Locally scale invariant proper scoring rules

LUND
UNIVERSITY

Jonas Wallin,
Lund University

Joint work with David Bolin, KAUST

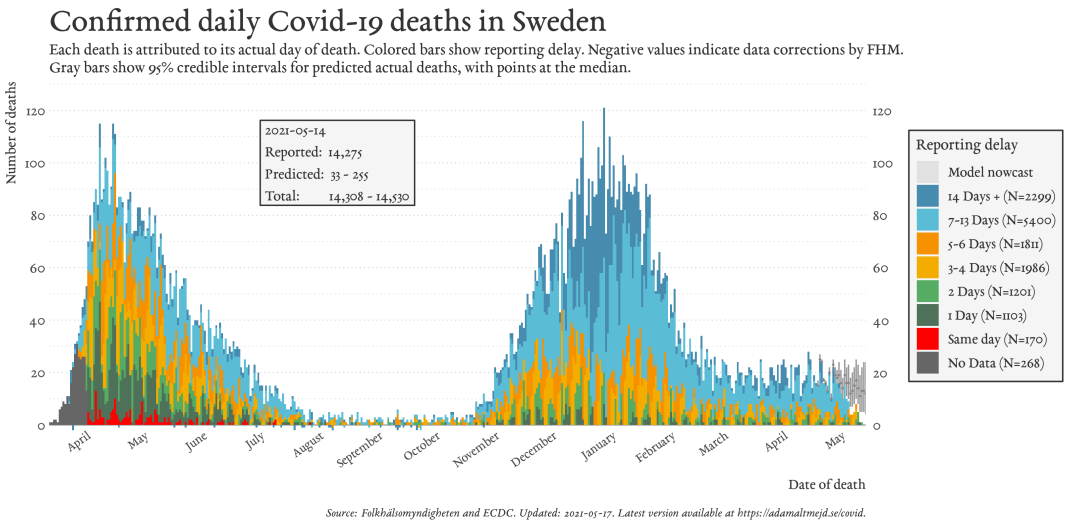
April 21, 2026

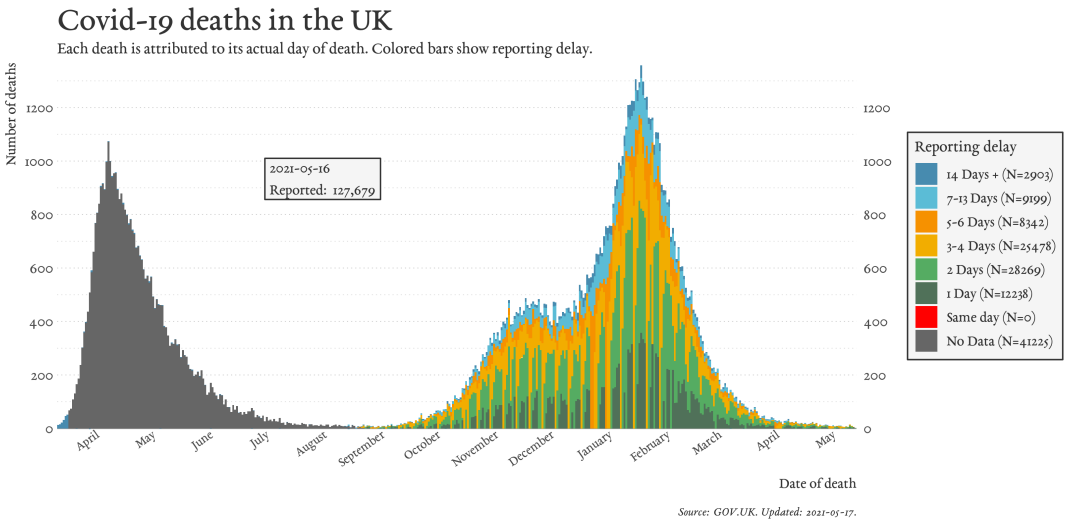
Main articles

This talk is mainly centered around the following three articles:

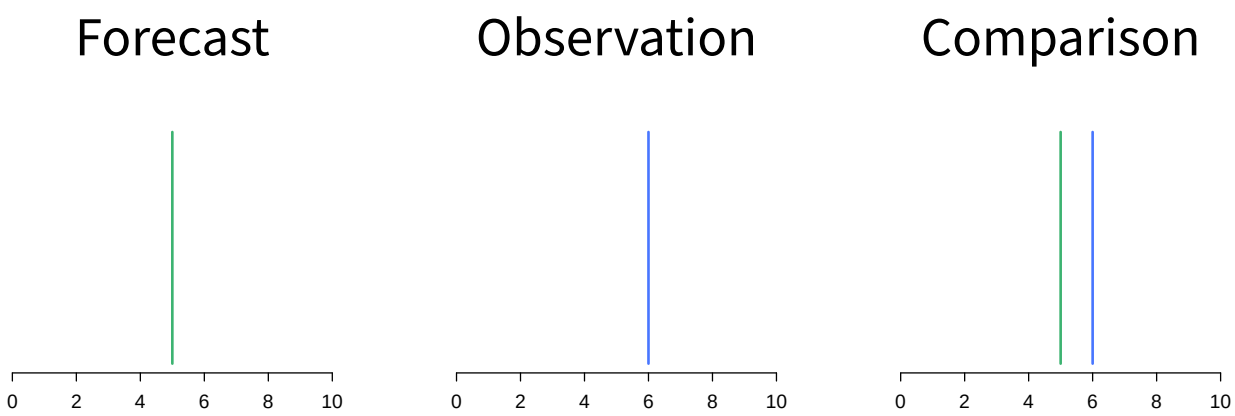
1. Bolin, D. and Wallin, J. (2023). *Local scale invariance and robustness of proper scoring rules*. *Statistical Science*, 38(1), 140–159.
2. Altmejd, A., Rocklöv, J. and Wallin, J. (2023). *Nowcasting COVID-19 statistics reported with delay: a case-study of Sweden and the UK*. *International Journal of Environmental Research and Public Health*, 20(4), 3040.
3. Bosse, N. I., Abbott, S., Cori, A., van Leeuwen, E., Bracher, J. and Funk, S. (2023). *Scoring epidemiological forecasts on transformed scales*. *PLOS Computational Biology*, 19(8), e1011393.

Nowcasting COVID deaths in Sweden and UK





Comparing models

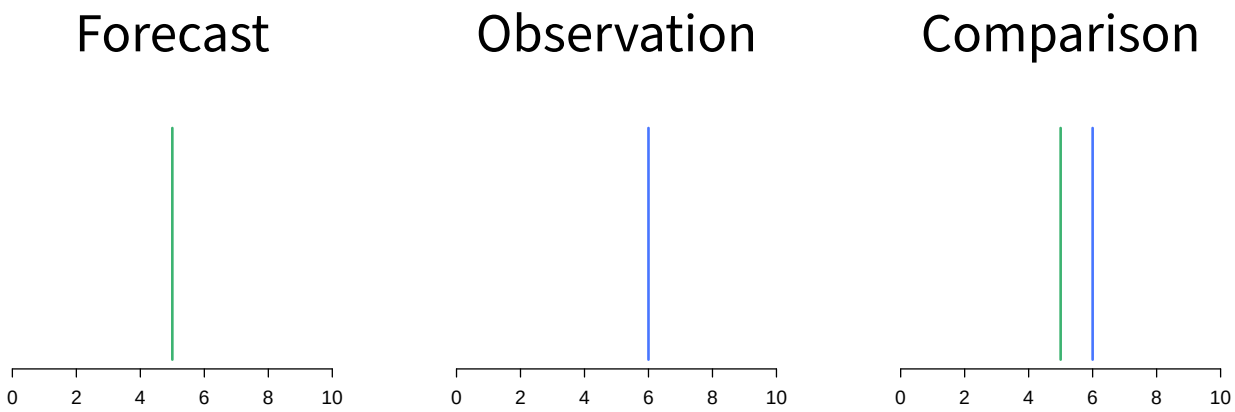


- Observation: y. Predictions: $\hat{y}^A$ and $\hat{y}^B$.
- Model selection: select Model A if

$$(y - \hat{y}^A)^2 < (y - \hat{y}^B)^2,$$

or $$|y - \hat{y}^A| < |y - \hat{y}^B|.$$

Comparing models

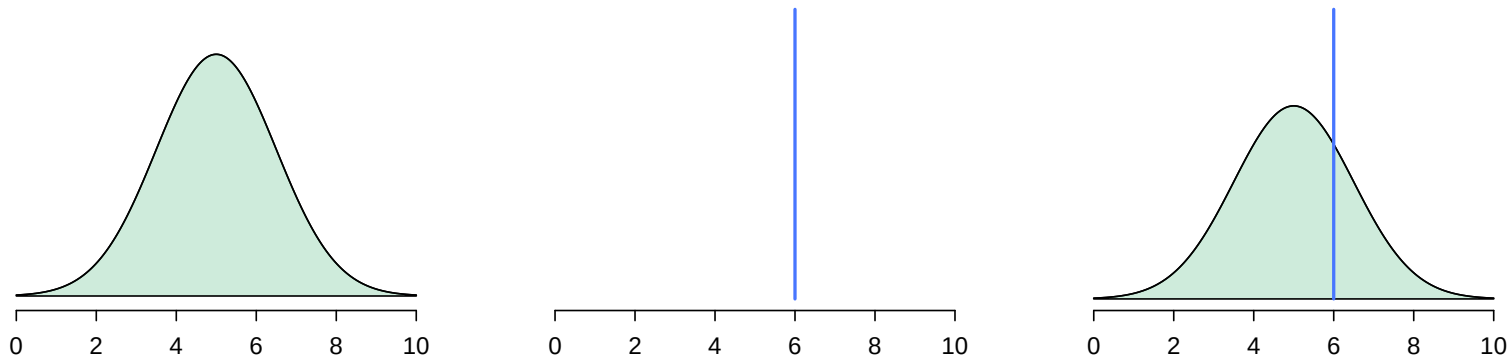


- Observation: y. Predictions: $\hat{y}^A$ and $\hat{y}^B$.
- Model selection: select Model A if

$$(y - \hat{y}^A)^2 < (y - \hat{y}^B)^2,$$

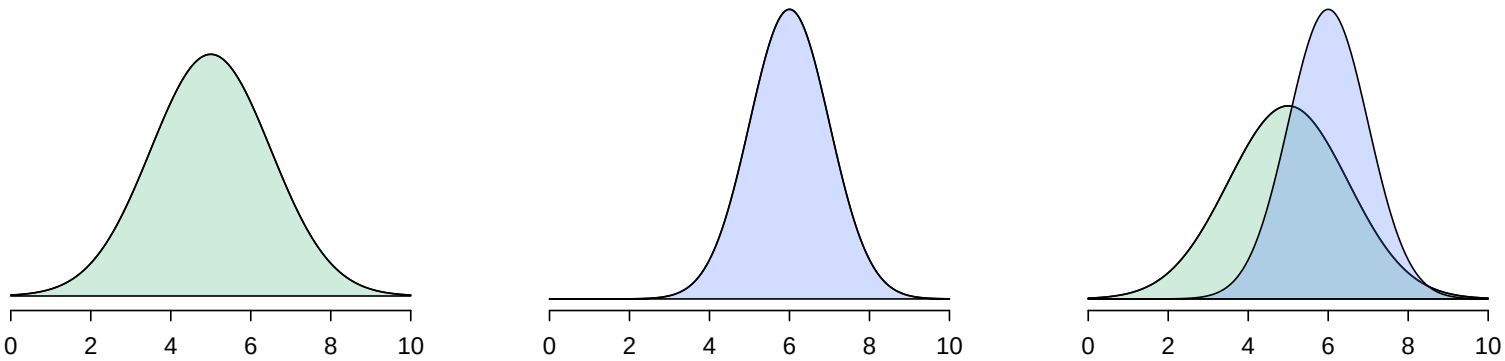
or $$|y - \hat{y}^A| < |y - \hat{y}^B|.$$

