Scaling of scoring rules

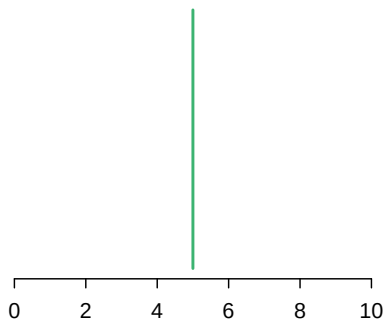
Jonas Wallin
joint work with
David Bolin (KAUST)

CIRM virtual conference
2020-06-02

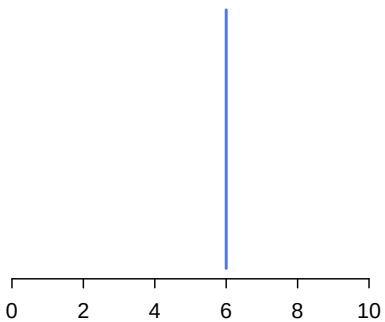
LUND
UNIVERSITY

Forecast and observation classes

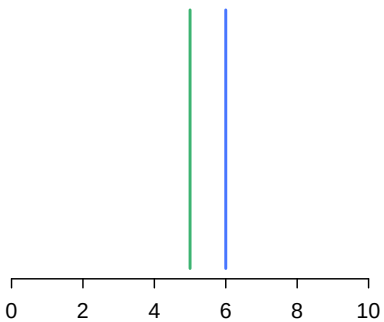
(a) Forecast



(b) Observation



(c) Comparison



Scoring functions apply to deterministic forecasts

The forecast x is evaluated against the observation y using **scoring functions** such as

negative Squared Error (SE)	$\mathrm{S}(x, y) = -(x - y)^2$
negative Absolute Error (AE)	$\mathrm{S}(x, y) = -\lvert x - y \rvert$

Bayes predictors should be used for probilistic forecasts

For a probabilistic forecast $\mathbb{P}$, **decision theory** tells us that if the scoring function S is given, we should issue the **Bayes predictor**,

$$\hat{x} = \arg\min_x \mathsf{E}_{\mathbb{P}}\left[\mathrm{S}(x, Y)\right]$$

as the point forecast, where the expectation is with respect to $\mathbb{P}$.

Squared Error (SE)	$\mathrm{S}(x, y) = -(x-y)^2$	$\hat{x} = \mathsf{mean}(\mathbb{P})$
Absolute Error (AE)	$\mathrm{S}(x, y) = -\lvert x-y\rvert$	$\hat{x} = \mathsf{median}(\mathbb{P})$

The basic idea

Assume we have a prediction $p \in \mathcal{P}$ and an observation $o \in \mathcal{O}$ where we wish to measure the skill of the prediction by applying a function

$$s : \mathcal{P} \times \mathcal{O} \longrightarrow \mathbb{R}$$

with a higher function value indicating a better skill.

- What are good theoretical properties for s?

General framework without any formulas...

- Assume $\mathbb{Q}$ is Nature's distribution of some event y and denote our forecast for y by $\mathbb{P}$.

General framework without any formulas...

- Assume $\mathbb{Q}$ is Nature's distribution of some event y and denote our forecast for y by $\mathbb{P}$.
- For forecast evaluation, we should use performance metrics that follow the principle

in the long run, we will obtain the optimal performance for $\mathbb{P} = \mathbb{Q}$

Probabilistic forecasts should generally be evaluated using proper scoring rules

A consistent scoring function is a special case of a proper scoring rule for probabilistic forecasts

Definition (Murphy and Winkler, 1968)

If $\mathcal{F}$ denotes a class of probabilistic forecasts on $\mathbb{R}$, a **proper scoring rule** is any function

$$\mathrm{S} : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$$

such that

$$\mathrm{S}(\mathbb{Q}, \mathbb{Q}) := \mathrm{E}_{\mathbb{Q}}\, \mathrm{S}(\mathbb{Q}, Y) \geq \mathrm{E}_{\mathbb{Q}}\, \mathrm{S}(\mathbb{P}, Y) =: \mathrm{S}(\mathbb{P}, \mathbb{Q})$$

for all $\mathbb{P}, \mathbb{Q} \in \mathcal{F}$.

The class of proper scoring rules is large

$$\mathrm{S}(\mathbb{P}, y) = -(\mathsf{mean}(\mathbb{P}) - y)^2$$
$$\mathrm{S}(\mathbb{P}, y) = -|\mathsf{median}(\mathbb{P}) - y|$$

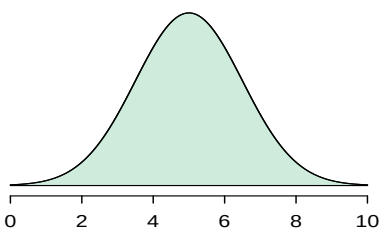
Gneiting, T. and Raftery, A.E. (2007): Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102, 359-178.

Optimally, forecasts should be probabilistic

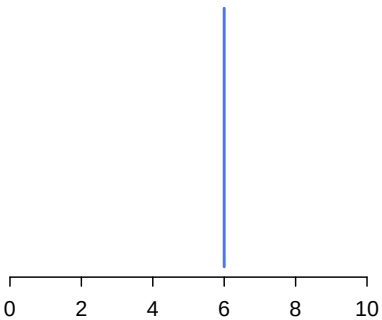
All those whose duty it is to issue regular daily forecasts know that there are times when they feel very confident and other times when they are doubtful as to coming weather. It seems to me that the condition of confidence or otherwise forms a very important part of the prediction.

