Allocation scoring rules as a generalization of quantile scoring rules

2022-11-28

Quantile Scoring Rules

A forecaster is asked to recommend an appropriate supply x of some good or investment in some precautionary measure intended to satisfy a random future demand or need Y. Suppose there is an incremental loss $O \geq 0$ incurred when overprediction leads to unused supply and an incremental loss $U > 0$ incurred when underprediction leads to unmet demand or need. Let g be a non-decreasing function that expresses a utility associated with the good or misfortune to which Y refers.

By rewarding the forecaster for their perfomance according to the (negatively oriented) scoring function

$$\begin{aligned} s(x,y) = s_{O,U}(x,y) &= O(g(x)-g(y))_+ + U(g(x)-g(y))_-. && (1)\\ &= \kappa((1-\alpha)(g(x)-g(y))_+ + \alpha(g(x)-g(y))_-) && (2)\\ &= \kappa(\mathbb{1}\{x>y\}-\alpha)(g(x)-g(y)) && (3)\\ &:= s_{\kappa,\alpha}(x,y) := \kappa s_\alpha(x,y) && (4)\end{aligned}$$

we elicit the $\alpha = U/(U+O)$ quantile Q of the distribution $F_Y = F$ of Y, where $u_+ := \max(u,0)$, $u_- := \max(-u,0) = (-u)_+$, and $\mathbb{1}\{A\}$ is the indicator function of the event A.

That is, if a forecaster believes that Y has cdf F and density f, then according to this belief they minimize their expected score

$$\begin{aligned} Z_F(x) = Z(x) = E_F[s(x,Y)] &= OE_F[(g(x)-g(Y))_+] + UE_F[(g(x)-g(Y))_-] && (5)\\ &= O\int_{-\infty}^{x}(g(x)-g(y))f(y)dy + U\int_{x}^{\infty}(g(y)-g(x))f(y)dy && (6)\end{aligned}$$

by forecasting the quantile $Q = F^{-1}(\alpha)$ which solves the first order equation

$$\begin{aligned} 0 = \frac{dZ}{dx}(x) &= g'(x)\left(OF(x) - U(1-F(x))\right)\\ &= g'(x)(O+U)\left(F(x) - \frac{U}{O+U}\right) && (7)\\ &= \kappa g'(x)\left(F(x)-\alpha\right) && (8)\end{aligned}$$

and is actually a minimum when $g'(Q) > 0$, which makes

$$\begin{aligned} \frac{d^2Z}{dx^2}(x)\big|_{x=Q} &= \kappa\left(g'(x)f(x) + g''(x)\left(F(x)-\alpha\right)\right)\big|_{x=Q} && (9)\\ &= \kappa(g'(Q)f(Q) + g''(Q)\cdot 0) > 0. && (10)\end{aligned}$$

Note that $\frac{dZ}{dx}(x) = \frac{d}{dx}E_F[s(x,Y)] = h(x)E_F[V(x,Y)]$, where $V(x,y) := \mathbb{1}\{x > y\} - \alpha$. By virtue of $E_F[V(Q_F(\alpha),Y)] = 0$, V is said to be an *identification function* for the α quantile. The fact that, generally speaking, any elicitable functional (such as a quantile) has an identification function is known as *Osband's principle.*

Forecasting the minimizer of $Z(x) = E_F[s_{\kappa,\alpha}(x,Y)]$ in this situation is known as the *Bayes act* a_F under s_α for the forecaster. Assuming a forecaster is informed, rational, and risk-neutral enough to take a_F, we can evaluate F, implicitly, as a distributional forecast, by the *scoring rule* S_α induced by s_α:

$$S_\alpha(F, y) := s_\alpha(Q(\alpha), y). \tag{11}$$

If y is generated by $Y \sim G$ then $S_\alpha(F, y)$ is an estimate of

$$S_\alpha(F, G) := E_G[S_\alpha(F, Y)] = E_G[s_\alpha(Q_F(\alpha), Y)]. \tag{12}$$

This defines a functional on $\mathcal{F}_0$, the class of measures on the Borel-Lebesgue sets of $\mathbb{R}$ by

$$\mathcal{Z}_{G,\alpha}(F) := Z_G \circ Q_\alpha(F) \tag{13}$$

where Q_α is the α quantile functional.

If $g'(Q) = 0$ or does not exist, we'll need to adopt conventions as to whether Q is optimal, similarly to how the lowest x with $F(x) \geq \alpha$ is chosen as $Q(\alpha)$ when F is constant on an interval (and $g(x) = x$). One example of such a g is the *power curve* of a wind turbine which becomes constant above a certain wind speed. Here x would be a forecast of future wind speed Y at time t and $g(x)$ would be the power a turbine operator commits to supplying the electric grid at t.

Taking g to be constant below a certain point could also serve to express the futility or lack of meaning of forecasts below that point, such as when a forecast user is only able to sell a commodity for which demand can become negative, turning into supply.

Using $u_- = u_+ - u$, the objective function (5) can also be helpfully rewritten as

$$\begin{aligned}
Z(x) &= OE_F[(g(x) - g(Y))_+] + UE_F[(g(x) - g(Y))_+] - UE_F[(g(x) - g(Y))] \\
&= (O + U)E_F[(g(x) - g(Y))_+] - UE_F[(g(x) - g(Y))] \\
&= (O + U)\left(g(x)F(x) - \overline{g}_F(x)\right) - U(g(x) - \overline{g}_F) \qquad (14) \\
&= \kappa\left[g(x)(F(x) - \alpha) - \overline{g}_F(x) + \alpha\overline{g}_F\right] \qquad (15)
\end{aligned}$$

where $\overline{g}_F(x) = E_F[g(Y)\mathbb{1}\{Y \leq x\}]$, and $\overline{g}_F = E_F[g(Y)]$.

The function $\overline{g}_F(x)$ sometimes goes by the name of "partial expectation" (Schlaifer Raiffa, p. 109) and can often be made more explicit for certain F and g.