Proper scoring rules



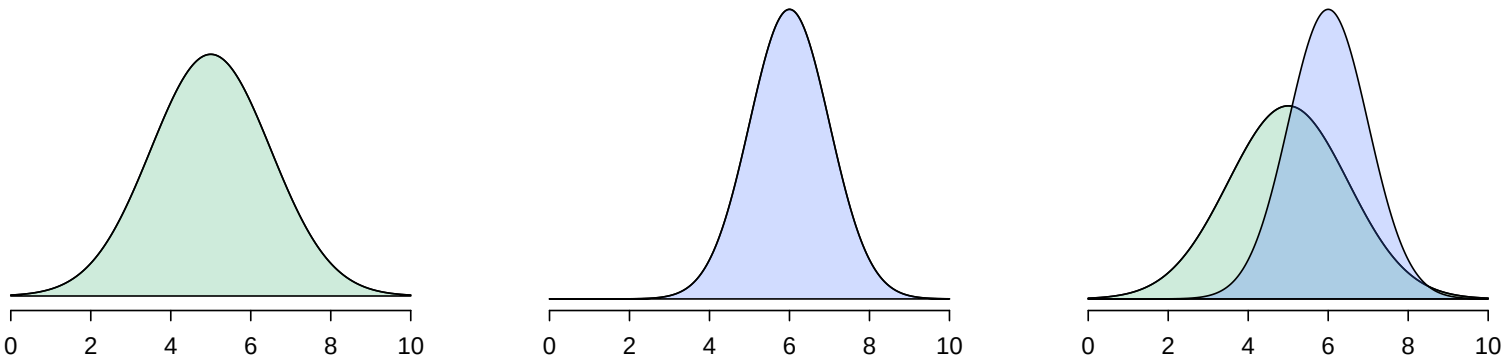
- A scoring rule, $S(\mathbb{P}, y)$ ($|\hat{y}_i(\mathbb{P}) - y_i|^2$) gives an ordering of a forecast based on a predictive distribution $\mathbb{P}$.
- A scoring rule is proper if $\mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{Q}, Y)] \geq \mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{P}, Y)]$ for all $\mathbb{P}$ and $\mathbb{Q}$.
- MSE are proper scoring rules, $S(\mathbb{P}, y_i) = |\hat{y}_i(\mathbb{P}) - y_i|^2$ if we use $\hat{y}_i(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[Y_i]$.

Proper scoring rules



- A scoring rule, $S(\mathbb{P}, y)$ ($|\hat{y}_i(\mathbb{P}) - y_i|^2$) gives an ordering of a forecast based on a predictive distribution $\mathbb{P}$.
- A scoring rule is proper if $\mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{Q}, Y)] \geq \mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{P}, Y)]$ for all $\mathbb{P}$ and $\mathbb{Q}$.
- MSE are proper scoring rules, $S(\mathbb{P}, y_i) = |\hat{y}_i(\mathbb{P}) - y_i|^2$ if we use $\hat{y}_i(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[Y_i]$.

Proper scoring rules



- A scoring rule, $S(\mathbb{P}, y)$ ($|\hat{y}_i(\mathbb{P}) - y_i|^2$) gives an ordering of a forecast based on a predictive distribution $\mathbb{P}$.
- A scoring rule is proper if $\mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{Q}, Y)] \geq \mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{P}, Y)]$ for all $\mathbb{P}$ and $\mathbb{Q}$.
- MSE are proper scoring rules, $S(\mathbb{P}, y_i) = |\hat{y}_i(\mathbb{P}) - y_i|^2$ if we use $\hat{y}_i(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[Y_i]$.

Epidemic example: model and forecasts

Observed weekly counts

$$y_1 = 20 \qquad \text{(low-incidence week)}, \qquad y_2 = 400 \qquad \text{(peak week)}.$$

Forecast model

For each week $t \in \{1, 2\}$, let the predictive distribution be

$$Y_t \sim \mathrm{NegBin}(\mu_t, s_t),$$

with

$$\mathbb{E}(Y_t) = \mu_t, \qquad \mathrm{Var}(Y_t) = \mu_t + \frac{\mu_t^2}{s_t}, \qquad \mathrm{SD}(Y_t) = \sqrt{\mu_t + \frac{\mu_t^2}{s_t}}.$$

Two-week epidemic forecasts

Model	Week	μ	s	SD	Obs. y	Interpretation
A	1	18	80	4.68	20	low week, close and precise
A	2	320	6	136.63	400	peak week, low-biased and uncertain
B	1	10	80	3.61	20	low week, too low and precise
B	2	380	6	157.48	400	peak week, close and uncertain

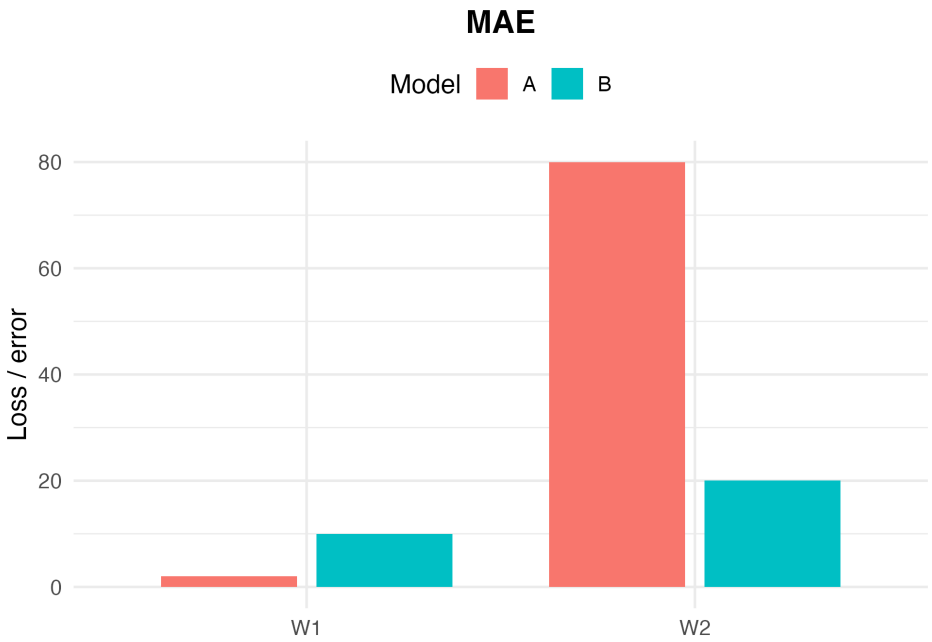
- Raw errors:

$$\text{A: } (|20-18|, |400-320|) = (2, 80), \qquad \text{B: } (|20-10|, |400-380|) = (10, 20).$$

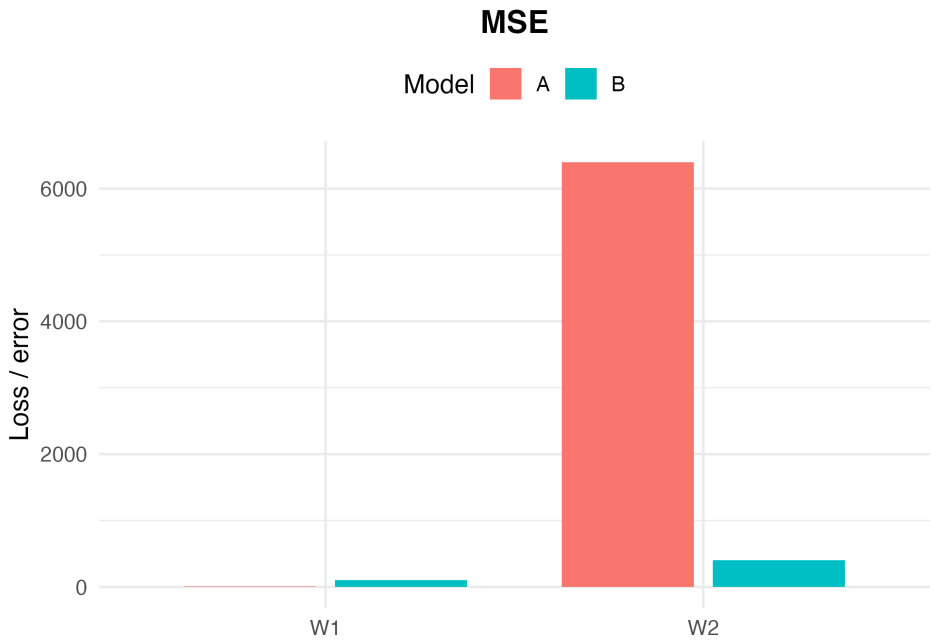
- Standardized errors (approximately):

$$\text{A: } \left(\frac{2}{4.68}, \frac{80}{136.63}\right) \approx (0.43, 0.59), \qquad \text{B: } \left(\frac{10}{3.61}, \frac{20}{157.48}\right) \approx (2.77, 0.13).$$