Cooke (Monthly Weather Review, 1906)

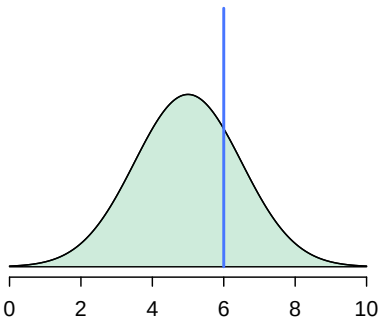
(d) Forecast



(e) Observation



(f) Comparison



The class of proper scoring rules is large

The perhaps the two most common proper scoring rule is the **continuous ranked probability score (CRPS)**

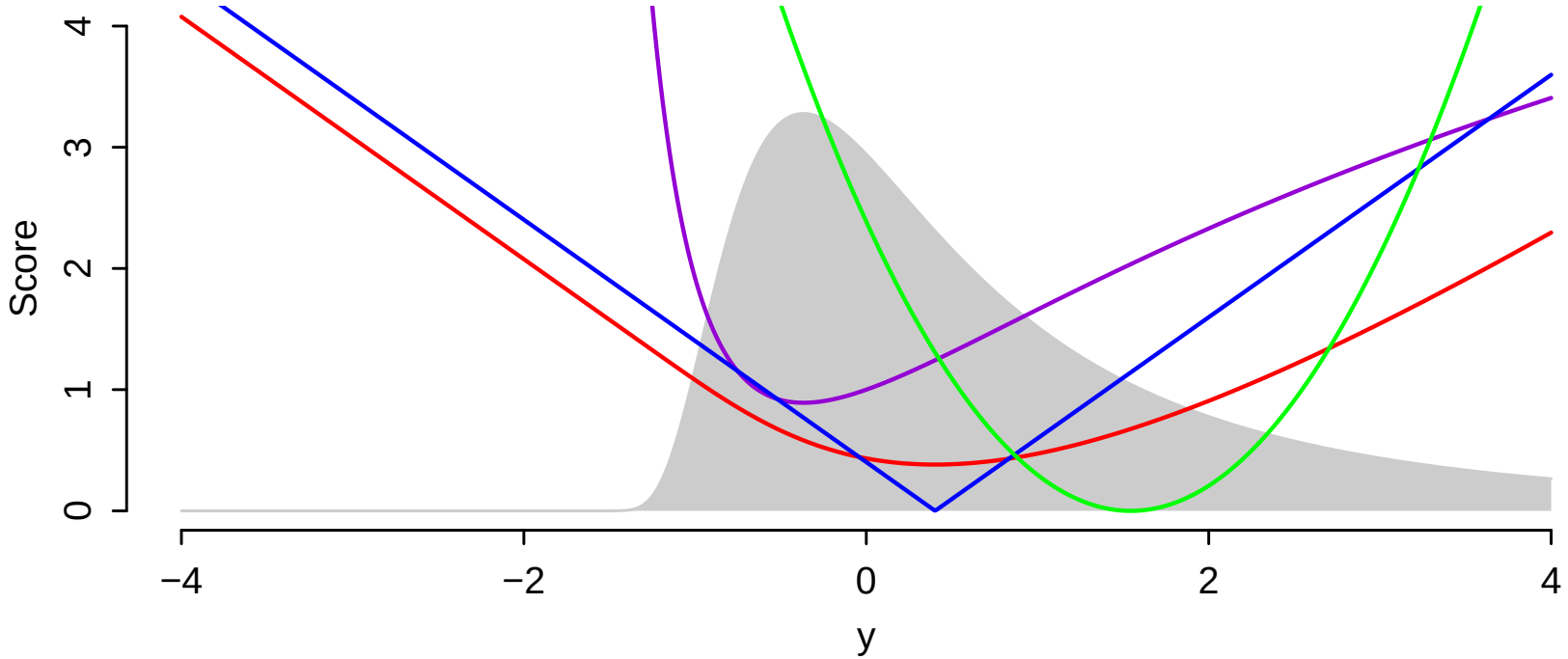
$$\mathrm{S}(\mathbb{P}, y) = -\mathsf{E}_{\mathbb{P}}|X - y| + \frac{1}{2}\mathsf{E}_{\mathbb{P}}\mathsf{E}_{\mathbb{P}}|X - X'|$$

and the log score

$$\mathrm{S}(\mathbb{P}, y) = -\log(f(y)),$$

Gneiting, T. and Raftery, A.E. (2007): Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102, 359-178.

The different scores behave somewhat differently



SE AE CRPS IGN

Average scores facilitate comparison across methods

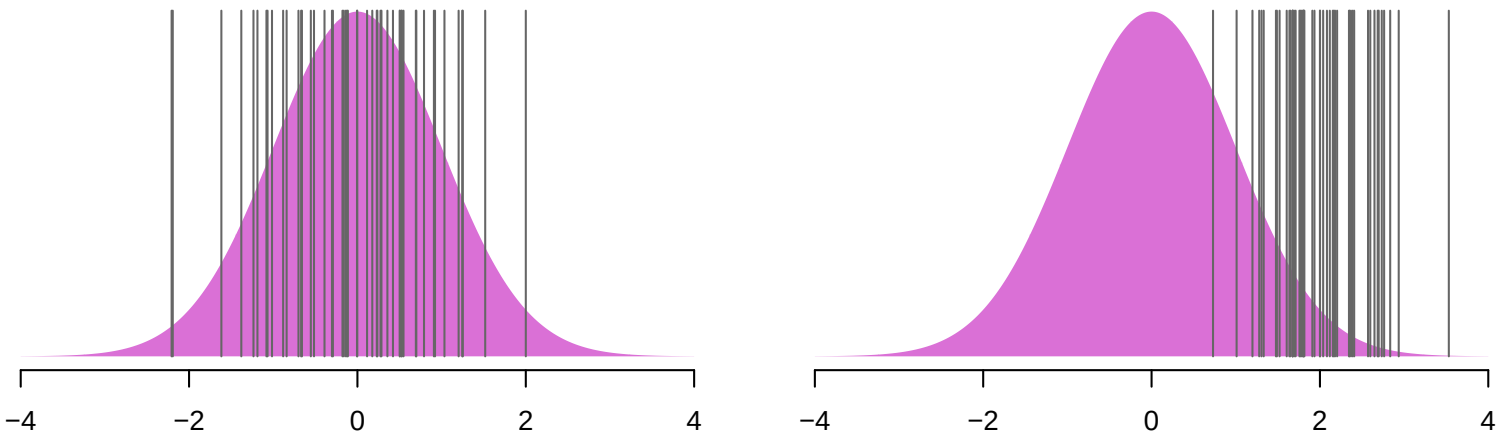
Assume we have two forecasting methods $m = 1, 2$.

