# Proper Scoring Rules

Luiz Max de Carvalho[lmax.fgv@gmail.com],

School of Applied Mathematics, Getúlio Vargas Foundation (FGV), Rio de Janeiro.

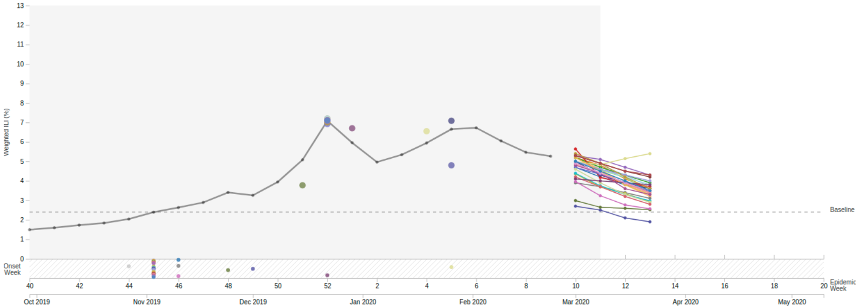
October 13, 2020

# Plan for today

- Motivation;
- Definitions and maths;
- Examples;
- Why I am interested in these things (maybe);

## Why?

- How do you assess a prediction from a model?
  **Desiderata**:
  - (Well-calibrated) Probabilistic predictions;
  - Encourage careful and honest predictions;



## Fixing notation

Let $\mathcal{P}$ be a convex class of probability measures on $(\Omega, \mathcal{A})$. We call $P \in \mathcal{P}$ a probabilistic forecast.

### Definition 1 (Scoring rule)

We say $S(P, \cdot) : \Omega \to [-\infty, \infty]$ is a **scoring rule** if it is measurable and $S(P, \cdot)$ is $P$-quasi-integrable for all $P \in \mathcal{P}$.

The expected score under $Q \in \mathcal{P}$ if the forecast is $P$ is

$$S(P, Q) := \int_{\Omega} S(P, \omega) dQ(\omega).$$

### Definition 2 (Strictly proper scoring rule)

We say $S$ is **proper** if $S(Q, Q) \geq S(P, Q)$ for all $P, Q \in \mathcal{P}$. In addition, we say $S$ is **strictly proper** if equality is achieved only for $P = Q$.

## Definition 3 (Strong equivalence)

If $S$ is a (strictly) proper scoring rule and $c \geq 1$ is a constant and $h$ is a $\mathcal{P}$-integrable function, then Gneiting & Raftery, (2007), Eq.2:

$$S^*(P, \omega) = cS(P, \omega) + h(\omega) \tag{1}$$

is also a (strictly) proper scoring rule and we say $S^*$ and $S$ are **strongly equivalent** if $c = 1$.

**Connection with convex functions:**

## Theorem 4 (Convex functions and proper scoring rules)

*A regular scoring rule $S : \mathcal{P} \times \Omega \to \bar{\mathbb{R}}$ is proper relative to the class $\mathcal{P}$ if and only if there exists convex $G : \mathcal{P} \to \mathbb{R}$ such that*

$$S(P, \omega) = G(P) - \int G^*(P, \omega) dP(\omega) + G^*(P, \omega),$$

*where $G^*(P, \omega)$ is a subtangent of $G$ at $P$.*

## Information measures and (Bregman) divergences

The function

$$G(P) = \sup_{Q \in \mathcal{P}} S(Q, P), P \in \mathcal{P} \tag{2}$$

is the **information measure** (or generalised entropy function) associated with $S$. Subject to regularity conditions on $S$, we can define

$$d(P, Q) = S(Q, Q) - S(P, Q),\ P, Q \in \mathcal{P}, \tag{3}$$

as the **divergence function** associated with $S$ (and $G$). Under regularity conditions on $\Omega$, $d(\cdot, \cdot)$ is called a *Bregman* divergence.
**Note:** if $S$ is strictly proper, $d(P, Q) \geq 0$ with equality iff $P = Q$.