MAE and MSE by week

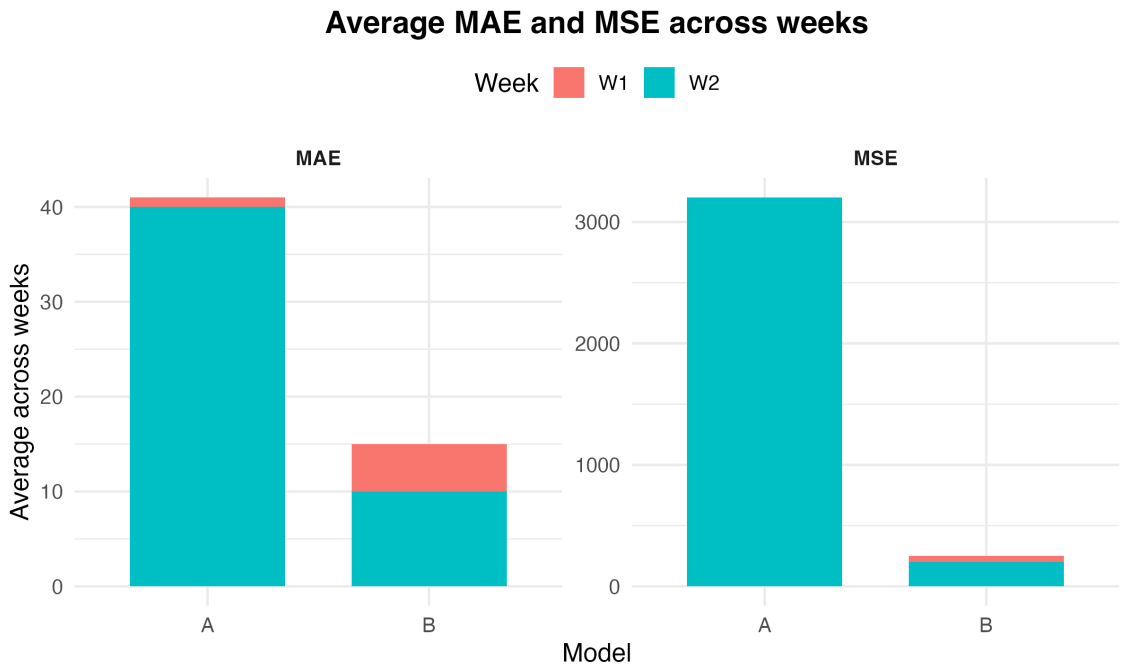


MAE



MSE

MAE and MSE by average



- Raw errors:

  A: $(|20-18|, |400-320|) = (2, 80)$, B: $(|20-10|, |400-380|) = (10, 20)$.

- Standardized errors (approximately):

  A: $\left(\frac{2}{4.68}, \frac{80}{136.63}\right) \approx (0.43, 0.59)$, B: $\left(\frac{10}{3.61}, \frac{20}{157.48}\right) \approx (2.77, 0.13)$.

Scoring rules

- For $S(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{P}, Y)]$ we have

$$\mathbb{E}_{\mathbb{P}_1}[Y] = \mathbb{E}_{\mathbb{P}_2}[Y] \implies S_{MSE}(\mathbb{P}_1, \mathbb{Q}) = S_{MSE}(\mathbb{P}_2, \mathbb{Q})$$

- Today it is almost mandatory to also use some proper scoring rule, such as the CRPS, that evaluates the quality of the entire predictive distribution.

$$\mathrm{CRPS}(\mathbb{P}, y) = \frac{1}{2}\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X - Y|] - \mathsf{E}_{\mathbb{P}}[|X - y|]$$

Scoring rules

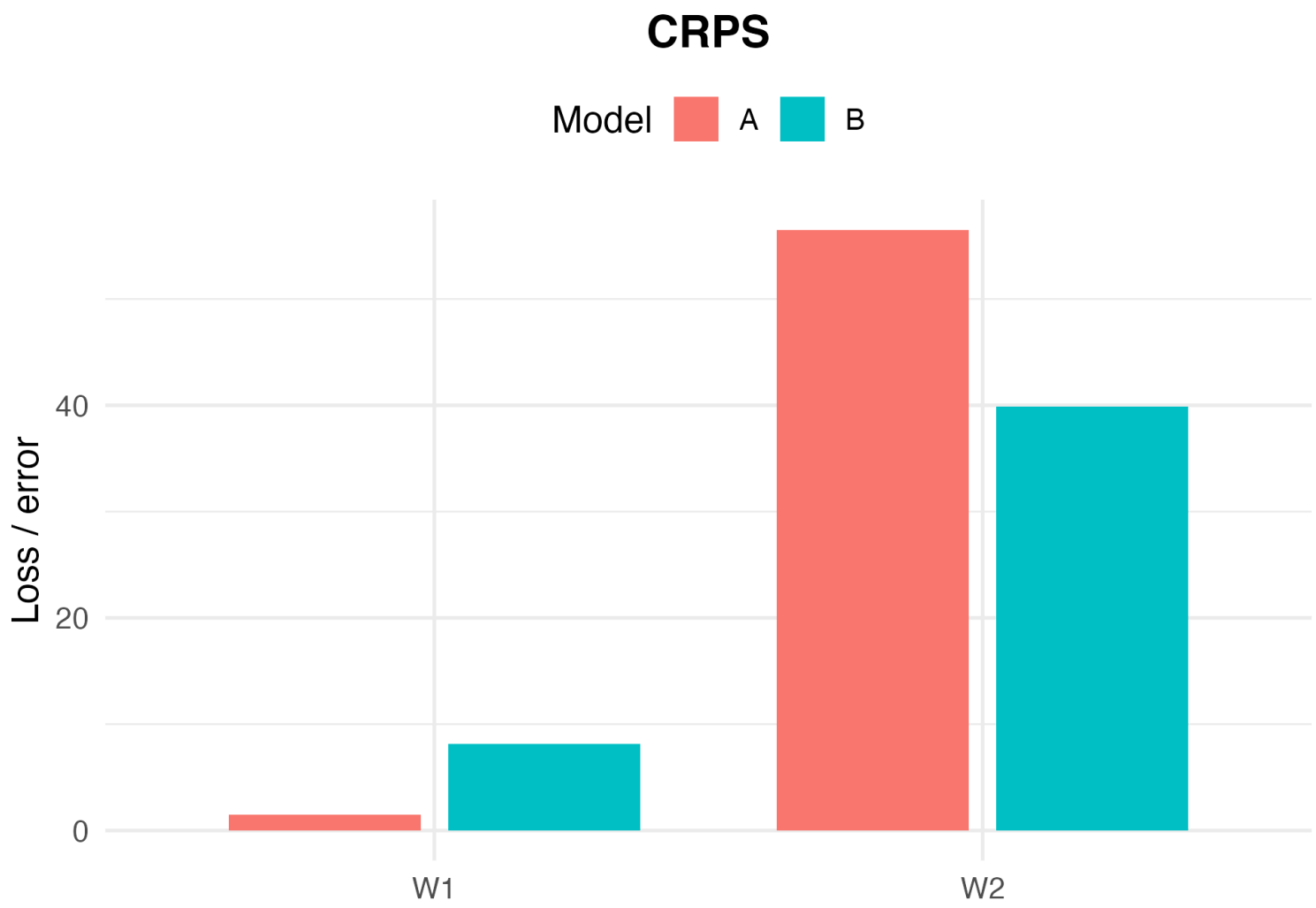
- For $S(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{Y\sim\mathbb{Q}}[S(\mathbb{P}, Y)]$ we have

$$\mathbb{E}_{\mathbb{P}_1}[Y] = \mathbb{E}_{\mathbb{P}_2}[Y] \implies S_{MSE}(\mathbb{P}_1, \mathbb{Q}) = S_{MSE}(\mathbb{P}_2, \mathbb{Q})$$

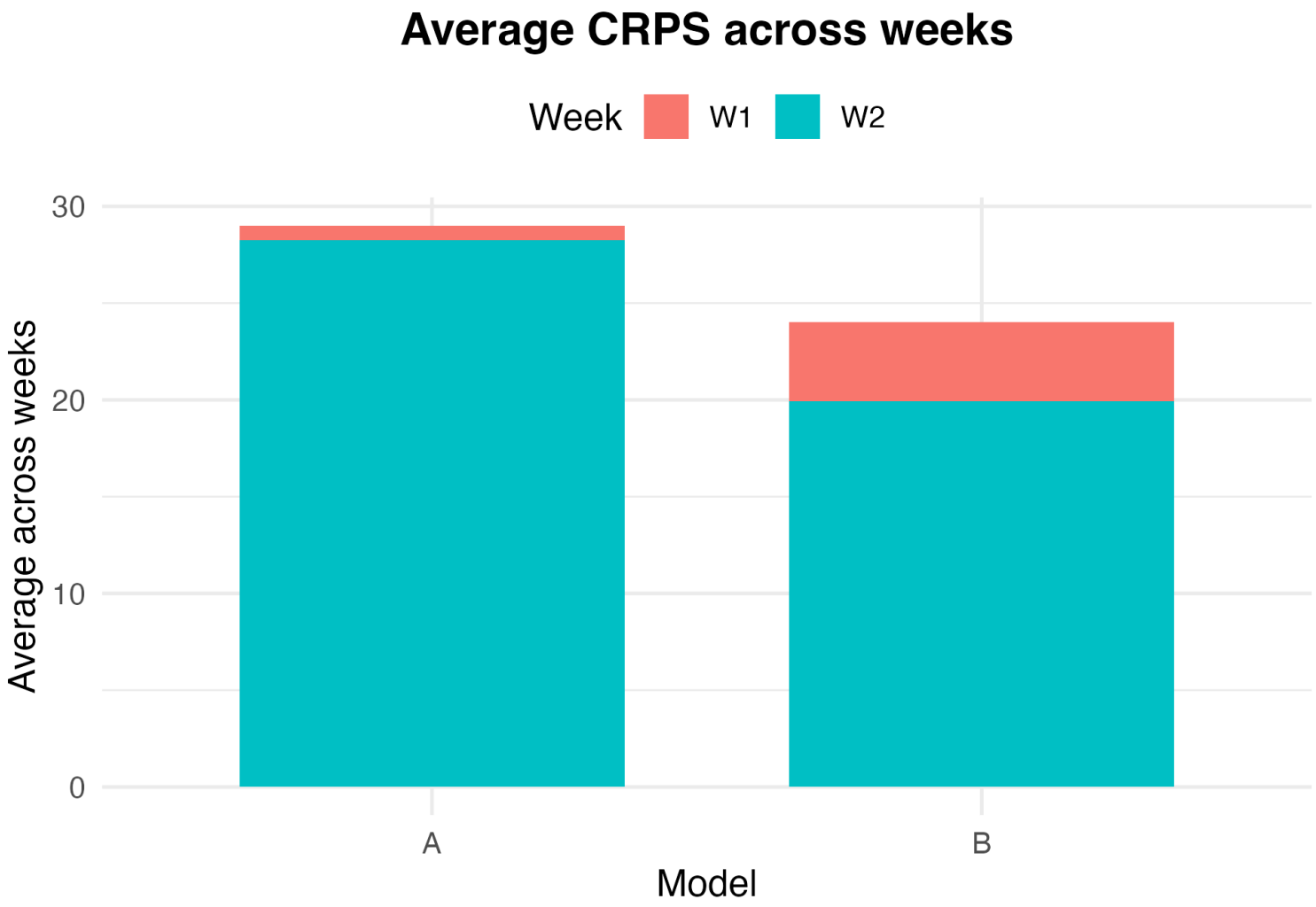
- Today it is almost mandatory to also use some proper scoring rule, such as the CRPS, that evaluates the quality of the entire predictive distribution.

$$\mathrm{CRPS}(\mathbb{P}, y) = \frac{1}{2}\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X - Y|] - \mathsf{E}_{\mathbb{P}}[|X - y|]$$