They issue point forecasts $\mathbb{P}_{mi}$ with observed values y_i, at a finite set of times, locations or instances $i = 1, \ldots, n$

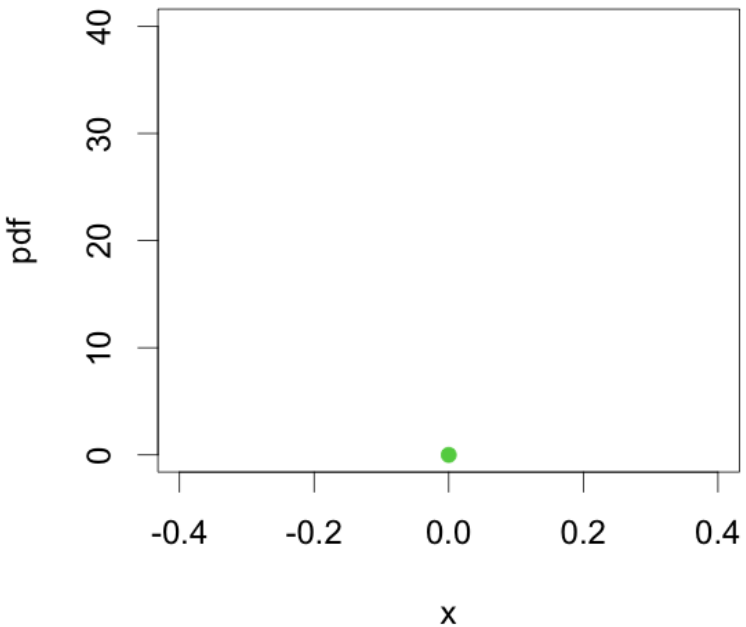
The methods are assessed and ranked by the mean score (our contribution starts here)

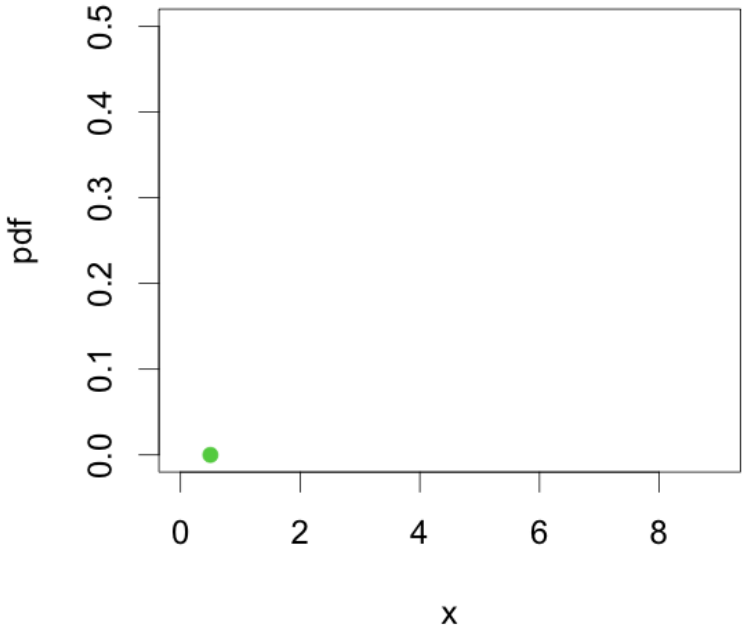
$$\bar{\mathrm{S}}_n^m = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(\mathbb{P}_{mi}, y_i) \qquad \text{for } m = 1, 2.$$

