau_{L,j} + \tau_{U,j})$ and continue the search on the narrowed sub-interval

$$[\tau_{L,j+1}, \tau_{U,j+1}] = \begin{cases} [\tau_{L,j}, \tau_{M,j}] & \text{if } \sum_{i=1}^{N} q_{\tau_{M,j}, F_i} \geq K \\ [\tau_{M,j}, \tau_{U,j}] & \text{if } \sum_{i=1}^{N} q_{\tau_{M,j}, F_i} < K. \end{cases}$$

This search continues until $\tau_{U,j+1} < (1 + \varepsilon)\tau_{L,j+1}$ for a suitably small $\varepsilon > 0$. We have implemented this procedure along with the resulting score computations in the R package `alloscore` which provided all allocation score values used in the analysis of section **??** in the main text.

Subtleties can arise when the forecast densities f_i vanish or are very small, in which case quantiles are non-unique or highly variable near a probability level, leading to ambiguity or numerical instabilities in the evaluation of $\sum_{i=1}^{N} q_{\tau, F_i}$. Additionally, if point masses are present in any of the F_i, (9) will not have a unique solution for some discrete set of constraint levels K. We have adopted conventions for detecting such levels and enforcing consistency in score calculations near them. Through extensive experimentation, we have determined that these conditions seem to address these challenges with the forecasts we are working with, but we leave a more rigorous approximation error analysis for later work.

6 Computing allocations from finite quantile forecast representations

In section **??** of the manuscript, we used the allocation score to evaluate forecasts of COVID-19 hospitalizations that have been submitted to the US COVID-19 Forecast Hub. These forecasts are submitted to the Hub using a set of 23 quantiles of the forecast distribution at the 23 probability levels in the set $\mathcal{T} = \{0.01, 0.025, 0.05, 0.1, 0.15, \ldots, 0.9, 0.95, 0.975, 0.99\}$, which specify a predictive median and the endpoints of central $(1 - \alpha) \times 100\%$ prediction intervals at levels $\alpha = 0.02, 0.05, 0.1, 0.2,$ $0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. For a given week and target date, we use $q_{i,k}$ to denote the submitted quantiles for location i and probability level $\tau_k \in \mathcal{T}$, $k = 1, \ldots, 23$.

In the event that there is some $k \in \{1, \ldots, 23\}$ for which $\sum_i q_{i,k} = K$, i.e., the provided predictive quantiles at level τ_k sum across locations to the resource constraint K, the solution to the allocation problem is given by those quantiles. However, generally this will not be the case; the optimal allocation will typically be at some probability level $\tau^\star \notin \mathcal{T}$.

To address this situation and support the numerical allocation algorithm outlined in section 5, we need a mechanism to approximate the full cumulative distribution functions F_i, $i = 1, \ldots, N$ based on the provided quantiles. We have developed functionality for this purpose in the `distfromq` package for R.[**?**] This functionality represents a distribution as a mixture of discrete and continuous parts, and it works in two steps:

1. Identify a discrete component of the distribution consisting of zero or more point masses, and create an adjusted set of predictive quantiles for the continuous part of the distribution by subtracting the point mass probabilities and rescaling.

2. For the continuous part of the distribution, different approaches are used on the interior and exterior of the provided quantiles:

   (a) On the interior, a monotonic cubic spline interpolates the adjusted quantiles representing the continuous part of the distribution.

   (b) A location-scale parametric family is used to extrapolate beyond the provided quantiles. The location and scale parameters are estimated separately for the lower and upper tails so as to obtain a tail distribution that matches the two most extreme quantiles in each tail. In this work, we use normal distributions for the tails.

The resulting distributional estimate exactly matches all of the predictive quantiles provided by the forecaster. We use the cumulative distribution function resulting from this procedure as an input to the allocation score algorithm.