## A first "example"

**Statistical decision problems**. If we let $U(\omega, a)$ be the utility for outcome $\omega$ under action $a$ and $\mathcal{P}$ be a convex family of probability measures, then

$$S(P, \omega) = U(\omega, a_P),$$

where $a_P$ is the Bayes act for $P \in \mathcal{P}$, is a proper scoring rule.

## Skill scores

If we want to rank $n$ forecasts, *as long as they refer to the* *<u>same</u>* *set of forecast situations*, we can compute

$$\mathcal{S}_n := \frac{1}{n}\sum_{i=1}^{n} S(P_i, x_i).$$

Since forecasts are likely to vary – in quality – spatially and temporally, we can compute the **skill score**

$$\mathcal{S}_n^{\text{skill}} := \frac{\mathcal{S}_n^{\text{fcst}} - \mathcal{S}_n^{\text{ref}}}{\mathcal{S}_n^{\text{opt}} - \mathcal{S}_n^{\text{ref}}}, \tag{4}$$

where

- $\mathcal{S}_n^{\text{fcst}}$ is the forecaster's score;
- $\mathcal{S}_n^{\text{opt}}$ is a hypothetical optimal forecast;
- $\mathcal{S}_n^{\text{ref}}$ is the score for a reference (model or) strategy.

## A slightly more enlightening example

Suppose $\Omega = \{1, 2, \ldots, m\}$ consisting of mutually exclusive events and that $\mathcal{P}$ is the set of open $m-$dimensional unit simplices. If a forecaster quotes a vector $\boldsymbol{p} \in \mathcal{P}$ and event $i$ materialises, then their reward is $S(\boldsymbol{p}, i)$.

### Remark 1 (Convexity)

*A (regular) scoring rule is proper if and only if $G(\boldsymbol{p}) = S(\boldsymbol{p}, \boldsymbol{p})$ is convex.*

**Example:** Brier score. If $G(\boldsymbol{p}) = \sum_{j=1}^{m} p_j^2 - 1$, then we have the *Brier score*:

$$S(\boldsymbol{p}, i) = -\sum_{j=1}^{m} (\mathbb{I}_j(i) - p_j)^2 - \sum_{j=1}^{m} p_j^2 - 1 \tag{5}$$

- **Spherical score**.
  For $\alpha > 1$ we can define

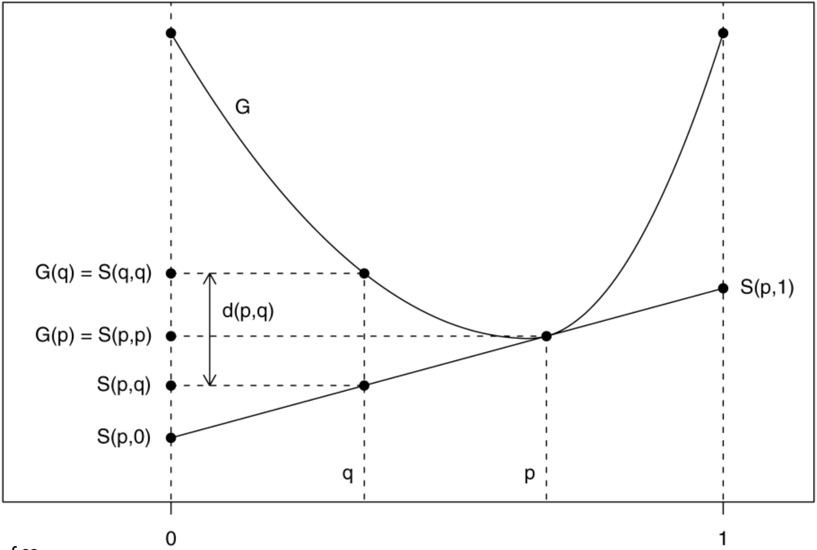
$$S(\boldsymbol{p}, i) = \frac{p_i^{\alpha-1}}{\left(\sum_{j=1}^{m} p_j^{\alpha}\right)^{\frac{\alpha-1}{\alpha}}} \tag{6}$$