Average scores facilitate comparison across methods

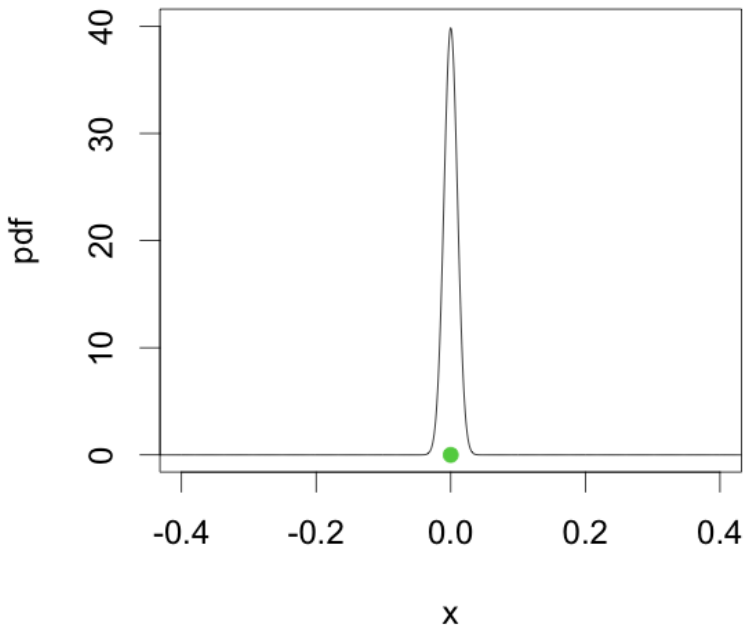


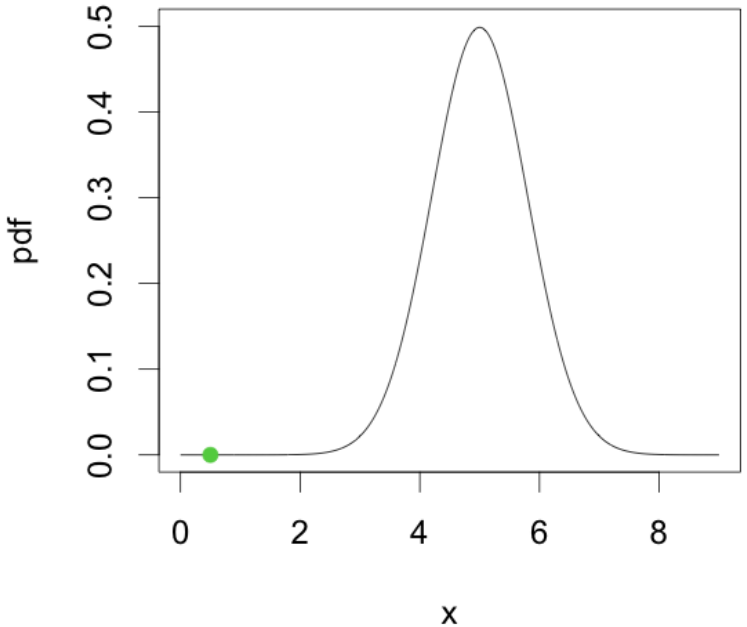
Two observations, two models



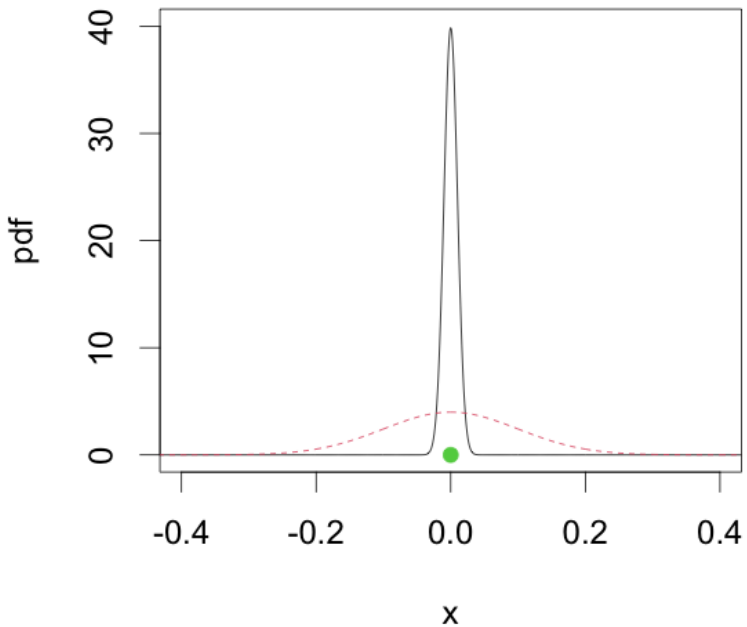


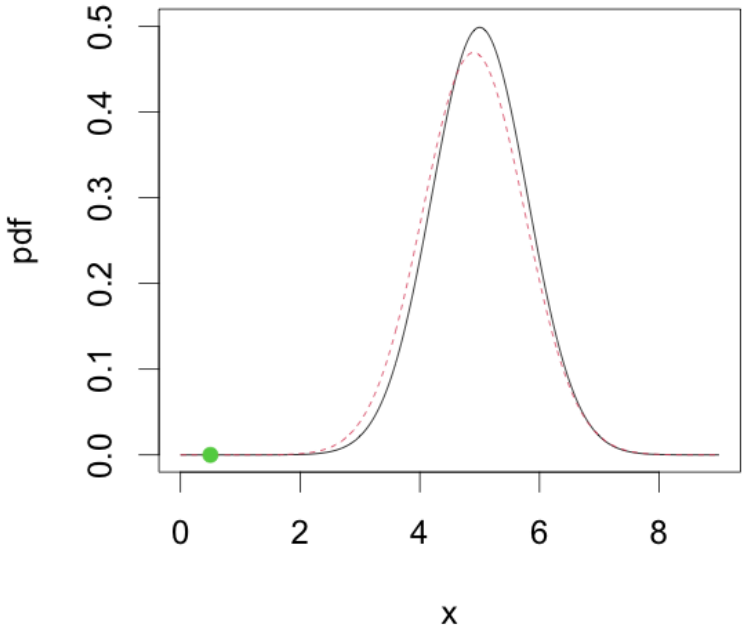
Two observations, two models





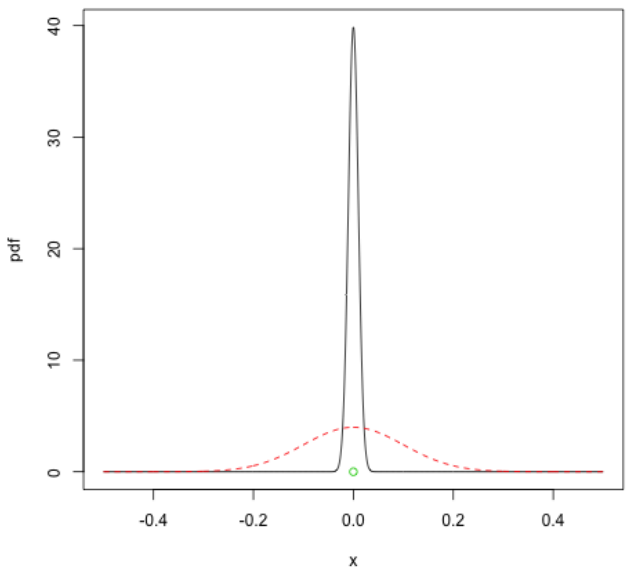
Two observations, two models

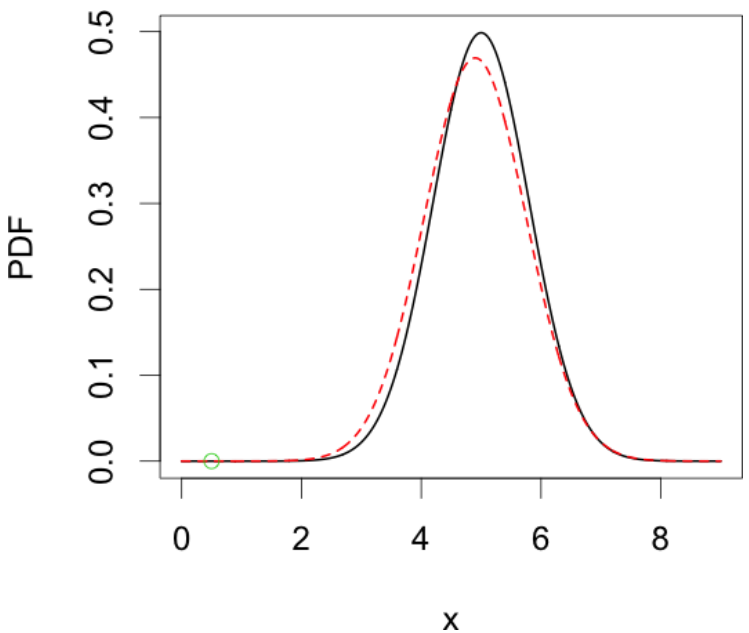




Two observations, two models, result

	Model 1 CRPS	Model 2 CRPS
Y_1	0.0023	0.02346
Y_2	4.0486	3.920
mean	2.0255	1.9719





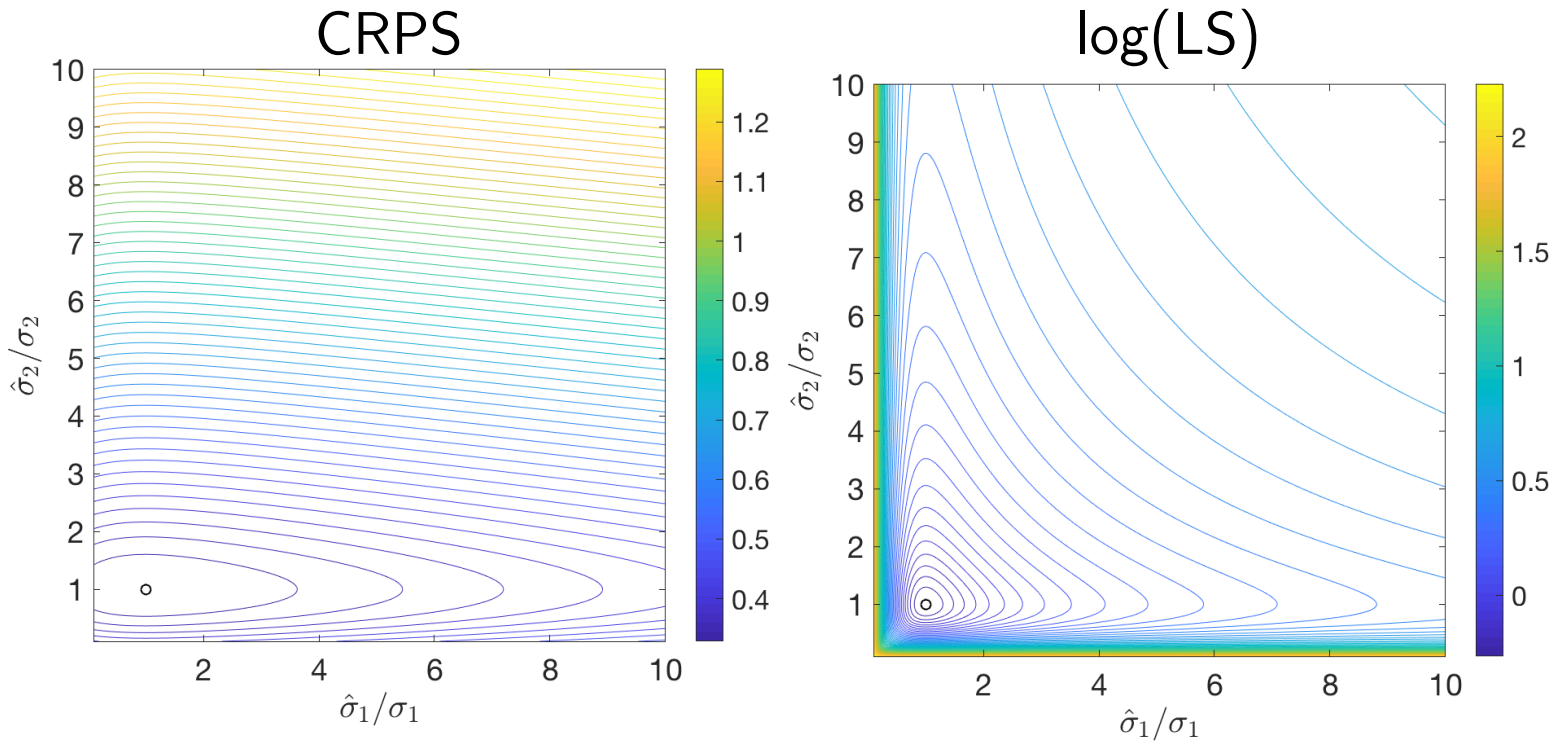
Other example

- Consider a situation with two observations $Y_i \sim \mathbb{Q}_{\theta_i} = \mathsf{N}(0, \sigma_i^2), i = 1, 2$, with $\sigma_1 = 0.1$ and $\sigma_2 = 1$.
- Assume that we want to evaluate a model which has predictive distributions $\mathbb{P}_i = \mathsf{N}(0, \hat{\sigma}_i^2)$ for Y_i, using the average of a proper scoring rule for the two observations.

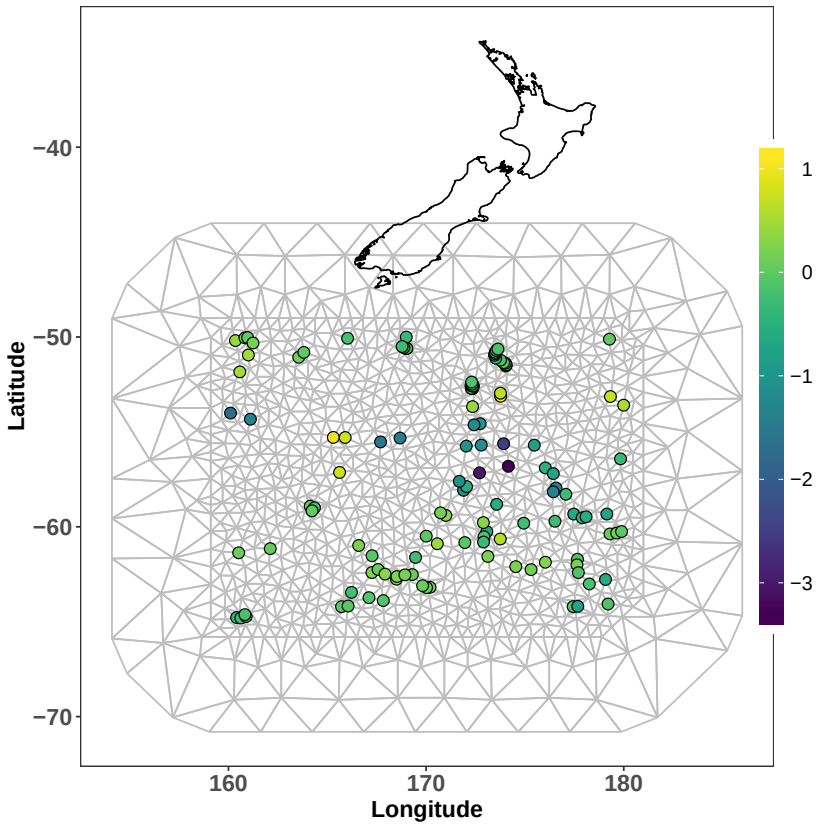
Other example

- Consider a situation with two observations $Y_i \sim \mathbb{Q}_{\theta_i} = \mathsf{N}(0, \sigma_i^2), i = 1, 2$, with $\sigma_1 = 0.1$ and $\sigma_2 = 1$.
- Assume that we want to evaluate a model which has predictive distributions $\mathbb{P}_i = \mathsf{N}(0, \hat{\sigma}_i^2)$ for Y_i, using the average of a proper scoring rule for the two observations.

Other example



Varying scale in practice?



Kuusela, M. and Stein, M.L. (2018): Locally stationary spatio-temporal interpolation of argo profiling float data. *Proceedings of the Royal Society A*, 474

Bolin, D. and Wallin, J. (2020):Multivariate type-G Matérn-SPDE random fields, *JRSSB*

Example spatial statistics

We will now go through how model evaluation using a scoring rule is typically done is spatial statistics. We start with the basic setup

Example spatial statistics

We will now go through how model evaluation using a scoring rule is typically done is spatial statistics. We start with the basic setup

- Let $\mathbf{s}_i$, $i = 1, \ldots, n$ be a set of, typically irregular, locations.