Examples for various forecaster beliefs

1) $Y \sim U[a, b]$

Here $F(x) = \mathbb{1}\{a < x < b\}(x - a)/(b - a) + \mathbb{1}\{b \leq x\}$ in (15) gives

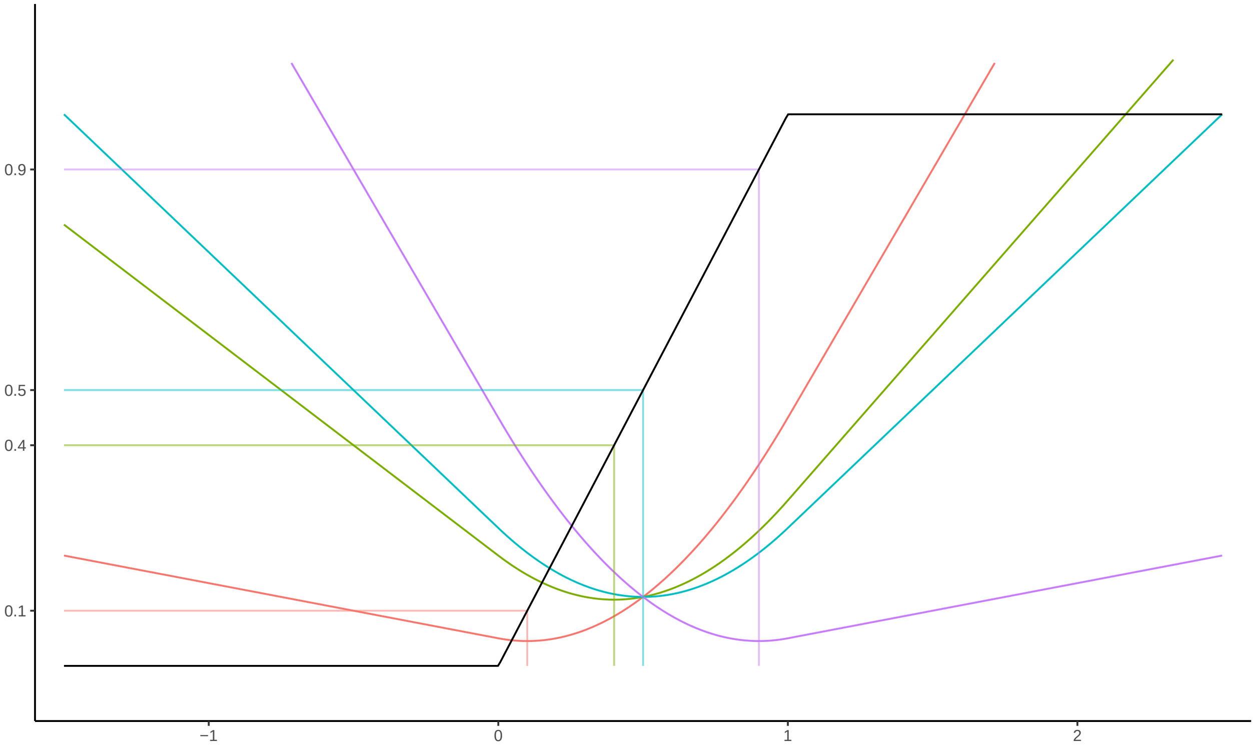
$$Z(x) = \begin{cases} \frac{\kappa}{b-a}\left(-\alpha(b-a)g(x) + \alpha\int_a^b g(y)dy\right) & x \leq a \\ \frac{\kappa}{b-a}\left((x - a - \alpha(b-a))g(x) - \int_a^x g(y)dy + \alpha\int_a^b g(y)dy\right) & a < x < b \\ \frac{\kappa}{b-a}\left((1-\alpha)(b-a)g(x) - (1-\alpha)\int_a^b g(y)dy\right) & b \leq x \end{cases} \tag{16}$$

When $g(x) = x$ this becomes

$$Z(x) = \begin{cases} \frac{\kappa}{b-a}\left(-\alpha(b-a)x + \frac{\alpha}{2}(b^2 - a^2)\right) & x \leq a \\ \frac{\kappa}{b-a}\left((x - a - \alpha(b-a))x - \frac{1}{2}(x^2 - a^2) + \frac{\alpha}{2}(b^2 - a^2)\right) & a < x < b \\ \frac{\kappa}{b-a}\left((1-\alpha)(b-a)x - \frac{(1-\alpha)}{2}(b^2 - a^2)\right) & b \leq x \end{cases} \tag{17}$$

$$= \begin{cases} -\kappa\alpha\left[x - \frac{a+b}{2}\right] & x \leq a \\ \frac{\kappa}{2}\left[\frac{1}{b-a}(x - (a + \alpha(b-a)))^2 + \alpha(1-\alpha)(b-a)\right] & a < x < b \\ \kappa(1-\alpha)\left[x - \frac{a+b}{2}\right] & b \leq x \end{cases} \tag{18}$$

Thus $Z(x)$ has the form of pinball loss centered on the mean $\mu_Y = \frac{a+b}{2}$ and interpolated on (a, b) by a quadratic centered at the quantile $a + \alpha(b - a)$ with slopes matching the linear tails at a and b.



2) $g(x) = x$ **and** $F_Y = F_{\mu,\sigma}$ **lies in a location-scale family**

That is, $Y = \mu + \sigma Z$ and $Z \sim F_Z = F_0$. Then $f_Y(y) = \frac{d}{dz} F_Z\left(\frac{y-\mu}{\sigma}\right) \frac{dz}{dy} = f_Z(z)/\sigma$ and so

$$\overline{g}_F(x) = \int_a^x y f_Y(y) dy = \int_{\frac{a-\mu}{\sigma}}^{\frac{x-\mu}{\sigma}} (\mu + \sigma z) f_Z(z) dz \tag{19}$$

$$= \mu F_Z\left(\frac{x-\mu}{\sigma}\right) + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{x-\mu}{\sigma}} z f_Z(z) dz \tag{20}$$