CRPS by week



CRPS average



Proper scoring rules

- For a set of observations, we typically use an average score

$$S_n = \frac{1}{n}\sum_{i=1}^{n} S(\mathbb{P}_i, y_i).$$

- But any weighting $w_i > 0$ yields a valid scoring rule

$$S_n^w = \frac{1}{n}\sum_{i=1}^{n} w_i S(\mathbb{P}_i, y_i).$$

Proper scoring rules

- For a set of observations, we typically use an average score

$$S_n = \frac{1}{n}\sum_{i=1}^{n} S(\mathbb{P}_i, y_i).$$

- But any weighting $w_i > 0$ yields a valid scoring rule

$$S_n^w = \frac{1}{n}\sum_{i=1}^{n} w_i S(\mathbb{P}_i, y_i).$$

Scale dependence

- The predictive measures $\mathbb{P}_i$ and data $\mathbb{Q}_i$ often have varying uncertainty.
- A scoring rule is scale dependent if the magnitude of the value given by the scoring rule at each location can depend on the magnitude of this uncertainty.
- For average scores, the scale dependence will make the different observations in the sum be of different importance.
- The CRPS, MSE, and MAE are all scale dependent measures!
- $$MSE(\mathrm{N}(\mu, \sigma^2), y) = (\mu - y)^2 = \sigma^2 \left(\frac{\mu - y}{\sigma} \right)^2$$
- $$\frac{1}{N} \sum_{i=1}^{N} \mathrm{CRPS}(\mathrm{N}(\mu_i, \sigma_i^2), y_i) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \mathrm{CRPS}\left(\mathrm{N}(0,1), \frac{y_i - \mu_i}{\sigma_i} \right)$$

Scale dependence

- The predictive measures $\mathbb{P}_i$ and data $\mathbb{Q}_i$ often have varying uncertainty.
- A scoring rule is scale dependent if the magnitude of the value given by the scoring rule at each location can depend on the magnitude of this uncertainty.
- For average scores, the scale dependence will make the different observations in the sum be of different importance.
- The CRPS, MSE, and MAE are all scale dependent measures!
- $$MSE(\mathrm{N}(\mu, \sigma^2), y) = (\mu - y)^2 = \sigma^2 \left(\frac{\mu - y}{\sigma} \right)^2$$
- $$\frac{1}{N} \sum_{i=1}^{N} \mathrm{CRPS}(\mathrm{N}(\mu_i, \sigma_i^2), y_i) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \mathrm{CRPS}\left(\mathrm{N}(0,1), \frac{y_i - \mu_i}{\sigma_i} \right)$$

Scale dependence

- The predictive measures $\mathbb{P}_i$ and data $\mathbb{Q}_i$ often have varying uncertainty.
- A scoring rule is scale dependent if the magnitude of the value given by the scoring rule at each location can depend on the magnitude of this uncertainty.
- For average scores, the scale dependence will make the different observations in the sum be of different importance.
- The CRPS, MSE, and MAE are all scale dependent measures!
- $$MSE(\mathrm{N}(\mu, \sigma^2), y) = (\mu - y)^2 = \sigma^2 \left(\frac{\mu - y}{\sigma} \right)^2$$
- $$\frac{1}{N} \sum_{i=1}^{N} \mathrm{CRPS}(\mathrm{N}(\mu_i, \sigma_i^2), y_i) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \mathrm{CRPS}\left(\mathrm{N}(0,1), \frac{y_i - \mu_i}{\sigma_i} \right)$$

Scale dependence

- The predictive measures $\mathbb{P}_i$ and data $\mathbb{Q}_i$ often have varying uncertainty.
- A scoring rule is scale dependent if the magnitude of the value given by the scoring rule at each location can depend on the magnitude of this uncertainty.
- For average scores, the scale dependence will make the different observations in the sum be of different importance.
- The CRPS, MSE, and MAE are all scale dependent measures!

$$MSE(\mathrm{N}(\mu, \sigma^2), y) = (\mu - y)^2 = \sigma^2 \left(\frac{\mu - y}{\sigma} \right)^2$$

$$\frac{1}{N} \sum_{i=1}^{N} \mathrm{CRPS}(\mathrm{N}(\mu_i, \sigma_i^2), y_i) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \mathrm{CRPS} \left(\mathrm{N}(0, 1), \frac{y_i - \mu_i}{\sigma_i} \right)$$

Scale dependence

- The predictive measures $\mathbb{P}_i$ and data $\mathbb{Q}_i$ often have varying uncertainty.
- A scoring rule is scale dependent if the magnitude of the value given by the scoring rule at each location can depend on the magnitude of this uncertainty.
- For average scores, the scale dependence will make the different observations in the sum be of different importance.
- The CRPS, MSE, and MAE are all scale dependent measures!
- $$MSE(\mathsf{N}(\mu, \sigma^2), y) = (\mu - y)^2 = \sigma^2 \left(\frac{\mu - y}{\sigma} \right)^2$$
- $$\frac{1}{N} \sum_{i=1}^{N} \mathrm{CRPS}(\mathsf{N}(\mu_i, \sigma_i^2), y_i) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \mathrm{CRPS}\left(\mathsf{N}(0,1), \frac{y_i - \mu_i}{\sigma_i} \right)$$

Scale dependence

- The predictive measures $\mathbb{P}_i$ and data $\mathbb{Q}_i$ often have varying uncertainty.
- A scoring rule is scale dependent if the magnitude of the value given by the scoring rule at each location can depend on the magnitude of this uncertainty.
- For average scores, the scale dependence will make the different observations in the sum be of different importance.
- The CRPS, MSE, and MAE are all scale dependent measures!
- $$MSE(\mathsf{N}(\mu, \sigma^2), y) = (\mu - y)^2 = \sigma^2 \left(\frac{\mu - y}{\sigma} \right)^2$$
- $$\frac{1}{N} \sum_{i=1}^{N} \mathrm{CRPS}(\mathsf{N}(\mu_i, \sigma_i^2), y_i) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \mathrm{CRPS}\left(\mathsf{N}(0, 1), \frac{y_i - \mu_i}{\sigma_i} \right)$$

Local scale invariance

- Let $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta, \theta = (\mu, \sigma)$, be two distributions with location μ and scale σ.
- One could require that the scoring rule is scale independent, in the sense that $S(\mathbb{P}_\theta, \mathbb{Q}_\theta) = S(\mathbb{P}_{(0,1)}, \mathbb{Q}_{(0,1)})$.
- This is not a good idea! We would then disregard the sharpness of the forecast.
- As an alternative, we propose the concept of local scale invariance:

Definition (Scale function and local scale invariance)

Let S be a proper scoring rule. Assume that there exist a constant $p \in \mathbb{R}$ and a function $s(\mathbb{Q}_\theta) : \mathcal{F} \to \mathbb{R}^+$, such that

$$S(\mathbb{Q}_\theta, \mathbb{Q}_\theta) - S(\mathbb{Q}_{\theta+t\sigma}, \mathbb{Q}_\theta) = s(\mathbb{Q}_\theta)t^p + o(t^p).$$

Then

1. s is the scale function of S.
2. S is locally scale invariant if $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$.

If one makes a small location-scale misspecification proportional to the scale of the true distribution, the difference in the scoring rule should not depend on the scale.

Local scale invariance

- Let $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta, \theta = (\mu, \sigma)$, be two distributions with location μ and scale σ.
- One could require that the scoring rule is scale independent, in the sense that $S(\mathbb{P}_\theta, \mathbb{Q}_\theta) = S(\mathbb{P}_{(0,1)}, \mathbb{Q}_{(0,1)})$.
- This is not a good idea! We would then disregard the sharpness of the forecast.
- As an alternative, we propose the concept of local scale invariance:

Definition (Scale function and local scale invariance)

Let S be a proper scoring rule. Assume that there exist a constant $p \in \mathbb{R}$ and a function $s(\mathbb{Q}_\theta) : \mathcal{F} \to \mathbb{R}^+$, such that

$$S(\mathbb{Q}_\theta, \mathbb{Q}_\theta) - S(\mathbb{Q}_{\theta+t\sigma}, \mathbb{Q}_\theta) = s(\mathbb{Q}_\theta)t^p + o(t^p).$$

Then

1. s is the scale function of S.
2. S is locally scale invariant if $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$.

If one makes a small location-scale misspecification proportional to the scale of the true distribution, the difference in the scoring rule should not depend on the scale.

Local scale invariance

- Let $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta, \theta = (\mu, \sigma)$, be two distributions with location μ and scale σ.
- One could require that the scoring rule is scale independent, in the sense that $S(\mathbb{P}_\theta, \mathbb{Q}_\theta) = S(\mathbb{P}_{(0,1)}, \mathbb{Q}_{(0,1)})$.
- This is not a good idea! We would then disregard the sharpness of the forecast.
- As an alternative, we propose the concept of local scale invariance:

Definition (Scale function and local scale invariance)

Let S be a proper scoring rule. Assume that there exist a constant $p \in \mathbb{R}$ and a function $s(\mathbb{Q}_\theta) : \mathcal{F} \to \mathbb{R}^+$, such that

$$S(\mathbb{Q}_\theta, \mathbb{Q}_\theta) - S(\mathbb{Q}_{\theta+t\sigma}, \mathbb{Q}_\theta) = s(\mathbb{Q}_\theta)t^p + o(t^p).$$

Then

1. s is the scale function of S.
2. S is locally scale invariant if $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$.

If one makes a small location-scale misspecification proportional to the scale of the true distribution, the difference in the scoring rule should not depend on the scale.

Local scale invariance

- Let $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta, \theta = (\mu, \sigma)$, be two distributions with location μ and scale σ.
- One could require that the scoring rule is scale independent, in the sense that $S(\mathbb{P}_\theta, \mathbb{Q}_\theta) = S(\mathbb{P}_{(0,1)}, \mathbb{Q}_{(0,1)})$.
- This is not a good idea! We would then disregard the sharpness of the forecast.
- As an alternative, we propose the concept of local scale invariance:

Definition (Scale function and local scale invariance)

Let S be a proper scoring rule. Assume that there exist a constant $p \in \mathbb{R}$ and a function $s(\mathbb{Q}_\theta) : \mathcal{F} \to \mathbb{R}^+$, such that

$$S(\mathbb{Q}_\theta, \mathbb{Q}_\theta) - S(\mathbb{Q}_{\theta+t\sigma}, \mathbb{Q}_\theta) = s(\mathbb{Q}_\theta)t^p + o(t^p).$$

Then

1. s is the scale function of S.
2. S is locally scale invariant if $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$.

If one makes a small location-scale misspecification proportional to the scale of the true distribution, the difference in the scoring rule should not depend on the scale.