Example spatial statistics

We will now go through how model evaluation using a scoring rule is typically done is spatial statistics. We start with the basic setup

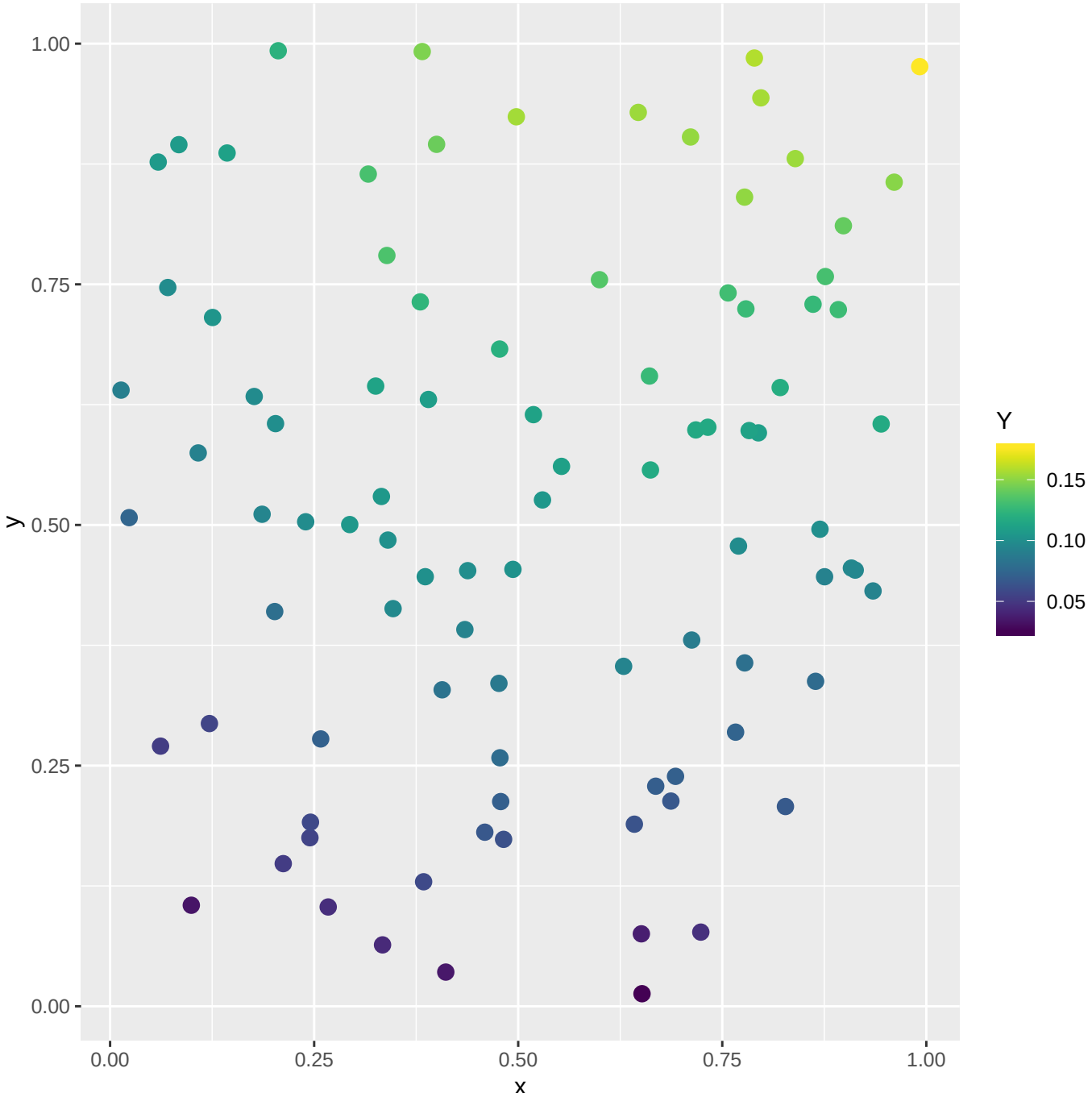
- Let $\mathbf{s}_i$, $i = 1, \ldots, n$ be a set of, typically irregular, locations.
- We have a set of observations $\{y_i\}_{i=1}^n$ at the locations $\{\mathbf{s}_i\}_{i=1}^n$.

Example spatial statistics

We will now go through how model evaluation using a scoring rule is typically done is spatial statistics. We start with the basic setup

- Let $\mathbf{s}_i$, $i = 1, \ldots, n$ be a set of, typically irregular, locations.
- We have a set of observations $\{y_i\}_{i=1}^n$ at the locations $\{\mathbf{s}_i\}_{i=1}^n$.
- The score of the model, $\mathbb{P}$, is given by $\bar{s} = \frac{1}{n} \sum_{i=1}^n S(\mathbb{P}_i, y_i)$.