$$= \mu F_Y(x) + \sigma \mu_{F_0}\left(\frac{x-\mu}{\sigma}\right). \tag{21}$$

where $\mu_{F_0}(u) = \int_{\frac{a-\mu}{\sigma}}^u z f_0(z) dz$. This gives

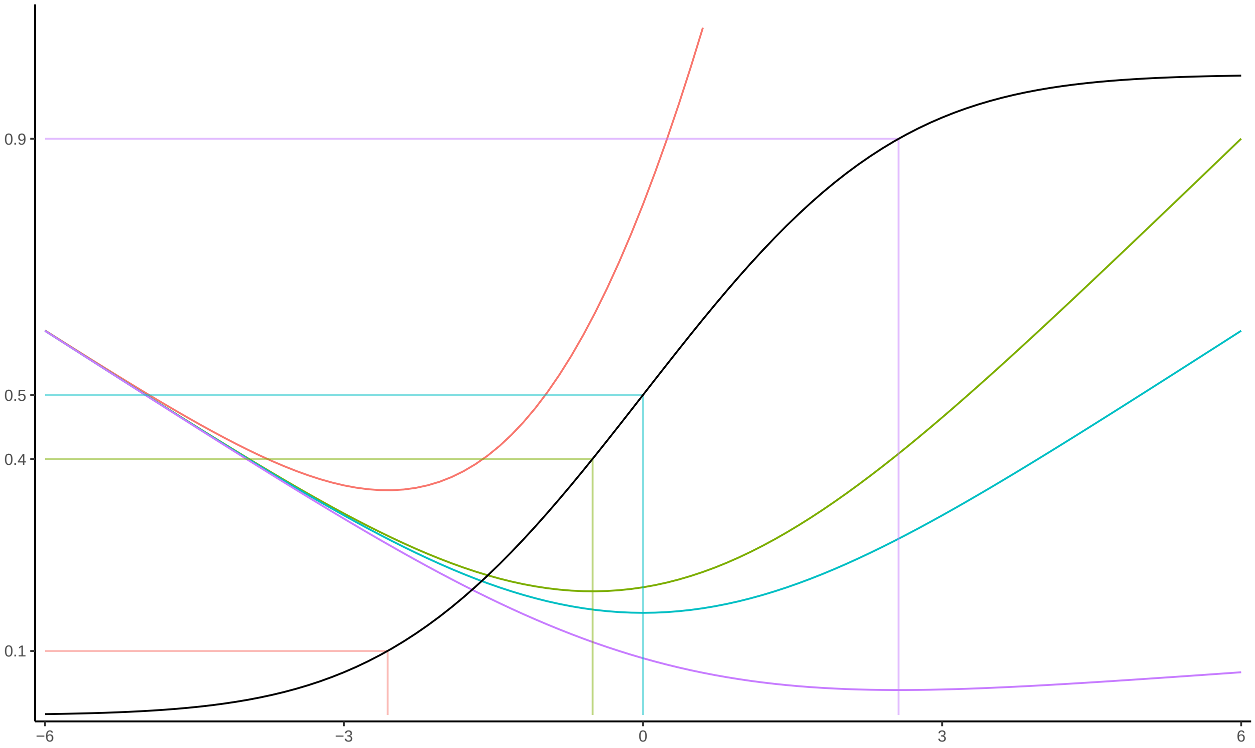
$$\begin{aligned} Z(x) &= \kappa \left[x(F(x) - \alpha) - \mu F(x) - \sigma \mu_{F_0}\left(\frac{x-\mu}{\sigma}\right) + \alpha\mu \right] \\ &= \kappa \left[(F(x) - \alpha)(x - \mu) - \sigma \mu_{F_0}\left(\frac{x-\mu}{\sigma}\right) \right] \end{aligned} \tag{22}$$

3) $g(x) = x$ **and** $Y \sim N(\mu, \sigma)$

When $F = \Phi_{\mu,\sigma}$, $\mu_{F_0}\left(\frac{x-\mu}{\sigma}\right) = -\sigma \varphi_{\mu,\sigma}(x)$, giving $q(\alpha)$ as the minimum of

$$Z(x) = \kappa \left[(\Phi_{\mu,\sigma}(x) - \alpha)(x - \mu) + \sigma^2 \varphi_{\mu,\sigma}(x) \right].$$

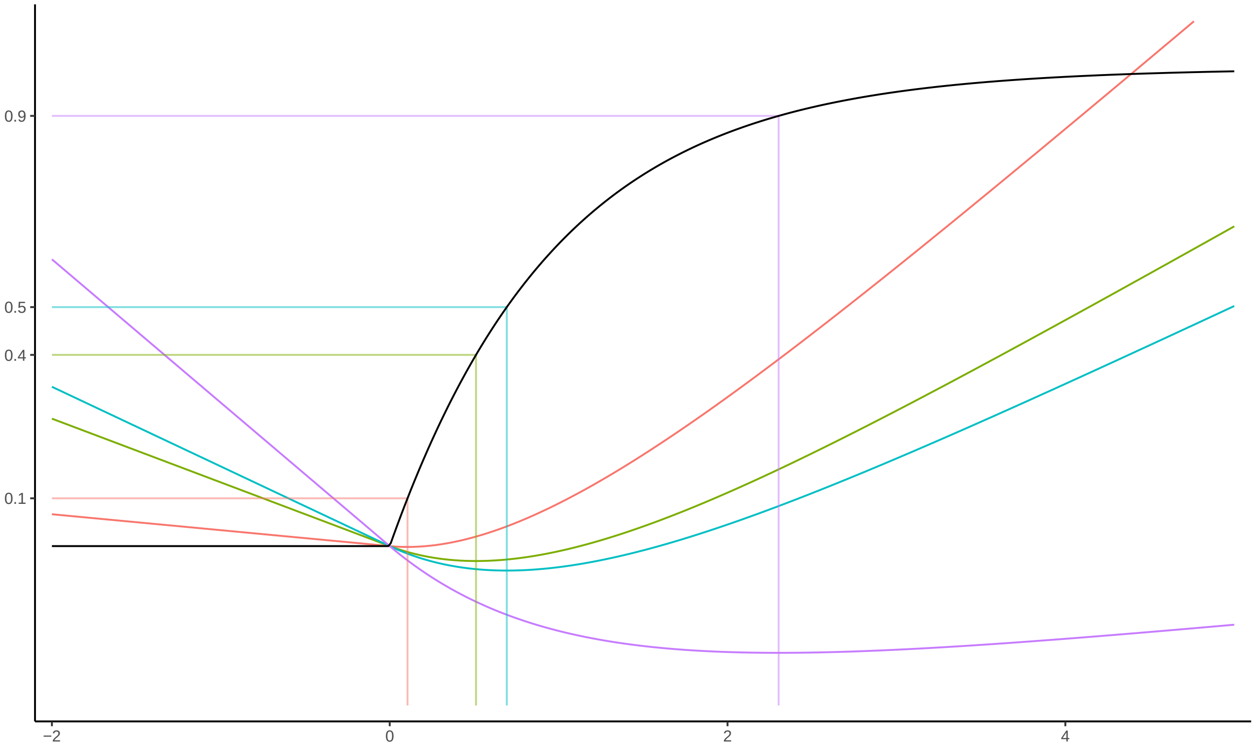
Plot of several such functions when $\mu = 0, \sigma = 2$ along with $\Phi_{0,2}(x)$:



4) $g(x) = x$ **and** $Y \sim \text{Exp}(1/\sigma)$

Taking $\mu \equiv 0, Y = \sigma Z$ and $F_Z(x) = F_0(x) = \mathbb{1}\{0 \le x\}(1 - e^{-x})$ we have $\mu_{F_0}(x) = \mathbb{1}\{0 \le x\}(1 - e^{-x}(1 + x))$. Then (22) is

$$\begin{aligned} Z(x) &= \mathbb{1}\{0 \le x\}\kappa \left[F(x)x - \sigma\mu_{F_0}\left(\frac{x}{\sigma}\right)\right] - \kappa\alpha x \\ &= \mathbb{1}\{0 \le x\}\kappa \left[\left(1 - e^{-x/\sigma}\right)x - \sigma\left(1 - e^{-x/\sigma}\left(1 + \frac{x}{\sigma}\right)\right)\right] - \kappa\alpha x \qquad (23) \\ &= \mathbb{1}\{0 \le x\}\kappa \left[\sigma e^{-x/\sigma} + x - \sigma\right] - \kappa\alpha x \qquad (24) \end{aligned}$$



Another form for $Z(x)$

There also may be situations where it is convenient to integrate the $(O + U)$-term in (14) by parts to get

$$Z(x) = \kappa \left[\int_a^x g'(y) F(y) dy - \alpha(g(x) - \overline{g}_F) \right]. \tag{25}$$

Newsvendor parameters

In the inventory management literature, minimization of an expectation of the form of $Z(x)$ for a given F arises in the *newsvendor problem*. This involves the ordering decision faced by a retailer of a perishable good (such as newpapers) when customer demand is uncertain. Because the focus here is on revenues and expenditures, terms other than just the over- and underprediction penalties in (1) often appear, giving a retailer's scoring function as

$$s_n(x, y) = a_x x + a_y y + a_+(x - y)_+ + a_-(x - y)_-. \tag{26}$$

We might have, for example, the score given by net value transfers

$$s_n(x, y) = cx - py - u(x - y)_+ + (p + r)(x - y)_- \tag{27}$$

where x newspapers are ordered at price c, $y-(x-y)_-$ are sold at retail price p, $(x-y)_+$ are sold at salvage price u, and $r(x - y)_-$ of customer "goodwill" is lost. This can be rewritten as

$$\begin{aligned} s_n(x, y) &= (a_+ + a_x)(x - y)_+ + (a_- - a_x)(x - y)_- + (a_x + a_y)y \qquad (28) \\ &= (c - u)(x - y)_+ + (p + r - c)(x - y)_- - (p - c)y \qquad (29) \end{aligned}$$

giving the retailer's objective function under a demand distribution F as

$$Z_n(x) = (a_+ + a_x)E_F[(x - Y)_+] + (a_- - a_x)E_F[(x - Y)_-] + (a_x + a_y)E[Y]. \tag{30}$$

This differs by a constant in x from our $Z(x)$ in (5) after taking $g(x) = x$, $O = a_+ + a_x$ and $U = a_- - a_x$, and so has the same minimizer

$$Q = F^{-1}\left(\frac{U}{O+U}\right) = F^{-1}\left(\frac{a_- - a_x}{a_- + a_+}\right) = F^{-1}\left(\frac{p+r-c}{p+r-u}\right). \tag{31}$$

Meteorologist parameters

Similarly, a weather forecaster might be judged by the cost Cx of recommended protection x against a level y of adverse weather (e.g., rainfall) added to any loss $L(x-y)_-$ resulting from underprediction, leading to the scoring function

$$s_m(x,y) = Cx + L(x-y)_-, \tag{32}$$

which rearranges to

$$s_m(x,y) = C(x-y)_+ + (L-C)(x-y)_- + Cy. \tag{33}$$

The minimizer of $Z_m(x) = E_F[s_m(x,Y)]$ given the forecaster's belief $Y \sim F$ is now

$$Q = F^{-1}\left(\frac{L-C}{C+L-C}\right) = F^{-1}\left(1 - \frac{C}{L}\right). \tag{34}$$

In particular, faced with the classical binary decision problem of whether to recommend an additional unit of protection given a current level of protection x, the forecaster's optimal decision rule under s_m is to recommend adding protection if $x < F^{-1}\left(1 - \frac{C}{L}\right)$, that is, if

$$1 - F(x) = \mathbb{P}_F\{y > x\} > \frac{C}{L}, \tag{35}$$

the *cost-loss ratio* of the problem.