- **Logarithmic score**.
  When $G(\boldsymbol{p}) =$ is the Shannon entropy, we have $S(\boldsymbol{p}, i) = \log p_i$ and

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sum_{j=1}^{m} \log\left(\frac{q_j}{p_j}\right), \tag{7}$$

  as the Kullback-Leibler divergence.

# A (helpful?) visualisation



# Scoring rules for density forecasts

Let $\mu$ be a $\sigma$-finite measure on $(\Omega, \mathcal{A})$. For $\alpha > 1$ define $\mathcal{L}_\alpha$ be the space of probability measures on $(\Omega, \mathcal{A})$ such that $\nu \ll \mu$ and $p(\omega) = \frac{d\nu}{d\mu}(\omega)$ and

$$||p||_\alpha = \left( \int_\Omega p(\omega)^\alpha d\mu(\omega) \right)^\alpha < \infty.$$

We establish a correspondence between the forecast $P$ and its $\mu$-density, $p$.

**Examples:**

- **Quadratic:**

$$\text{QS}(p, \omega) = 2p(\omega) - ||p||_2^2, \tag{8}$$

  is strictly proper relative to $\mathcal{L}_2$ class of probability measures.
- **Pseudo-spherical:**

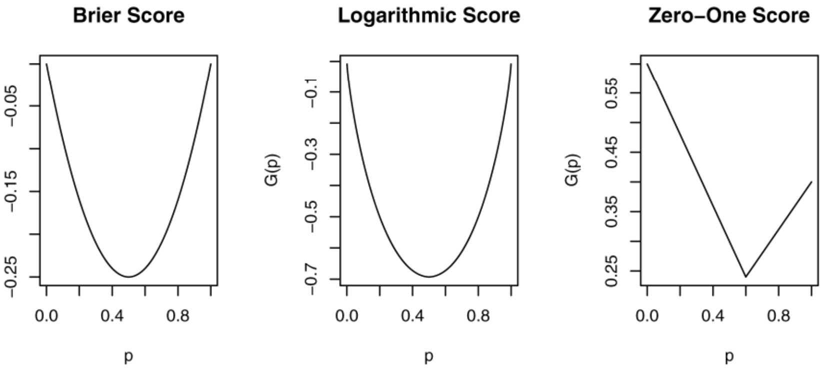
$$\text{PseudoS}(p, \omega) = \frac{p(\omega)^{\alpha-1}}{||p||_\alpha^{\alpha-1}}, \tag{9}$$

- **Logarithmic score:**

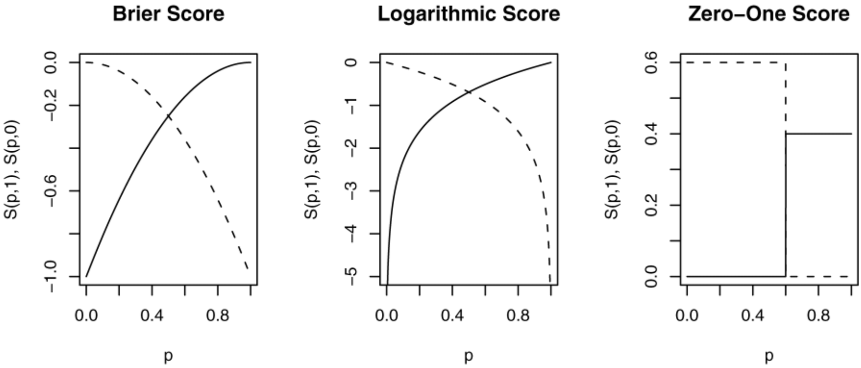
$$\text{LogS}(p, \omega) = \log p(\omega), \tag{10}$$

  is what happens to the pseudo-spherical score when $\alpha \to 1$.