Local scale invariance

- Let $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta, \theta = (\mu, \sigma)$, be two distributions with location μ and scale σ.
- One could require that the scoring rule is scale independent, in the sense that $S(\mathbb{P}_\theta, \mathbb{Q}_\theta) = S(\mathbb{P}_{(0,1)}, \mathbb{Q}_{(0,1)})$.
- This is not a good idea! We would then disregard the sharpness of the forecast.
- As an alternative, we propose the concept of local scale invariance:

Definition (Scale function and local scale invariance)

Let S be a proper scoring rule. Assume that there exist a constant $p \in \mathbb{R}$ and a function $s(\mathbb{Q}_\theta) : \mathcal{F} \to \mathbb{R}^+$, such that

$$S(\mathbb{Q}_\theta, \mathbb{Q}_\theta) - S(\mathbb{Q}_{\theta+t\sigma}, \mathbb{Q}_\theta) = s(\mathbb{Q}_\theta)t^p + o(t^p).$$

Then

1. s is the scale function of S.
2. S is locally scale invariant if $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$.

If one makes a small location-scale misspecification proportional to the scale of the true distribution, the difference in the scoring rule should not depend on the scale.

Local scale invariance

- Let $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta, \theta = (\mu, \sigma)$, be two distributions with location μ and scale σ.
- One could require that the scoring rule is scale independent, in the sense that $S(\mathbb{P}_\theta, \mathbb{Q}_\theta) = S(\mathbb{P}_{(0,1)}, \mathbb{Q}_{(0,1)})$.
- This is not a good idea! We would then disregard the sharpness of the forecast.
- As an alternative, we propose the concept of local scale invariance:

Definition (Scale function and local scale invariance)

Let S be a proper scoring rule. Assume that there exist a constant $p \in \mathbb{R}$ and a function $s(\mathbb{Q}_\theta) : \mathcal{F} \to \mathbb{R}^+$, such that

$$S(\mathbb{Q}_\theta, \mathbb{Q}_\theta) - S(\mathbb{Q}_{\theta+t\sigma}, \mathbb{Q}_\theta) = s(\mathbb{Q}_\theta)t^p + o(t^p).$$

Then

1. s is the scale function of S.
2. S is locally scale invariant if $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$.

If one makes a small location-scale misspecification proportional to the scale of the true distribution, the difference in the scoring rule should not depend on the scale.

Local scale invariance

It is easy to show that

- The log score is locally scale invariant, $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$
- The CRPS has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MAE has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MSE has $s(\mathbb{Q}_\theta) = \sigma^2 s(\mathbb{Q})$

Question

Is there a locally scale invariant version of the CRPS?

Local scale invariance

It is easy to show that

- The log score is locally scale invariant, $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$
- The CRPS has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MAE has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MSE has $s(\mathbb{Q}_\theta) = \sigma^2 s(\mathbb{Q})$

Question

Is there a locally scale invariant version of the CRPS?

Local scale invariance

It is easy to show that

- The log score is locally scale invariant, $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$
- The CRPS has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MAE has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MSE has $s(\mathbb{Q}_\theta) = \sigma^2 s(\mathbb{Q})$

Question

Is there a locally scale invariant version of the CRPS?

Local scale invariance

It is easy to show that

- The log score is locally scale invariant, $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$
- The CRPS has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MAE has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MSE has $s(\mathbb{Q}_\theta) = \sigma^2 s(\mathbb{Q})$

Question

Is there a locally scale invariant version of the CRPS?

Local scale invariance

It is easy to show that

- The log score is locally scale invariant, $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$
- The CRPS has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MAE has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MSE has $s(\mathbb{Q}_\theta) = \sigma^2 s(\mathbb{Q})$

Question

Is there a locally scale invariant version of the CRPS?

Local scale invariance

It is easy to show that

- The log score is locally scale invariant, $s(\mathbb{Q}_\theta) \equiv s(\mathbb{Q})$
- The CRPS has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MAE has $s(\mathbb{Q}_\theta) = \sigma s(\mathbb{Q})$
- The MSE has $s(\mathbb{Q}_\theta) = \sigma^2 s(\mathbb{Q})$

Question

Is there a locally scale invariant version of the CRPS?

Kernel scores

The CRPS is a special case of the larger class of kernel scoring rules.

Theorem (Gneiting and Raftery, 2007)

Let $\mathbb{P}$ be a Borel probability. Assume that g is a non-negative, continuous negative definite kernel then

$$S_g^{ker}(\mathbb{P}, y) := \frac{1}{2}\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X, Y)\right] - \mathsf{E}_{\mathbb{P}}\left[g(X, y)\right]$$

is a proper scoring rule on $\mathcal{P}$.

One example of a family of negative definite kernels that can be used is $g_\alpha(x, y) = |x - y|^\alpha$ for $\alpha \in (0, 2]$. The CRPS is the special case with $\alpha = 1$.

Kernel scores

The CRPS is a special case of the larger class of kernel scoring rules.

Theorem (Gneiting and Raftery, 2007)

Let $\mathbb{P}$ be a Borel probability. Assume that g is a non-negative, continuous negative definite kernel then

$$S_g^{ker}(\mathbb{P}, y) := \frac{1}{2}\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X, Y)\right] - \mathsf{E}_{\mathbb{P}}\left[g(X, y)\right]$$

is a proper scoring rule on $\mathcal{P}$.

One example of a family of negative definite kernels that can be used is $g_\alpha(x, y) = |x - y|^\alpha$ for $\alpha \in (0, 2]$. The CRPS is the special case with $\alpha = 1$.

Generalized Kernel scores

Theorem (B. and Wallin)

Let $\mathbb{P}$ be a Borel probability. Assume that g is a non-negative, continuous negative definite kernel and that $h > 0$ is a monotonically decreasing convex differentiable function. Then the scoring rule

$$S_g^h(\mathbb{P}, y) := h\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right) + \\ +2h'\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)\left(\mathsf{E}_{\mathbb{P}}\left[g(X,y)\right] - \mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)$$

- With $h(x) = -x/2$ we obtain the usual kernel scores:

$$S_g^h(\mathbb{P}, y) := h\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right) + \\ +2h'\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)\left(\mathsf{E}_{\mathbb{P}}\left[g(X,y)\right] - \mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)$$

- and $g(x,y) = |x-y|$ we obtain the CRPS.

Generalized Kernel scores

Theorem (B. and Wallin)

Let $\mathbb{P}$ be a Borel probability. Assume that g is a non-negative, continuous negative definite kernel and that $h > 0$ is a monotonically decreasing convex differentiable function. Then the scoring rule

$$S_g^h(\mathbb{P}, y) := h\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right) + \\ +2h'\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)\left(\mathsf{E}_{\mathbb{P}}\left[g(X,y)\right] - \mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)$$

- With $h(x) = -x/2$ we obtain the usual kernel scores:

$$S_g^h(\mathbb{P}, y) := h\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right) + \\ +2h'\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)\left(\mathsf{E}_{\mathbb{P}}\left[g(X,y)\right] - \mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right]\right)$$

- and $g(x,y) = |x-y|$ we obtain the CRPS.

The standardized kernel scores

- With $h(x) = -\log(x)/2$ we obtain the scoring rule

$$S_g^{-\frac{1}{2}\log(x)}(\mathbb{P}, y) = -\frac{1}{2}\log(\mathsf{E}_{\mathbb{P},\mathbb{P}}[g(X,Y)]) - \frac{\mathsf{E}_{\mathbb{P}}[g(X,y)]}{\mathsf{E}_{\mathbb{P},\mathbb{P}}[g(X,Y)]} + 1.$$

  which we refer to as a standardized kernel score.

- Choosing $g(x,y) = |x-y|^2$ gives us the Dawid-Sebastiani score

$$S_2^{\text{sta}}(\mathbb{P}, y) = -\frac{(y - \mathsf{E}_{\mathbb{P}}[Y])^2}{\mathsf{V}_{\mathbb{P}}[Y]} - \frac{1}{2}\log\left(\mathsf{V}_{\mathbb{P}}[Y]\right).$$

- Choosing $g(x,y) = |x-y|$ gives us a standardized version of the CRPS, which we refer to as the SCRPS:

$$\text{SCRPS}(\mathbb{P}, y) := -\frac{\mathsf{E}_{\mathbb{P}}[|X-y|]}{\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X-Y|]} - \frac{1}{2}\log\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X-Y|]\right).$$

The standardized kernel scores

- With $h(x) = -\log(x)/2$ we obtain the scoring rule

$$S_g^{-\frac{1}{2}\log(x)}(\mathbb{P}, y) = -\frac{1}{2}\log(\mathsf{E}_{\mathbb{P},\mathbb{P}}[g(X,Y)]) - \frac{\mathsf{E}_{\mathbb{P}}[g(X,y)]}{\mathsf{E}_{\mathbb{P},\mathbb{P}}[g(X,Y)]} + 1.$$

  which we refer to as a standardized kernel score.
- Choosing $g(x,y) = |x-y|^2$ gives us the Dawid-Sebastiani score

$$S_2^{\mathsf{sta}}(\mathbb{P}, y) = -\frac{(y - \mathsf{E}_{\mathbb{P}}[Y])^2}{\mathsf{V}_{\mathbb{P}}[Y]} - \frac{1}{2}\log\left(\mathsf{V}_{\mathbb{P}}[Y]\right).$$

- Choosing $g(x,y) = |x-y|$ gives us a standardized version of the CRPS, which we refer to as the SCRPS:

$$\mathrm{SCRPS}(\mathbb{P}, y) := -\frac{\mathsf{E}_{\mathbb{P}}[|X-y|]}{\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X-Y|]} - \frac{1}{2}\log\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X-Y|]\right).$$

The standardized kernel scores

- With $h(x) = -\log(x)/2$ we obtain the scoring rule

$$S_g^{-\frac{1}{2}\log(x)}(\mathbb{P}, y) = -\frac{1}{2}\log(\mathsf{E}_{\mathbb{P},\mathbb{P}}[g(X,Y)]) - \frac{\mathsf{E}_{\mathbb{P}}[g(X,y)]}{\mathsf{E}_{\mathbb{P},\mathbb{P}}[g(X,Y)]} + 1.$$

  which we refer to as a standardized kernel score.
- Choosing $g(x,y) = |x-y|^2$ gives us the Dawid-Sebastiani score

$$S_2^{\mathsf{sta}}(\mathbb{P}, y) = -\frac{(y - \mathsf{E}_{\mathbb{P}}[Y])^2}{\mathsf{V}_{\mathbb{P}}[Y]} - \frac{1}{2}\log\left(\mathsf{V}_{\mathbb{P}}[Y]\right).$$

- Choosing $g(x,y) = |x-y|$ gives us a standardized version of the CRPS, which we refer to as the SCRPS:

$$\mathrm{SCRPS}(\mathbb{P}, y) := -\frac{\mathsf{E}_{\mathbb{P}}[|X-y|]}{\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X-Y|]} - \frac{1}{2}\log\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X-Y|]\right).$$

The SCRPS

$$\mathrm{CRPS}(\mathbb{P}, y) = \frac{1}{2}\mathsf{E}_{\mathbb{P},\mathbb{P}}[|X - Y|] - \mathsf{E}_{\mathbb{P}}[|X - y|]$$

$$\mathrm{SCRPS}(\mathbb{P}, y) := -\frac{\mathsf{E}_{\mathbb{P}}\left[|X - y|\right]}{\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[|X - Y|\right]} - \frac{1}{2}\log\left(\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[|X - Y|\right]\right)$$

Important advantages:

- Locally scale invariant.
- As easy to evaluate as the CRPS. (Arguably as easy to interpret as well.)
- Is strictly proper over the same class of measures as CRPS.