Multipoint Quantile Scoring Rules

Suppose we ask a forecaster to provide a set of point forecasts $\{x_1, x_2\}$ to be rewarded according to the aggregate scoring function

$$s(x_1, x_2, y) = s_{O_1,U_1}(x_1, y) + s_{O_2,U_2}(x_2, y). \tag{36}$$

Allocation scoring rules

We now develop *allocation scoring functions* that elicit forecasts x_i for outcomes $y_i, i = 1, \dots, N$, each with it's own incremental cost O_i and loss U_i as for a quantile scoring rule, but with the additional constraint

$$\sum_{i=1}^{N} w_i x_i = \mathbf{w}^T \mathbf{x} \leq K \tag{37}$$

on the total provision available with $w_i > 0$. We assume $K > 0$, that only non-negative forecasts x_i, i.e., recommended allocations, are accepted, and that

$$g_i'(x_i) > 0 \text{ for } x_i \geq 0, \tag{38}$$

though it may be interesting to introduce regions where $g_i(x_i)$ is constant, such as for the wind speed power curve mentioned above.

According to (6) and (15) the objective function is now

$$Z(\mathbf{x}) = \sum Z_i(x_i) = \sum \left\{ O_i \int_{-\infty}^{x_i} (g_i(x_i) - g_i(y)) f_i(y) dy + U_i \int_{x_i}^{\infty} (g_i(y) - g_i(x_i)) f_i(y) dy \right\} \tag{39}$$

$$= \sum \kappa_i \left[g_i(x_i)(F_i(x_i) - \alpha_i) - \overline{g}_{i_{F_i}}(x_i) + \alpha_i \overline{g}_{i_{F_i}} \right] \tag{40}$$

and the elicited forecasts solve the allocation problem (AP)

$$\underset{\mathbf{x} \in \mathbb{R}^N}{\text{minimize}}\ Z(\mathbf{x}) \text{ subject to } \mathbf{x} \in C_K = \left\{ \mathbf{x} \mid D\mathbf{x} := \begin{bmatrix} \mathbf{w}^T \\ -\mathrm{Id}_N \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} K \\ 0 \end{bmatrix} \right\}. \tag{41}$$

To describe such forecasts, we begin by stating the necessary Karush-Kuhn–Tucker (KKT) conditions for a point $\mathbf{Q} \in C_K$ to be a solution of the AP. Since the feasible set C_K is a polyhedron on which Z is smooth, these are: there exists a vector of multipliers $\boldsymbol{\mu} = (\lambda, \mu_1, \ldots, \mu_N) \in \mathbb{R}^{1+N}$ such that

$$\lambda, \mu_1, \ldots, \mu_N \geq 0 \tag{42}$$

$$-\nabla Z(\mathbf{Q}) = D^T \boldsymbol{\mu} \tag{43}$$

$$\lambda(\mathbf{w}^T \mathbf{Q} - K) = 0 \tag{44}$$

$$\mu_i Q_i = 0, \quad i = 1, \ldots, N \tag{45}$$

(see for example, (Nocedal and Wright 2006), Theorem 12.1). These equations express the fact that at a minimizer, either the gradient ∇Z must vanish (so that $\boldsymbol{\mu} = \mathbf{0}$), or the direction of greatest descent $-\nabla Z$ must lie in the normal (outwardly pointing) cone of C_K, which is the non-negative span of some subset (generically a single one) of the columns of D^T. This subset corresponds to which constraints (i.e. “sides” of C_K) are active, allowing the corresponding multipliers in the “complementary slackness” equations (44) and (45) to be positive.

From (8), the “Lagrange multiplier” equation (43) becomes

$$-\sum_{i=1}^{N} \kappa_i g_i'(Q_i)(F_i(Q_i) - \alpha_i)\mathbf{e}_i = [\mathbf{w} | -\mathrm{Id}_N]\boldsymbol{\mu} = \sum_{i=1}^{N} (\lambda w_i - \mu_i)\mathbf{e}_i \tag{46}$$

which decomposes to

$$\lambda = w_i^{-1}(\kappa_i g_i'(Q_i)(\alpha_i - F_i(Q_i)) + \mu_i), \quad i = 1, \ldots, N. \tag{47}$$

Using (45), we can write this as

$$\lambda = \begin{cases} w_i^{-1} \kappa_i g_i'(Q_i)(\alpha_i - F_i(Q_i)) & \text{for } Q_i > 0 \quad (48) \\ w_i^{-1}(\kappa_i g_i'(0)(\alpha_i - F_i(0)) + \mu_i) & \text{for } Q_i = 0. \quad (49) \end{cases}$$

This leaves us with two alternatives.

(I) $\lambda = 0$. Then (48) forces

$$F_i(Q_i) = \alpha_i \text{ for } Q_i > 0, \tag{50}$$

and (49) forces

$$F_i(0) = \alpha_i + \mu_i / \kappa_i g_i'(0) \geq \alpha_i \text{ for } Q_i = 0. \tag{51}$$

That is, $\mathbf{Q}$ is a vector with entries $Q_i = \max(Q_i(\alpha_i), 0)$.

(II) $\lambda > 0$. In this case (48) forces

$$F_i(Q_i) < \alpha_i \text{ for } Q_i > 0, \tag{52}$$

and for any i with $Q_i = 0$, (49) along with $\mu_i \geq 0$ gives a lower bound on λ,

$$\lambda \geq w_i^{-1}\kappa_i g_i'(0)(\alpha_i - F_i(0)). \tag{53}$$

And from (44), the constraint must be active, i.e.,

$$\mathbf{w}^T\mathbf{Q} = K. \tag{54}$$

Now since Z is a continuous function on the compact set C_K, some solution $\mathbf{Q} \in C_K$ of the AP exists. Uniqueness, however, requires some additional conditions on the g_i over the simplex $\{\mathbf{x} \in C_K \mid \mathbf{w}^T\mathbf{x} = K\}$. One such set of conditions – that for convenience we'll require over all of C_K – is that the g_i have non-positive second derivatives,

$$g_i''(x_i) \leq 0 \text{ whenever } F_i(x_i) < \alpha_i \tag{55}$$

so that from (9), the second partial derivatives $Z_{x_ix_i}$ are all non-negative on C_K. Because $Z_{x_ix_j} = 0, i \neq j$, the Hessian $\nabla^2 Z$ is then positive semi-definite on C_K so that Z in convex on C_K and any local minimizer for the AP is actually global. (Note that we might have to consider minimizing sets if any of the densities f_i vanish at a minimizer, but we'll set aside this possibility for now.) And along with (47), requiring (55) gives the upper bound

$$\lambda \leq \min_i w_i^{-1}(\kappa_i g_i'(0)(\alpha_i - F_i(0)) + \mu_i). \tag{56}$$