We refer the reader to the `distfromq` documentation for further detail.[**?**]

7 Propriety of parametric approximation

In practice, open forecasting exercises are generally not able to collect a perfect description of the forecast distribution F other than in simple settings such as for a categorical variable with a relatively small number of categories. In settings where the outcome being forecasted is a continuous quantity (such as the proportion of outpatient doctor visits where the patient has influenza-like illness) or a count (such as influenza hospitalizations), forecasting exercises have therefore resorted to collecting summaries of a forecast distribution such as bin probabilities or predictive quantiles. In this section, we address two practical concerns raised by this. First, we discuss conditions under which it is possible to calculate the allocation score when only summaries of a forecast distribution are recorded in a submission to a forecast hub. Second, we show that a post hoc attempt to compute the allocation score based on submitted predictive quantiles may in fact compute an alternative score that is not proper.

7.1 Propriety when scoring methods are announced prospectively

We consider a setting where a forecasting exercise (such as a forecast hub) pre-specifies that forecasts will be represented using a parametric family of forecast distributions $G_\theta(y)$, and the task of the forecaster is to select a particular parameter value θ. We use $\mathcal{P}$ to denote the collection of all distributions G_θ in the given parametric family. For instance, it has recently been proposed that mixture distributions could be used to represent forecast distributions **?**. Additionally, we note that the functionality in `distfromq` can be viewed as specifying a parametric family $\mathcal{P}_{\text{dfq}}$ where the parameters θ of G_θ are its quantiles at pre-specified probability levels, and where the shape of any $G_\theta \in \mathcal{P}_{\text{dfq}}$ over the full range of its support is entirely controlled by these quantiles.

We find it helpful now to formally distinguish between two decision making problems. The first is the public health decision maker's allocation problem where the task is to select an allocation x, with the allocation loss $s_A(x, y) = \sum_{i=1}^{N} L \cdot \max(0, y_i - x_i)$ as described in section **??** of the main text. The second is the forecaster's reporting problem where the task is to select parameter values θ to report. The forecaster's loss is given by $s_R(\theta, y) = s_A(x^{G_\theta}, y)$, where x^{G_θ} is the Bayes act for the allocation problem under the distribution G_θ. In words, the loss associated with reporting θ is equal to the loss associated with taking the Bayes allocation corresponding to the distribution G_θ.

Following our usual construction, the Bayes act for the forecast reporting problem is the parameter set that minimizes the forecaster's expected loss. Breaking with our earlier notation for improved legibility, we use $\theta^\star(F)$ to denote this Bayes act:

$$\theta^\star(F) = \operatorname{argmin}_\theta \mathbb{E}_F[s_R(\theta, Y)]$$
$$= \operatorname{argmin}_\theta \mathbb{E}_F[s_A(x^{G_\theta}, Y)]$$

We then arrive at the scoring rule

$$S_R(F, y) = s_R(\theta^\star(F), y) = s_A(x^{G_{\theta^\star(F)}}, y).$$

It follows from the discussion in section 2.1 that this is a proper scoring rule for F. Although the full forecast distribution F is not available in the forecast submission, the score $S_R(F, y)$ can be calculated from the reported parameter values as long as the forecaster submits the optimal parameters $\theta^\star(F)$.

We emphasize that the forecaster's true predictive distribution F does not need to be a member of the specified parametric family $\mathcal{P}$ for this construction to yield a proper score. It is, however, necessary to specify the parametric family to use and the foundational scoring rule s_A (including any relevant problem parameters such as the resource constraint K) in advance, so that forecasters can identify the Bayes act parameter set $\theta^\star(F)$ to report.