Two-week epidemic forecasts

Model	Week	μ	s	SD	Obs. y	Interpretation
A	1	18	80	4.68	20	low week, close and precise
A	2	320	6	136.63	400	peak week, low-biased and uncertain
B	1	10	80	3.61	20	low week, too low and precise
B	2	380	6	157.48	400	peak week, close and uncertain

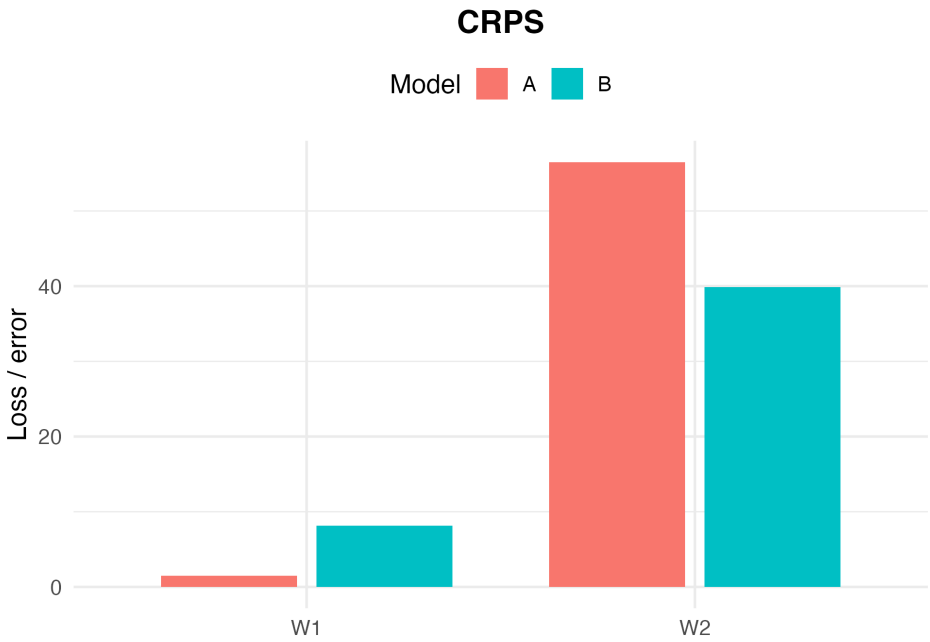
- Raw errors:

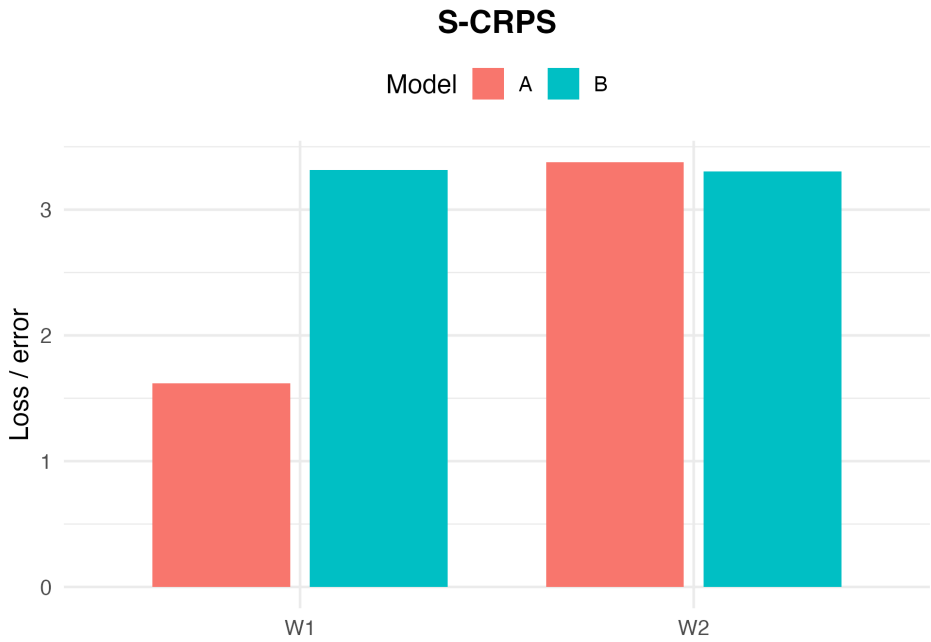
$$\text{A: } (|20-18|, |400-320|) = (2, 80), \qquad \text{B: } (|20-10|, |400-380|) = (10, 20).$$

- Standardized errors (approximately):

$$\text{A: } \left(\frac{2}{4.68}, \frac{80}{136.63}\right) \approx (0.43, 0.59), \qquad \text{B: } \left(\frac{10}{3.61}, \frac{20}{157.48}\right) \approx (2.77, 0.13).$$

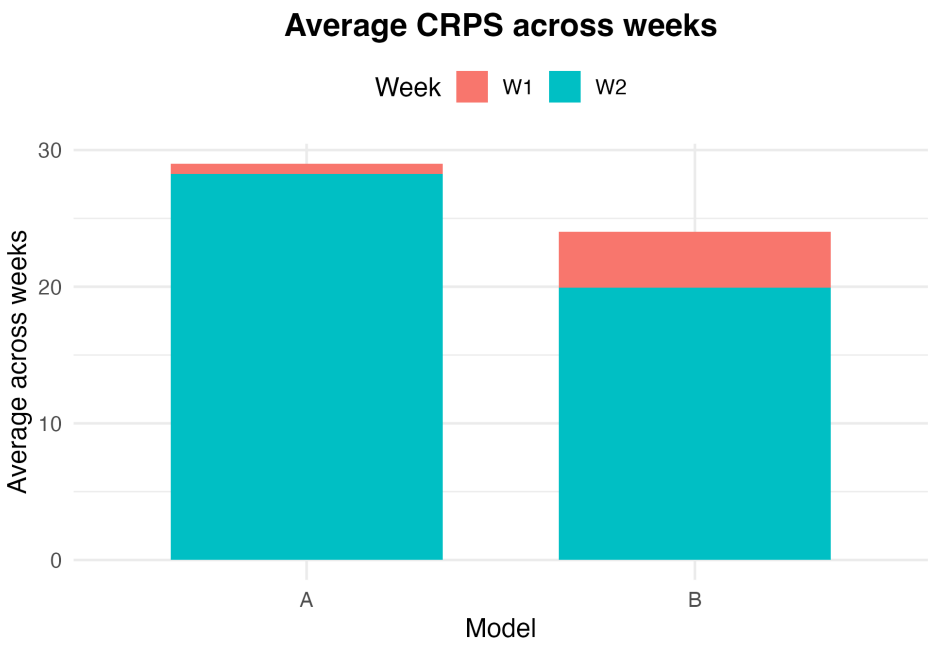
CRPS and S-CRPS by week

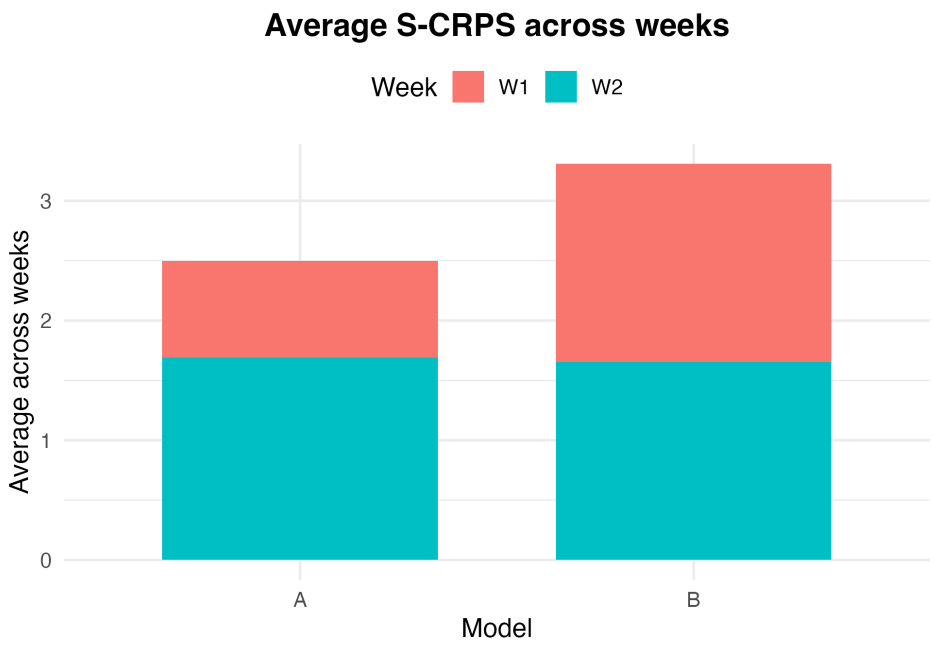




Lower values indicate better probabilistic forecasts.

CRPS and S-CRPS average





Lower values indicate better probabilistic forecasts.