## Illustrations: $G(p)$



## Illustrations: $S(p, 1)$ and $S(p, 0)$



## Continuous ranked probability score

The continuous ranked probability score (CRPS):

$$\text{CRPS}(F, x) = -\int_{-\infty}^{\infty} \left(F(y) - \mathbb{I}\{y \geq x\}\right)^2 \, dy, \tag{11}$$

can be seen as the integral of the Brier scores for the associated binarisation of the forecasts based on $x$ as cutoff.

**Example:**

$$\text{CRPS}\left(\text{Normal}(\mu, \sigma^2), x\right) = \sigma \left[ \frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma}\left(2\Phi\left(\frac{x-\mu}{\sigma}\right) - 1\right)\right],$$

where $\varphi$ and $\Phi$ are the probability density function and cumulative distribution function of a standard normal, respectively.

# Loads more!

The article lists a bunch more scoring rules:

- Energy score (Section 4.3), a generalisation of CRPS;
- Scoring rules that depend on the first two moments only (Section 4.4);
- Kernel scores (Section 5) – negative-definite functions and (Hoeffding) expectations inequalities;
- Random-fold cross-validation (Section 7.2).

## Quantiles and Bayes factors

A few more connections.

- **Quantiles:**
  If one quotes the quotes the quantiles $r_1, \dots, r_k$, we have

$$S(r_1, \dots, r_k; P) = \int S(r_1, \dots, r_k; x)\, dP(x)$$

  as a proper scoring rule under somewhat mild technical conditions (see Theorem 6 in Gneiting & Raftery, (2007)).

- **Bayes factors:**
  We have

$$B_{12} = \frac{P(\boldsymbol{X} \mid H_1)}{P(\boldsymbol{X} \mid H_2)},$$

  and thus

$$\log B_{12} = \operatorname{LogS}(H_1, \boldsymbol{X}) - \operatorname{LogS}(H_2, \boldsymbol{X}),$$

  is a proper scoring rule.

## Case study I: interval forecasts for heterokedastic processes

The (Markovian) model is

$$X_{t+1} = \frac{1}{2}X_t + \frac{1}{2}X_t\epsilon_t + \epsilon_t, \; \epsilon_t \sim \text{Normal}(0, 1).$$

Interval predictions for $X_{t+1}$ will be

$$I := \left[\frac{1}{2}X_t - c\left|1 + \frac{1}{2}X_t\right|, \frac{1}{2}X_t + c\left|1 + \frac{1}{2}X_t\right|\right], \tag{12}$$

with $c = \Phi^{-1}(\frac{1+\alpha}{2})$ and $\alpha = 0.95$.
Alternative forecast is

$$J := \left[F^{-1}\left(\frac{1-\alpha}{2}\right), F^{-1}\left(\frac{1+\alpha}{2}\right)\right], \tag{13}$$

where $F$ is the unconditional stationary distribution of $X_t$. Finally, consider also

$$K := \left[\frac{1}{2}X_t - \gamma\left(\left|1 + \frac{1}{2}X_t\right|\right), \frac{1}{2}X_t + \gamma\left(\left|1 + \frac{1}{2}X_t\right|\right)\right] \tag{14}$$

where $\gamma(a) = a\sqrt{2\left(\log 7.36 - \log a\right)}\mathbb{I}(a < 7.26)$, which minimises interval width subject to nominal coverage.