Finding this unique $\mathbf{Q}$ can sometimes be straightforward. Suppose, for example, there are no point masses, $F_i : [0, \varepsilon_i] \to [0, \delta_i]$ are bijective for some $\varepsilon_i, \delta_i > 0$ and all i, and g_i' are all identically 1, and that the constraint is active with $\mathbf{w}^T\mathbf{Q}(\boldsymbol{\alpha}) = K_1 > K$. Then we can write the equations (48) as

$$Q_i(\lambda) = F_i^{-1}(\alpha_i - \lambda w_i/\kappa_i) = F_i^{-1}(\alpha_i(1 - \lambda w_i/U_i)) \tag{57}$$

which are all defined for λ in the interval $I = [0, \lambda_0 = \min(U_i/w_i)]$. By our assuptions, the left hand side of the constraint equation (54)

$$\sum_{i=1}^{N} w_i Q_i(\lambda) = K. \tag{58}$$

is a continuous function mapping I to $[K_1, \sum_{i=1}^N w_i Q_i(\lambda_0)]$. If K is sufficently close to K_1, we will have a root $0 < \lambda^\star < \lambda_0$ which can be easily found using a formal or numerical search in I. And since our assumptions imply that $Q_i(\lambda^\star) > 0$, the KKT conditions (48) and (49) are met.

But if the constraint it too tight or the lower bounds of the supports of the F_i differ, we will potentially need to repeat the root search for each possible set $\{i \mid Q_i = 0\}$, facing intractability for large N. Our convexity assumption (55), however, besides ensuring existence and uniqueness, makes available a binary search method of (Zhang, Xu, and Hua 2009) which is only of polynomial complexity in N. In particular (55) guarantees that the functions

$$\lambda_i(x_i) := -\frac{1}{w_i}\frac{\partial Z}{\partial x_i}(x_i) = w_i^{-1}\kappa_i g_i'(x_i)(\alpha_i - F_i(x_i)) \tag{59}$$

are decreasing in x_i (cf. (9)) from

$$\lambda_i(0) = w_i^{-1}\kappa_i g_i'(0)(\alpha_i - F_i(0)) \quad \text{to} \quad \lambda_i(Q_i(\alpha_i)) = 0. \tag{60}$$

They can therefore can be inverted to give decreasing functions $x_i(\lambda_i)$ with

$$x_i(0) = Q_i(\alpha_i) \quad \text{and} \quad x_i(w_i^{-1}\kappa_i g_i'(0)(\alpha_i - F_i(0))) = 0. \tag{61}$$

ZXH binary search algorithm

```
**Input:** $F_i, Q_i, \kappa_i, \alpha_i, w_i, g_i, i = 1, \ldots, N,$
$K, \varepsilon_K, \varepsilon_\lambda, \mathrm{root}(f, I)$    ▷ root a function returning a root of function $f$ on interval $I$
**for** $i = 1$ to $N$ **do**
    $x_i \leftarrow \min(Q_i(\alpha_i), w_i^{-1} 2K)$    ▷ ensure constraint is violated if any $\alpha_i = 1$
**if** $\mathbf{w}^T\mathbf{x} \leq K$ **then return** $\mathbf{Q} = \mathbf{x}$    ▷ return quantiles if they satisfy constraint
$\lambda_1 \leftarrow \lambda_L \leftarrow 0$, $\lambda_U \leftarrow \max_i \lambda_i(0)$, $\tau \leftarrow 2$
**if** $\lambda_U \leq 0$ **then**
    **return** $\mathbf{Q} = \mathbf{0}$
**while** $|(\mathbf{w}^T\mathbf{x} - K)/K| > \varepsilon_K$ or $\lambda_U - \lambda_L > \varepsilon_\lambda$ **do**
    $\lambda_\tau \leftarrow (\lambda_U + \lambda_L)/2$
    **for** $i = 1$ to $N$ **do**
        **if** $\lambda_\tau < \lambda_i(0)$ **then**    ▷ that is, if, from (49) and (55), $\lambda_\tau$ cannot be a multiplier for $x_i = 0$
            **if** $\lambda_\tau < \lambda_{\tau-1}$ **then**
                $I_{i,\tau} \leftarrow [x_i, Q_i(\alpha_i)]$
            **else**
                $I_{i,\tau} \leftarrow [0, x_i]$
            $x_i \leftarrow \mathrm{root}(\lambda_i(x) - \lambda_\tau, I_{i,\tau})$    ▷ well-defined since $\lambda_i$, defined by (59), is decreasing
        **else**
            $x_i \leftarrow 0$    ▷ prevent decrease past 0 if $\lambda_\tau > \lambda_{\tau-1}$ and maintain at 0 if $\lambda_i(0) < \lambda_\tau < \lambda_{\tau-1}$
    **if** $\mathbf{w}^T\mathbf{x} < K$ **then**
        $\lambda_U \leftarrow \lambda_\tau$    ▷ unused capacity so $\lambda_{\tau+1} < \lambda_\tau$ and any $x_i$ could be increased but none will decrease
    **else**
        $\lambda_L \leftarrow \lambda_\tau$    ▷ capacity exceeded so $\lambda_{\tau+1} > \lambda_\tau$ and no $x_i$ will increase so any set to 0 will remain so
    $\tau \leftarrow \tau + 1$
**return** $\mathbf{Q} = \mathbf{x}$
```

Marginal benefit interpretation of multipliers

The functions (59) can each be interpeted as a *marginal expected unit benefit* of x_i. That is, if $F_i(x_i) < \alpha_i$ we receive the positive "benefit" of reducing Z at a rate of $\lambda_i(x_i)$ per additional unit of capacity allocated to the i'th target at "price" w_i. Note that with $\mathcal{S}_i$ the support of F_i we have

$$\lambda_i(x_i) = \begin{cases} \alpha_i w_i^{-1} \kappa_i g_i'(x_i) & x_i \leq \inf(\mathcal{S}_i) \\ (1-\alpha_i^{-1}) w_i^{-1} \kappa_i g_i'(x_i) & x_i \geq \sup(\mathcal{S}_i), \end{cases} \tag{62}$$

as in the uniform and exponential examples.