Realization



variation of the standard deviation

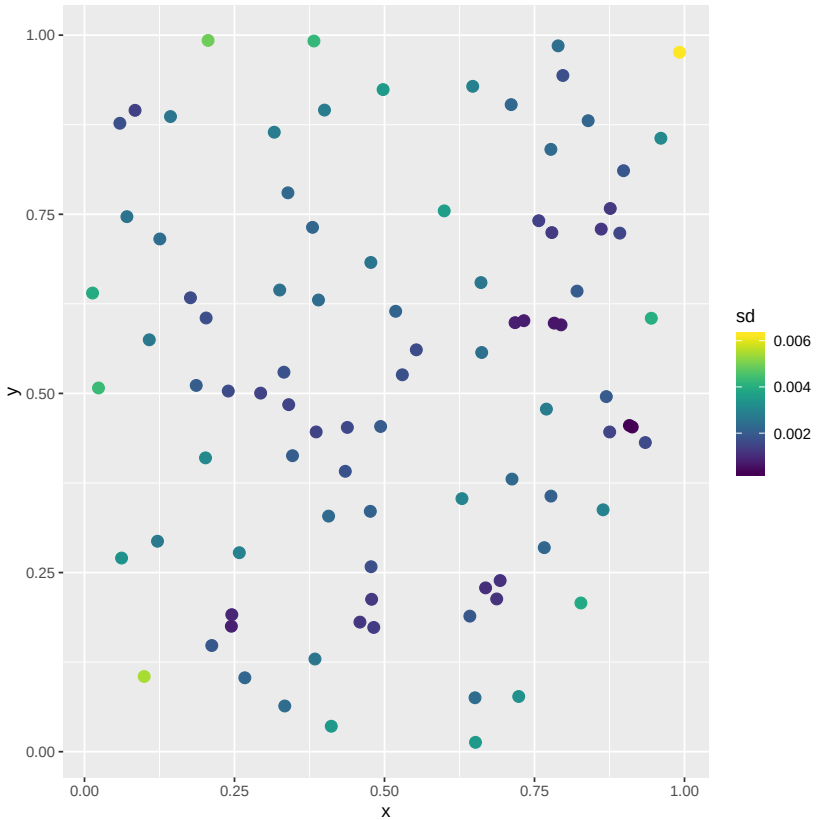


Figure: true kriging standard devation, by location

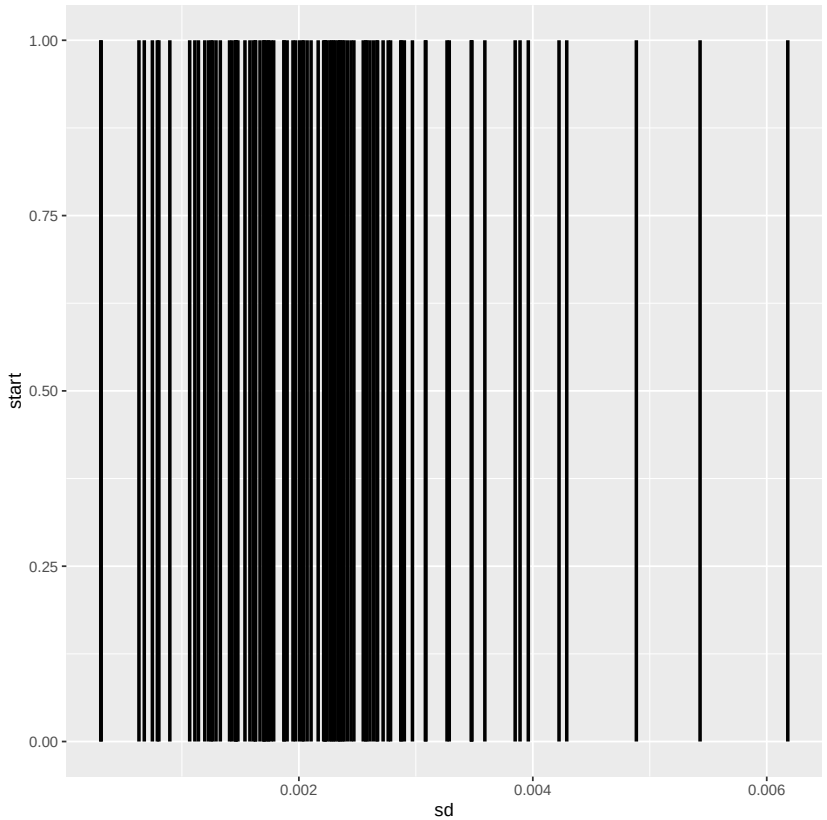


Figure: emperical density of the true kriging standard devation, σ_i

Mathematical framework

Definition

If S is a **proper scoring rule**. If $\mathbb{Q}_\sigma, \mathbb{P}_\sigma$ are probability measure with scaling σ then

$$\tilde{S}(\mathbb{P}_{\hat{\sigma}}, \mathbb{Q}_\sigma, \pi) = \int S\left(\mathbb{P}_{\hat{\sigma}(\sigma)}, \mathbb{Q}_\sigma\right) \pi(d\sigma),$$

is a proper scoring rule

Mathematical framework

Definition

If S is a **proper scoring rule**. If $\mathbb{Q}_\sigma, \mathbb{P}_\sigma$ are probability measure with scaling σ then

$$\tilde{S}(\mathbb{P}_{\hat{\sigma}}, \mathbb{Q}_\sigma, \pi) = \int S\left(\mathbb{P}_{\hat{\sigma}(\sigma)}, \mathbb{Q}_\sigma\right) \pi(d\sigma),$$

is a proper scoring rule

- The difference between this scoring rule and regular scoring rule is that there is no $\bar{S}(\mathbb{P}_{\hat{\sigma}}, y)$ function. It is a theortical construction.
- However if $\sigma_i \sim \pi$ and $Y_i \sim \mathbb{Q}_{\sigma_i}$ then

$$\frac{1}{n}\sum_{i=1}^{n} S\left(\mathbb{P}_{\hat{\sigma}_i}, Y_i\right) \to \tilde{S}(\mathbb{P}_{\hat{\sigma}}, \mathbb{Q}_\sigma, \pi)$$

Defining $\bar{s}$ mathematically

- What affects the shape of π be?

Defining $\bar{s}$ mathematically

- What affects the shape of π be?
- If Y is a Gaussian processes σ_i (and hence π) is bascially determined by the distance of the locations, $\mathbf{s}$.

Defining $\bar{s}$ mathematically

- What affects the shape of π be?
- If Y is a Gaussian processes σ_i (and hence π) is bascially determined by the distance of the locations, $\mathbf{s}$.
- if assume that the observations comes from some point processes, we can derive the true leave-one-out standard devations.