If the parametric family used to represent forecast distributions is flexible enough, the reporting scoring rule S_R and the allocation score are equivalent in the sense that they will yield the same score for any distribution F. Suppose that for a given resource constraint K, for any forecast distribution F it is possible to find a member $G_{\theta^\star}$ of the specified parametric family $\mathcal{P}$ with the same allocation as F (i.e., $x^F = x^{G_{\theta^\star}}$). Then $\theta^\star$ is a Bayes act for the reporting problem since for any other parameter value θ,

$$\mathbb{E}_F[s_R(\theta^\star, Y)] = \mathbb{E}_F[s_A(x^{G_{\theta^\star}}, Y)]$$
$$= \mathbb{E}_F[s_A(x^F, Y)] \quad (\text{since } x^F = x^{G_{\theta^\star}})$$
$$\leq \mathbb{E}_F[s_A(x^{G_\theta}, Y)] \quad (\text{by definition of } x^F)$$
$$= \mathbb{E}_F[s_R(\theta, Y)].$$

It therefore follows that

$$S_R(F, y) = s_R(\theta^\star, y)$$
$$= s_A(x^{G_{\theta^\star}}, y)$$
$$= s_A(x^F, y)$$
$$= S_A(F, y).$$

For the particular choice of the parametric family $\mathcal{P}_{\text{dfq}}$ (i.e., using the `distfromq` package), this flexibility requirement is satisfied. For instance, the forecaster could pick one required quantile level (such as 0.5, for which the corresponding predictions are predictive medians), and set the submitted quantiles of their forecast distribution in each location at that level to be the desired allocations, which sum to K across all locations. However, this representation of the forecast may be quite different from the actual forecast distribution F. For example, for the actual forecast distribution F the allocations may occur at some quantile level other than 0.5.

As another alternative for practical forecasting exercises, a forecast hub could ask forecasters to directly provide the Bayes allocations associated with their forecasts for one or more specified resource constraints K. At the cost of increasing the number of quantities solicited by the forecast hub, this would have several advantages: it would prevent any artificial distortion of the forecast distributions, allow for direct calculation of scores, and narrow the gap between model outputs and public health end users. For this to be feasible, implementations of the allocation algorithm would have to be provided to participating forecasters in the computational languages being used for modeling.

7.2   Impropriety of post hoc allocation scoring with quantile forecasts

A *post hoc* evaluation of quantile forecasts that combines the parametric family specified by `distfromq` with the allocation score does not yield the allocation score of the forecast distribution F. Instead, it computes an alternative score that is improper. This is because the forecast distribution F and the distribution $G^q \in \mathcal{P}_{\mathrm{dfq}}$ with the same quantiles as F may determine different resource allocations. In our investigations, these discrepancies appear to be relatively minor on the interior of the provided quantiles, but could be severe if the tail extrapolations performed by `distfromq` do not match the tail behavior of F and the allocations are in the tails of the predictive distribution.

We define

$$G^\star(F) := \operatorname*{argmin}_{G \in \mathcal{P}_{\mathrm{dfq}}} E_F[S_A(G, Y)],$$

the distribution corresponding to the optimal submission for the forecaster with predictive distribution F. Because the distribution $G^q(F)$ typically will not yield allocations that are optimal under the distribution F, in general $G^q(F)$ and $G^\star(F)$ are different distributions.

When an analyst attempts a post hoc computation of the allocation score using $G^q(F)$ (implicitly assuming that $G^q(F) = G^\star(F)$), they in fact compute the alternative score

$$\tilde{S}(F, y) = S_A(G^q(F), y) = s_A(x^{G^q(F)}, y).$$

This score is improper because $E_F[S_A(G, Y)]$ is minimized by $G^\star(F)$, not $G^q(F)$. In general, we have

$$E_F[\tilde{S}(G^\star(F), Y)] \leq E_F[S_A(G^q(F), Y)] \tag{10}$$
$$= E_F[\tilde{S}(F, Y)]$$

However, the inequality in (10) will typically be strict. For example, if F has heavy upper tails (such as for a lognormal distribution), but normal distributions are used for tail extrapolations in `distfromq`, then the resource allocations based on the distribution $G^q(F)$ may be quite different from the optimal allocations under the distribution F, leading to a strict inequality. This demonstrates that $\tilde{S}$ is improper.