We can demonstrate the KKT condition (48) directly in terms of the λ_i:

Proposition 1. *For a solution* $\mathbf{Q} = (Q_1, \dots, Q_N)$ *of the AP we have*

$$\lambda_i(Q_i) = \lambda_j(Q_j) \textit{ whenever } Q_i, Q_j > 0. \tag{63}$$

Proof. Assume instead we had $\lambda_i(Q_i) > \lambda_j(Q_j)$ and $Q_i, Q_j > 0$. Moving away from $\mathbf{Q}$ in the direction $\mathbf{d}_{ij} = w_i^{-1}\mathbf{e}_i - w_j^{-1}\mathbf{e}_j$ we remain in $\mathbb{R}_+^N$ while respecting the constraint ($\langle \nabla \mathbf{w}^T\mathbf{x}, \mathbf{d}_{ij}\rangle = 0$) but reducing Z:

$$\langle \nabla Z(\mathbf{Q}), \mathbf{d}_{ij}\rangle = w_i^{-1}\frac{\partial Z}{\partial x_i}(Q_i) - w_j^{-1}\frac{\partial Z}{\partial x_j}(Q_j) = \lambda_j(Q_j) - \lambda_i(Q_i) < 0. \tag{64}$$

Therefore $\mathbf{Q}$ cannot be a constrained minimum of Z. □

But in line with KKT complementary slack conditions, the argument does not apply when $Q_j = 0$ and $\mathbf{d}_{ij}$ takes us immediately out of $\mathbb{R}_+^N$. This shows again how the set of indices $\{i \big| Q_i = 0\}$ arises as another variable of the problem.

Return to scoring rule definition:

Collect the AP parameters into $\boldsymbol{\beta} := \{K, \boldsymbol{\alpha}, \boldsymbol{\kappa}\}$. The resulting $Q_i^{\boldsymbol{\beta}} = Q_i(\boldsymbol{\beta})$ form the Bayes act for the the forecast $\mathbf{F}$ and the realized losses and costs defines the scoring rule evaluated on $\mathbf{F}$:

$$\begin{aligned} S_{\boldsymbol{\beta}}(\mathbf{F}, \mathbf{y}) = s(\mathbf{Q}^{\boldsymbol{\beta}}, \mathbf{y}) &= \sum O_i(g_i(Q_i^{\boldsymbol{\beta}}) - g_i(y_i))_+ + U_i(g_i(Q_i^{\boldsymbol{\beta}}) - g_i(y_i))_- && (65) \\ &= \sum ((U_i + O_i)\mathbb{1}\{Q_i^{\boldsymbol{\beta}} > y_i\} - U_i)(g_i(Q_i^{\boldsymbol{\beta}}) - g_i(y_i)) && (66) \\ &= \sum \kappa_i(\mathbb{1}\{Q_i^{\boldsymbol{\beta}} > y_i\} - \alpha_i)(g_i(Q_i^{\boldsymbol{\beta}}) - g_i(y_i)) && (67) \end{aligned}$$

Remarks:

- Quantile rules allow you to vary cost ratios while allocation rules add capacity as a parameter.
- While quantile rules are undefined for zero overprediction costs, allocation rules (with $K < \infty$) are defined even when some $\alpha_i = 1$. This covers one of the problems originally motivating this project: find the Bayes act for allocation $\mathbf{x}$ of hospital supplies to locations $l_i, i = 1, \dots, N$ given distributions F_i of need Y_i, a total available stock K of supplies, and a loss function

$$l(\mathbf{x}, \mathbf{y}) = U \sum_{i=1}^{N} (x_i - y_i)_-. \tag{68}$$

Special Case I: F_i from same location-scale family and $w_i = 1$, $O_i = O$.

Also assume the constraint is active, i.e.,

$$K < \sum F_i^{-1}((1+O)^{-1}) = \sum \mu_i + \sigma_i \Phi^{-1}((1+O)^{-1}) \tag{69}$$

where Φ is the standard CDF for the family. Then

$$\frac{1-\lambda}{1+O} = \Phi\left(\frac{K-\sum\mu_j}{\sum\sigma_j}\right) := F(K) \tag{70}$$

and the Bayes act when the constraint is set to K is

$$\begin{aligned} Q_i(K) = F_i^{-1}(F(K)) &= \mu_i + \sigma_i\Phi^{-1}(F(K)) \\ &= \mu_i + \sigma_i\left(\frac{K-\sum\mu_j}{\sum\sigma_j}\right) \\ &= \mu_i + \tilde{\sigma}_i(K-\sum\mu_j) \end{aligned} \tag{71}$$

where $\tilde{\sigma_i}$ is the proportion of F's scale (e.g. SD) "due" to F_i.

Remarks:

- $Q_i(K)$ here does not depend on O unlike when the constraint is inactive or the O_i differ.
- An "excess or shortage in mean" $K-\sum\mu_j$ is divided among the components in proportion to their scale factors as adjustmments up or down from their locations μ_i. This suggests an interpretation of under-dispersion of a forecast in one component as leading to that component not receiving as much additional recources as it should when there is an excess or other components not recieving as much of scarce resources as they should when there is a shortage.

The score for the forecast $\{F_i\}$ is

$$s_K(\mathbf{Q},\mathbf{y}) = \sum(1-(1+O)\mathbb{1}\left\{\frac{K-\sum\mu_i}{\sum\sigma_i} > \frac{y_i-\mu_i}{\sigma_i}\right\})(g_i(y_i)-g_i(Q_i(K))) \tag{72}$$

In the motivating case that $O=0$, this becomes

$$s_K(\mathbf{Q},\mathbf{y}) = \sum\mathbb{1}\left\{\frac{K-\sum\mu_i}{\sum\sigma_i} \leq \frac{y_i-\mu_i}{\sigma_i}\right\}(g_i(y_i)-g_i(Q_i(K))) \tag{73}$$

That is, a Q_i is only penalized when the standardized observed excess demand in component i exceeds the standardized excess in resources or the standardized observed shortfall in demand is not as large as the standardized shortage of resources.