CRPS vs. SCRPS

Score	Epidemic interpretation	Tends to reward
CRPS	absolute error on the **case-count scale**	models that get **large peaks** right
SCRPS	count error **scaled by forecast uncertainty**	models that are fair across **low and high incidence**

Introducing log-CRPS

- Bosse, N. I., Abbott, S., Cori, A., van Leeuwen, E., Bracher, J. and Funk, S. (2023). *Scoring epidemiological forecasts on transformed scales*. *PLOS Computational Biology*, 19(8), e1011393.
- For a positive-valued forecast $\mathbb{P}$ and observation $y > 0$, define

$$\text{log-CRPS}(\mathbb{P}, y) := \text{CRPS}(\mathbb{P}^{\log}, \log y),$$

  where $\mathbb{P}^{\log}$ is the law of $\log X$ if $X \sim \mathbb{P}$.
- Equivalently,

$$\text{log-CRPS}(\mathbb{P}, y) = \frac{1}{2}\,\mathbb{E}|\log X - \log X'| - \mathbb{E}|\log X - \log y|.$$

$s_{\log \mathrm{CRPS}}(u, \sigma)$ for $Y = u + \sigma Z$

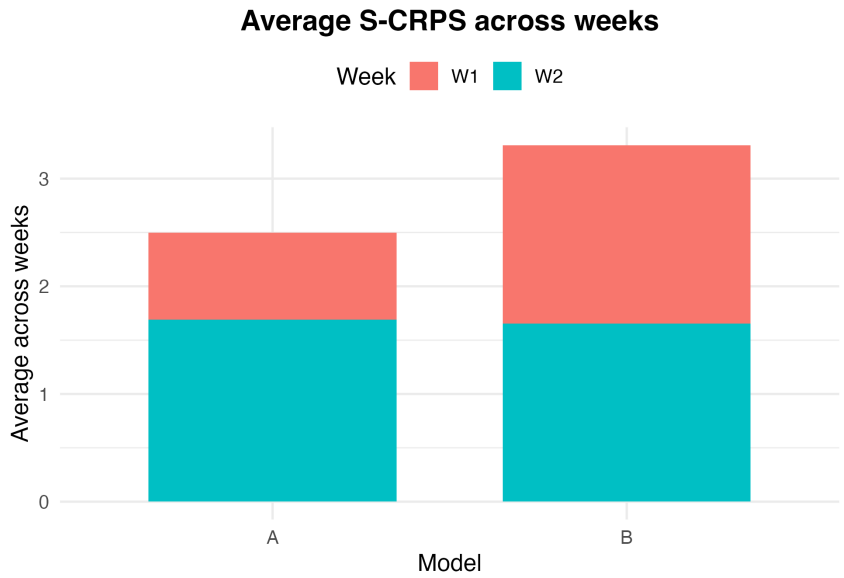
One can show that

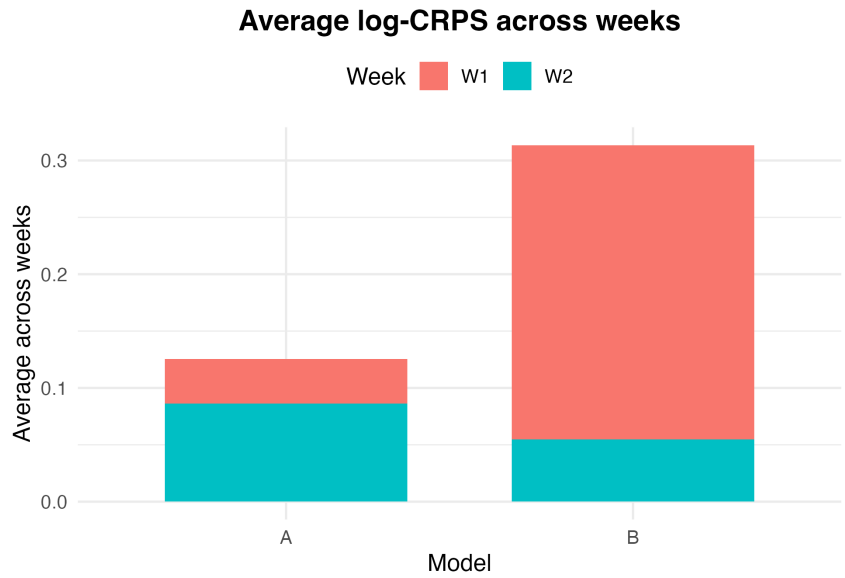
$$s_{\log \mathrm{CRPS}}(\mathbb{Q}_\theta) = B(u/\sigma)$$

So at least for $\mu = 0$ it is scale invariant.

- Thus log-transformation handles the scale dependence.
- The difference is that SCRPS scales with predictive uncertainty, while log-CRPS scales with the ratio of the location to the scale.

log-CRPS average





- Raw errors:

  A Raw: $(2, 80)$, Log: $(0.11,\ 0.22)$.

  B Raw: $(10, 20)$, Log: $(0.69,\ 0.05)$.

- Standardized errors (approximately):

  A: $\left(\frac{2}{4.68}, \frac{80}{136.63}\right) \approx (0.43, 0.59)$, B: $\left(\frac{10}{3.61}, \frac{20}{157.48}\right) \approx (2.77, 0.13)$.

Epidemic example: model and forecasts

Observed weekly counts

$$y_1 = 20 \qquad \text{(low-incidence week)}, \qquad y_2 = 400 \qquad \text{(peak week)}.$$

Forecast model

For each week $t \in \{1, 2\}$, let the predictive distribution be

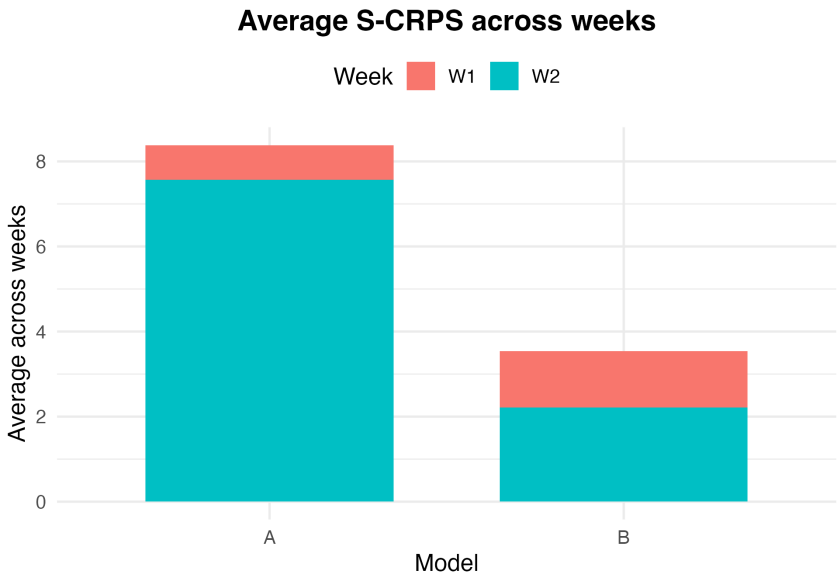
$$Y_t \sim \mathcal{N}(\mu_t, s^2),$$

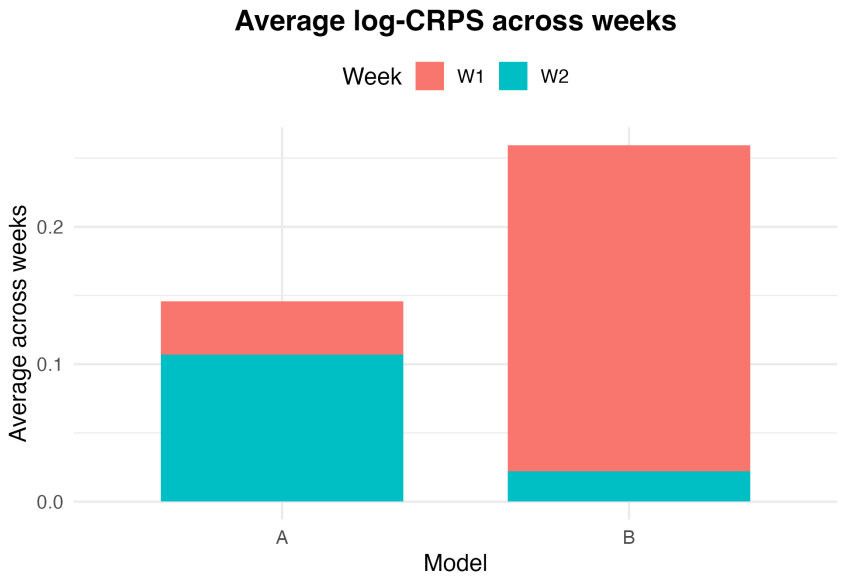
with

$$\mathbb{E}(Y_t) = \mu_t, \qquad \mathrm{Var}(Y_t) = s^2, \qquad \mathrm{SD}(Y_t) = s.$$

Here, the predictive standard deviation s is taken to be the same in both weeks, so the forecast variance is constant across time.

Average S-CRPS and log-CRPS





- Raw errors:

A Raw: $(2, 80)$, Log: $(0.11,\ 0.22)$.

B Raw: $(10, 20)$, Log: $(0.69,\ 0.05)$.

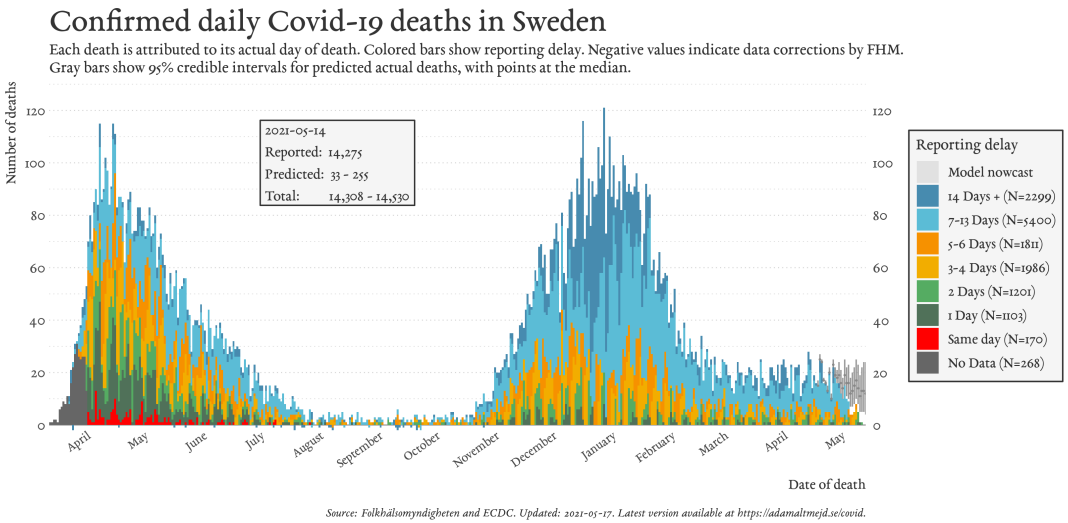
- Standardized errors ($s = 5$):

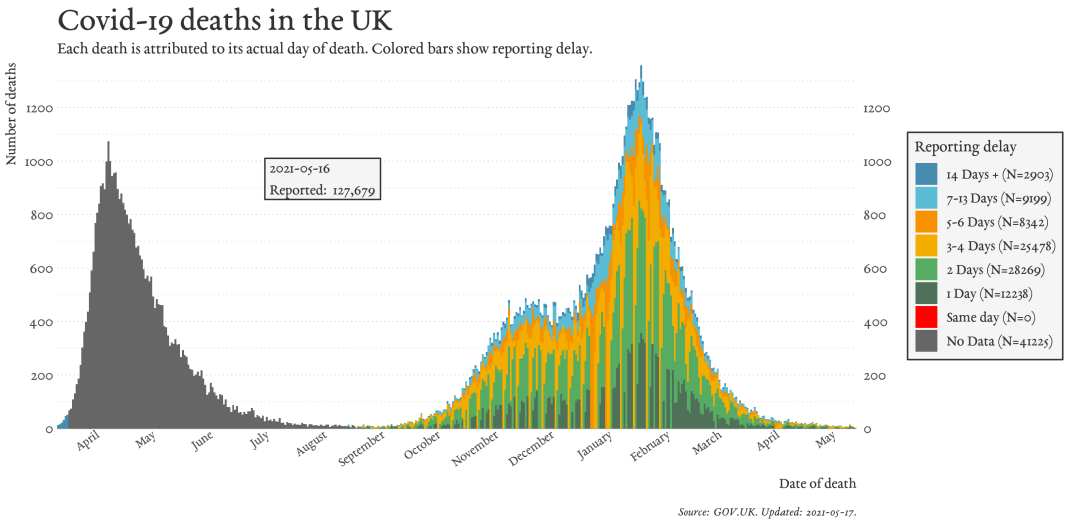
$$\text{A: } \left(\frac{2}{5}, \frac{80}{5}\right) = (0.4, 16), \qquad \text{B: } \left(\frac{10}{5}, \frac{20}{5}\right) = (2, 4).$$