# Results I

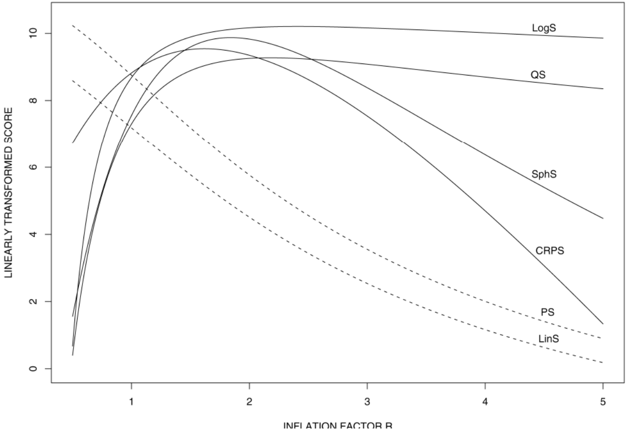
Table 2. *Comparison of One-Step-Ahead 95% Interval Forecasts for the Stationary Bilinear Process (44)*

| *Interval forecast* | | *Empirical coverage* | *Average width* | *Average interval score* |
|---|---|---|---|---|
| $I$ | (45) | 95.01% | 4.00 | 4.77 |
| $J$ | (46) | 95.08% | 5.45 | 8.04 |
| $K$ | (47) | 94.98% | 3.79 | 5.32 |

NOTE: The table shows the empirical coverage, the average width, and the average value of the negatively oriented interval score (43) for the prediction intervals $I$, $J$, and $K$ in 100,000 sequential forecasts in a sample path of length 100,001. See text for details.

# Case study II: forecasting sea-level pressure

Key idea: **model ensembles**.
Five models give 48-hour look ahead predictions for sea-level pressure in (some places of) the Pacific Ocean.
**Problem:** underdispersed predictions by the ensemble.
**Solution:** Come up with an inflation factor ($R$).

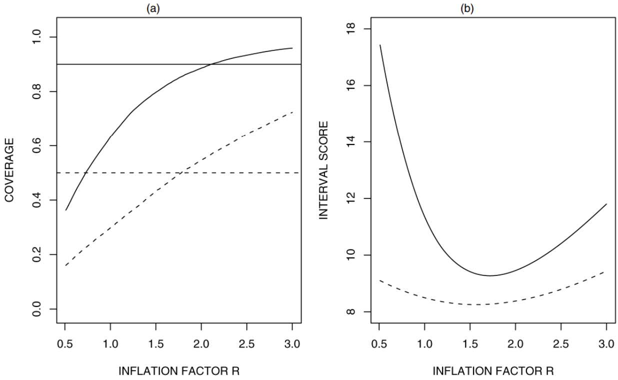


# Results IIb

For upper ($u_i$) and lower ($l_i$) predictions, we can evaluate interval forecasts for the sea-level pressure problem using, for $r > 0$,

$$s_\alpha(r) = \frac{1}{16,015} \sum_{i=1}^{16,015} S_\alpha^{\text{int}}(l_i, u_i; x),$$

as a scoring rule, which takes both calibration and sharpness into account.



- The obvious: assess COVID-19 forecasts;
- The not-so-obvious: InfoDengue and InfoGripe ("gripe" is Portuguese for the flu);

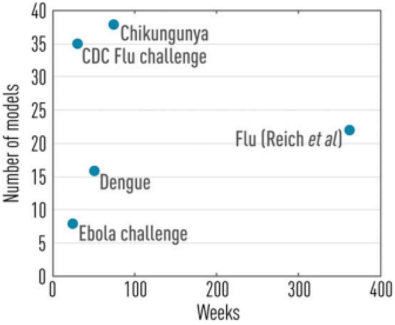



**Fig. 1. Past and present infectious disease forecasting challenges as a function of prediction horizon and number of models considered (data from refs. 10–14).**

Figure 1 of Viboud & Vespigiani (2019).

## Conclusions

- Proper scoring rules (PSRs) are cool! They permeate many seemingly unrelated things;
- A proper scoring rule encourages honest and well-calibrated forecasts;
- It is possible to define PSRs for interval, point and distributional forecasts;
- For more, see the work by Alexander P. Dawid, starting with Dawid (1984).