Special Case II: F_i uniform on $[a_i, b_i]$

Problem: This does not seem to work. Take $F_1 = F_2 = F_{[0,1]}, \alpha_1 = \alpha_2 = .5, w_1 = 1, w_2 = 2$. Then

$$\lambda = \frac{3/2-K}{1/2+4/2} = \frac{3-2K}{5} > \frac{1}{2}$$

for $K < 1/4$, so that $Q_2(\alpha_2(1-w_2\lambda))$ is undefined. The minimizer here is $Q_1 = K, Q_2 = 0$, which does satisfy the KKT equations (48) and (49). $Q_1 = 0, Q_2 = K/2$ does not.

Following (Abdel-Malek, Montanari, and Morales 2004), we can directly evaluate (63) to solve for λ in the case of the uniform location-scale family ($\mu_i = a_i, \sigma = b_i - a_i$):

$$\begin{aligned} K &= \sum w_i F^{-1}_{[a_i,b_i]}\left(\alpha_i(1 - w_i\lambda)\right) \qquad (74)\\ &= \sum w_i Q_{[a_i,b_i]}\left(\alpha_i(1 - w_i\lambda)\right) \qquad (75)\\ &= \sum w_i \left[a_i + (b_i - a_i)(\alpha_i - w_i\alpha_i\lambda)\right] \qquad (76)\\ &= \sum w_i Q_{[a_i,b_i]}(\alpha_i) - (b_i - a_i)w_i^2\alpha_i\lambda \qquad (77)\end{aligned}$$

Writing $D = \sum w_i Q_{[a_i,b_i]}(\alpha_i) - K$ for the “deficit” of the unconstrained solution, we have

$$\lambda = \frac{D}{\sum (b_j - a_j)w_j^2\alpha_j} \tag{78}$$

$$\tag{79}$$

and from (**??**),

$$\begin{aligned} Q_i &= F^{-1}_{[a_i,b_i]}\left(\alpha_i\left(1 - \frac{w_i D}{\sum_j (b_j - a_j)w_j^2\alpha_j}\right)\right) \qquad (80)\\ &= Q_{[a_i,b_i]}(\alpha_i) - \frac{(b_i - a_i)w_i\alpha_i D}{\sum (b_j - a_j)w_j^2\alpha_j}. \qquad (81)\end{aligned}$$

That is, each constrained Q_i is obtained by reducing the α_i quantile by an amount of the deficit proportional to the product of spread, resource weighting, and cost/loss ratio (what should this be called?) of the corresponding prediction target.

General solution method

Is there anything we can salvage here?

Again following (Abdel-Malek, Montanari, and Morales 2004), we can use this solution for uniform F_i as the basis of an iterative procedure for approximating the Q_i under a constraint for general predictive distributions. Assuming that the unconstrained solution $\{Q_i(\alpha_i)\}$ is not feasible and that we can evaluate F_i, F_i^{-1} and f_i at least approximately at arbitrary x_i, we first linearize the equations (63) at $\{(\alpha_i, Q_i(\alpha_i))\}$ by replacing the F_i^{-1} with the tangential approximations

$$\left(F_i^{(1)}\right)^{-1}(p) := Q_i(\alpha_i) + \frac{p - \alpha_i}{f_i(Q_i(\alpha_i))}, \tag{82}$$

that is, the quantile functions of the uniform cdf’s

$$F_i^{(1)} := \min(0, \max(1, \alpha_i + f_i(Q_i(\alpha_i))(x_i - Q_i(\alpha_i)))). \tag{83}$$

As before, these equations

$$\begin{aligned} K &= \left(F_i^{(1)}\right)^{-1}\left(\alpha_i(1 - w_i\lambda)\right) \qquad (84)\\ &= \sum w_i Q_i(\alpha_i) - \frac{w_i^2\alpha_i\lambda}{f_i(Q_i(\alpha_i))} \qquad (85)\end{aligned}$$

are linear in λ and have the solution

$$\lambda^{(1)} = \frac{D}{\sum \frac{w_j^2\alpha_j}{f_j(Q_j(\alpha_j))}}. \tag{86}$$

This provides the first iterate

$$Q_i^{(1)} = F_i^{-1}(\alpha_i(1 - w_i\lambda^{(1)})) \tag{87}$$

and a new deficit (or negative surplus?)

$$D^{(1)} = \sum w_i Q_i^{(1)} - K. \tag{88}$$

Now we can repeat the process with new tengential approximations

$$\left(F_i^{(2)}\right)^{-1}(p) := Q_i^{(1)} + \frac{p - \alpha_i^{(1)}}{f_i\left(Q_i^{(1)}\right)} \tag{89}$$

where $\alpha_i^{(1)} := \alpha_i(1 - w_i\lambda^{(1)})$ to get the next λ iterate

$$\lambda^{(2)} = \frac{D^{(1)}}{\sum \frac{w_j^2 \alpha_j^{(1)}}{f_j\left(Q_j^{(1)}\right)}}, \tag{90}$$

continuing with $\alpha_i^{(\tau)} := \alpha_i^{(\tau-1)}(1 - w_i\lambda^{(\tau)})$ until the relative error at the τ'th iteration

$$\varepsilon^{(\tau)} = D^{(\tau)}/K \tag{91}$$

is sufficiently small.

References

Abdel-Malek, Layek, Roberto Montanari, and Libia Cristina Morales. 2004. “Exact, Approximate, and Generic Iterative Models for the Multi-Product Newsboy Problem with Budget Constraint.” *International Journal of Production Economics* 91 (2): 189–98.

Nocedal, Jorge, and Stephen Wright. 2006. *Numerical Optimization*. Springer.

Zhang, Bin, Xiaoyan Xu, and Zhongsheng Hua. 2009. “A Binary Solution Method for the Multi-Product Newsboy Problem with Budget Constraint.” *International Journal of Production Economics* 117 (1): 136–41.

‘