Original-scale vs. log-scale scores

Score	Epidemic interpretation	Tends to reward
CRPS	absolute error on the **case-count scale**	models that get **large peaks** right
SCRPS	count error **scaled by forecast uncertainty**	models that are equal across **low and high incidence**
log-CRPS	error on the **log scale**, that is, in **relative / multiplicative** terms	models that get **relative error** or **growth** right
log-SCRPS	log-scale error **scaled by log-scale uncertainty**	models that are equal on the **log scale**

Nowcasting COVID deaths in Sweden and UK





Model

- Underlying (simplified) model:

$$\begin{aligned} X(t) &\sim GP(t; \theta_{GP}), \\ \lambda_t &= \exp(X(t)), \\ dea_t &\sim Poisson(\lambda_t), \\ y_{t,t:T} &\sim Binomal(dea_t, p_{t:T}). \end{aligned}$$

- dea_t latent variable of interest. We assume known at day $D + t$.
- Ideally we would like to weigh the observations with underlying number at risk:

$$S(\mathbb{P}, \mathbf{dea}) \approx \frac{1}{T} \sum_{i=1}^{T} \exp(X(i)) S(\mathbb{P}_i, dea_i)$$

  where S is locally scale invariant scoring rule.
- $X(t)$ not observed hence can't be used.
- Therefore better to CRPS where we weigh the observations implicitly through scale dependence $w_i \approx \exp(X(i)/2)$.

Model

- Underlying (simplified) model:

$$\begin{aligned} X(t) &\sim GP(t; \theta_{GP}), \\ \lambda_t &= \exp(X(t)), \\ dea_t &\sim Poisson(\lambda_t), \\ y_{t,t:T} &\sim Binomal(dea_t, p_{t:T}). \end{aligned}$$

- dea_t latent variable of interest. We assume known at day $D + t$.
- Ideally we would like to weigh the observations with underlying number at risk:

$$S(\mathbb{P}, \mathbf{dea}) \approx \frac{1}{T} \sum_{i=1}^{T} \exp(X(i)) S(\mathbb{P}_i, dea_i)$$

  where S is locally scale invariant scoring rule.
- $X(t)$ not observed hence can't be used.
- Therefore better to CRPS where we weigh the observations implicitly through scale dependence $w_i \approx \exp(X(i)/2)$.

Model

- Underlying (simplified) model:

$$\begin{aligned} X(t) &\sim GP(t; \theta_{GP}), \\ \lambda_t &= \exp(X(t)), \\ dea_t &\sim Poisson(\lambda_t), \\ y_{t,t:T} &\sim Binomal(dea_t, p_{t:T}). \end{aligned}$$

- dea_t latent variable of interest. We assume known at day $D + t$.
- Ideally we would like to weigh the observations with underlying number at risk:

$$S(\mathbb{P}, \mathbf{dea}) \approx \frac{1}{T} \sum_{i=1}^{T} \exp(X(i)) S(\mathbb{P}_i, dea_i)$$

  where S is locally scale invariant scoring rule.
- $X(t)$ not observed hence can't be used.
- Therefore better to CRPS where we weigh the observations implicitly through scale dependence $w_i \approx \exp(X(i)/2)$.

Model

- Underlying (simplified) model:

$$\begin{aligned} X(t) &\sim GP(t; \theta_{GP}), \\ \lambda_t &= \exp(X(t)), \\ dea_t &\sim Poisson(\lambda_t), \\ y_{t,t:T} &\sim Binomal(dea_t, p_{t:T}). \end{aligned}$$

- dea_t latent variable of interest. We assume known at day $D + t$.
- Ideally we would like to weigh the observations with underlying number at risk:

$$S(\mathbb{P}, \mathbf{dea}) \approx \frac{1}{T} \sum_{i=1}^{T} \exp(X(i)) S(\mathbb{P}_i, dea_i)$$

  where S is locally scale invariant scoring rule.
- $X(t)$ not observed hence can't be used.
- Therefore better to CRPS where we weigh the observations implicitly through scale dependence $w_i \approx \exp(X(i)/2)$.

Model

- Underlying (simplified) model:

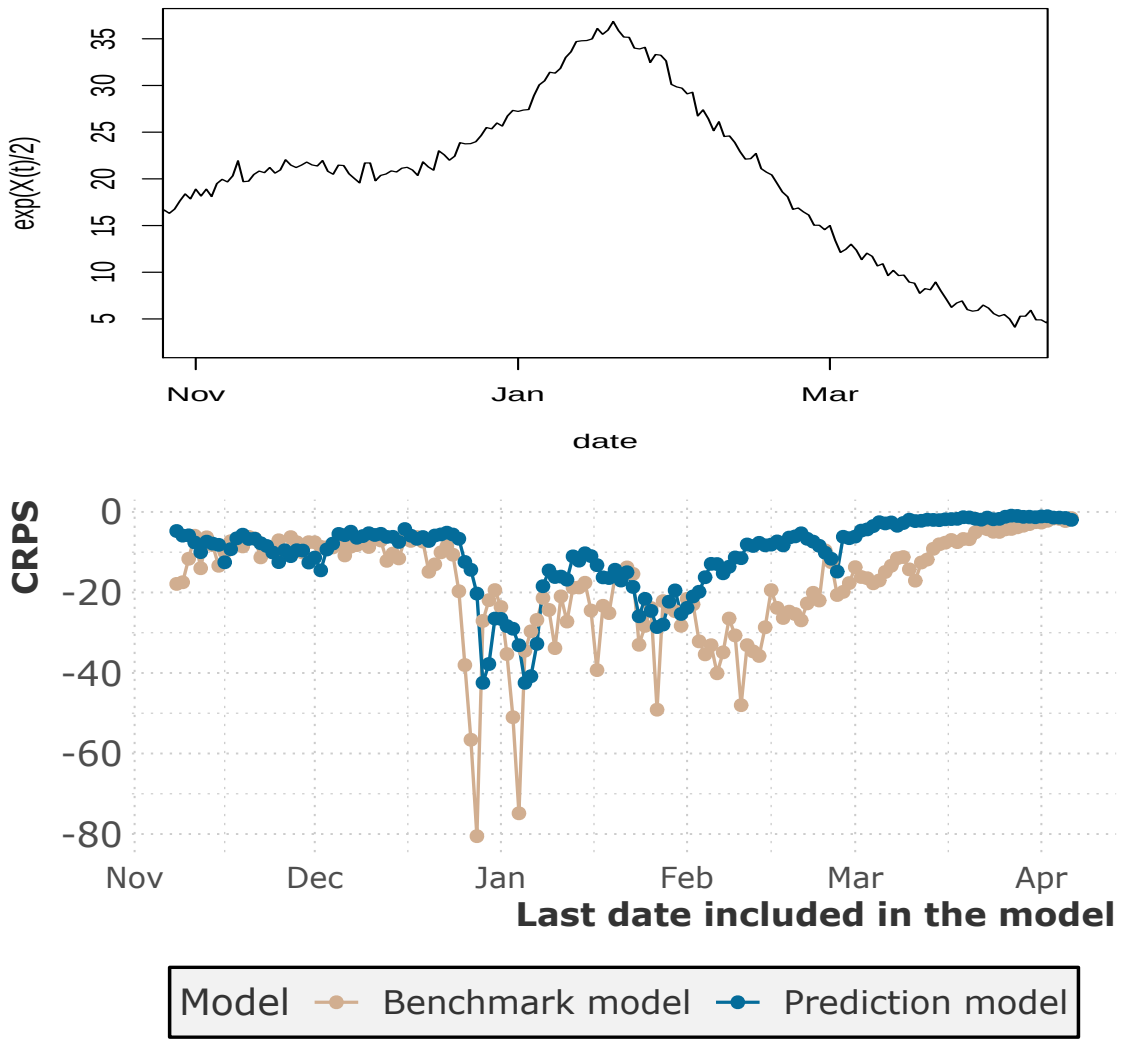
$$\begin{aligned} X(t) &\sim GP(t; \theta_{GP}), \\ \lambda_t &= \exp(X(t)), \\ dea_t &\sim Poisson(\lambda_t), \\ y_{t,t:T} &\sim Binomal(dea_t, p_{t:T}). \end{aligned}$$

- dea_t latent variable of interest. We assume known at day $D + t$.
- Ideally we would like to weigh the observations with underlying number at risk:

$$S(\mathbb{P}, \mathbf{dea}) \approx \frac{1}{T} \sum_{i=1}^{T} \exp(X(i)) S(\mathbb{P}_i, dea_i)$$

  where S is locally scale invariant scoring rule.
- $X(t)$ not observed hence can't be used.
- Therefore better to CRPS where we weigh the observations implicitly through scale dependence $w_i \approx \exp(X(i)/2)$.

Nowcasting COVID deaths in Sweden and UK



Discussion

- The correct choice of scoring rule is hard.

scale function of WIS

For the quantile representation

$$\mathrm{WIS}(F, y) = \frac{1}{K} \sum_{k=1}^{K} 2\Big(\mathbf{1}\{y \le q_{\tau_k}\} - \tau_k\Big)(q_{\tau_k} - y),$$

a location–scale transform $Y_{\mu,\sigma} = \mu + \sigma Y$ gives

$$q_{\tau_k}^{(\mu,\sigma)} = \mu + \sigma q_{\tau_k}, \qquad \mathrm{WIS}(F_{\mu,\sigma}, \mu + \sigma z) = \sigma\, \mathrm{WIS}(F, z).$$

Hence WIS is homogeneous of order 1, just like CRPS. Therefore, under the local perturbation $\theta + t\sigma$,

$$S_{\mathrm{WIS}}(\mathbb{Q}_\theta, \mathbb{Q}_\theta) - S_{\mathrm{WIS}}(\mathbb{Q}_{\theta+t\sigma}, \mathbb{Q}_\theta) = \sigma\, s_{\mathrm{WIS}}(\mathbb{Q}) t^p + o(t^p),$$

so

$$s_{\mathrm{WIS}}(\mathbb{Q}_\theta) = \sigma\, s_{\mathrm{WIS}}(\mathbb{Q}).$$