Defining $\bar{s}$ mathematically

$$\mathbf{s} \sim PPois(\lambda)$$

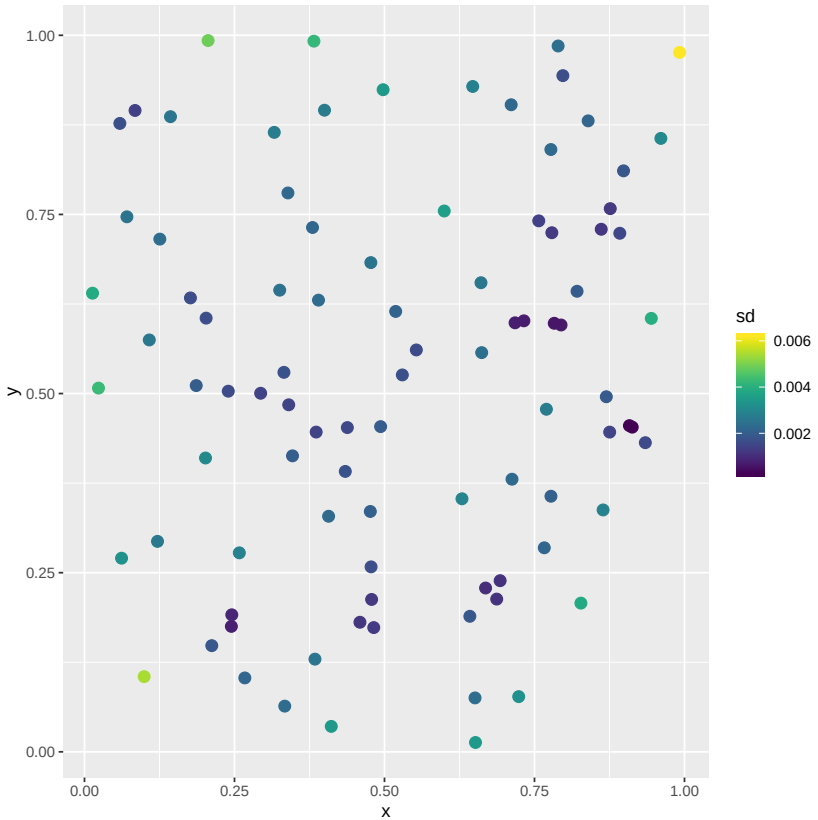


Figure: Standard devation by location

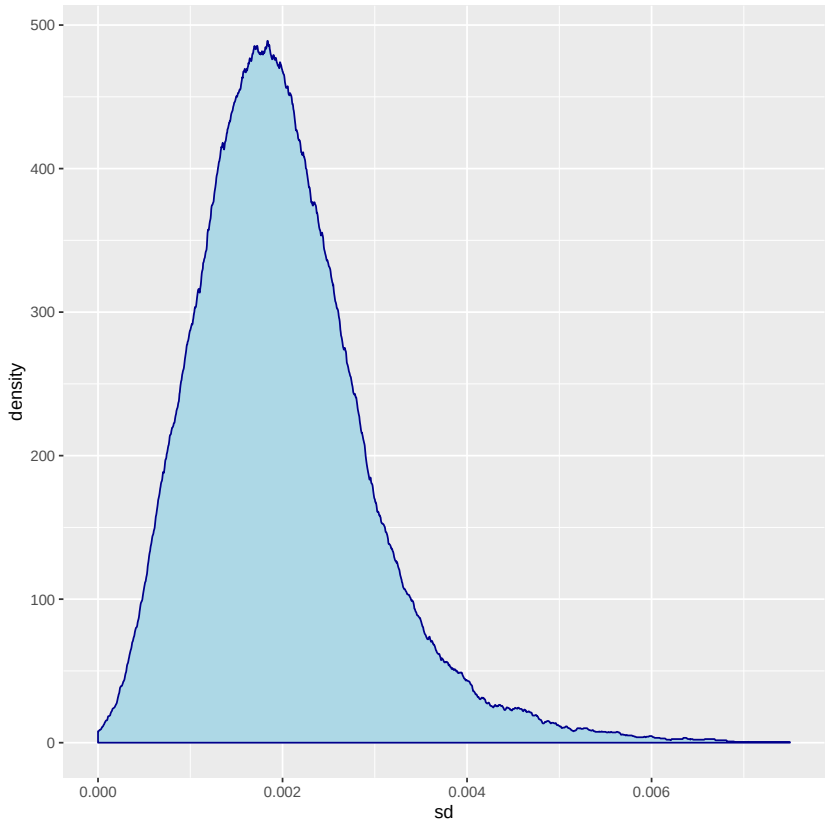
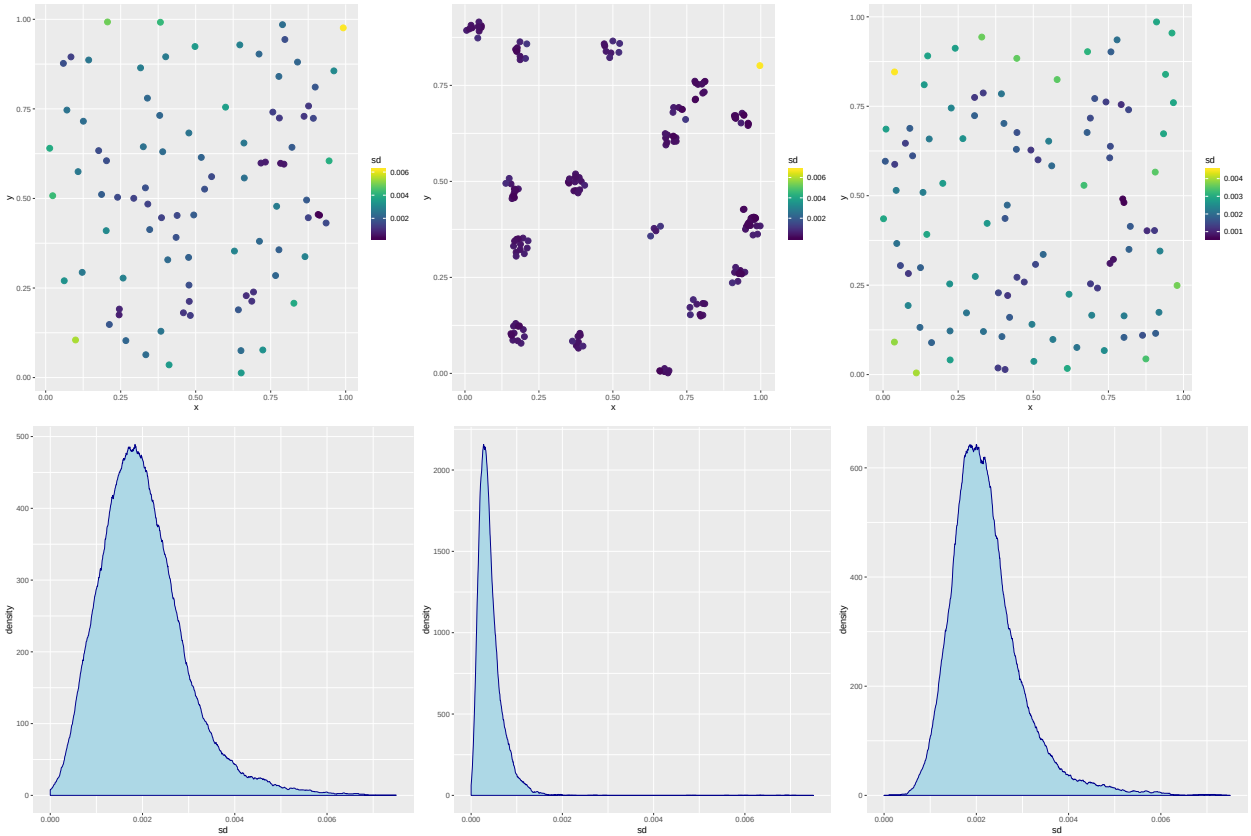
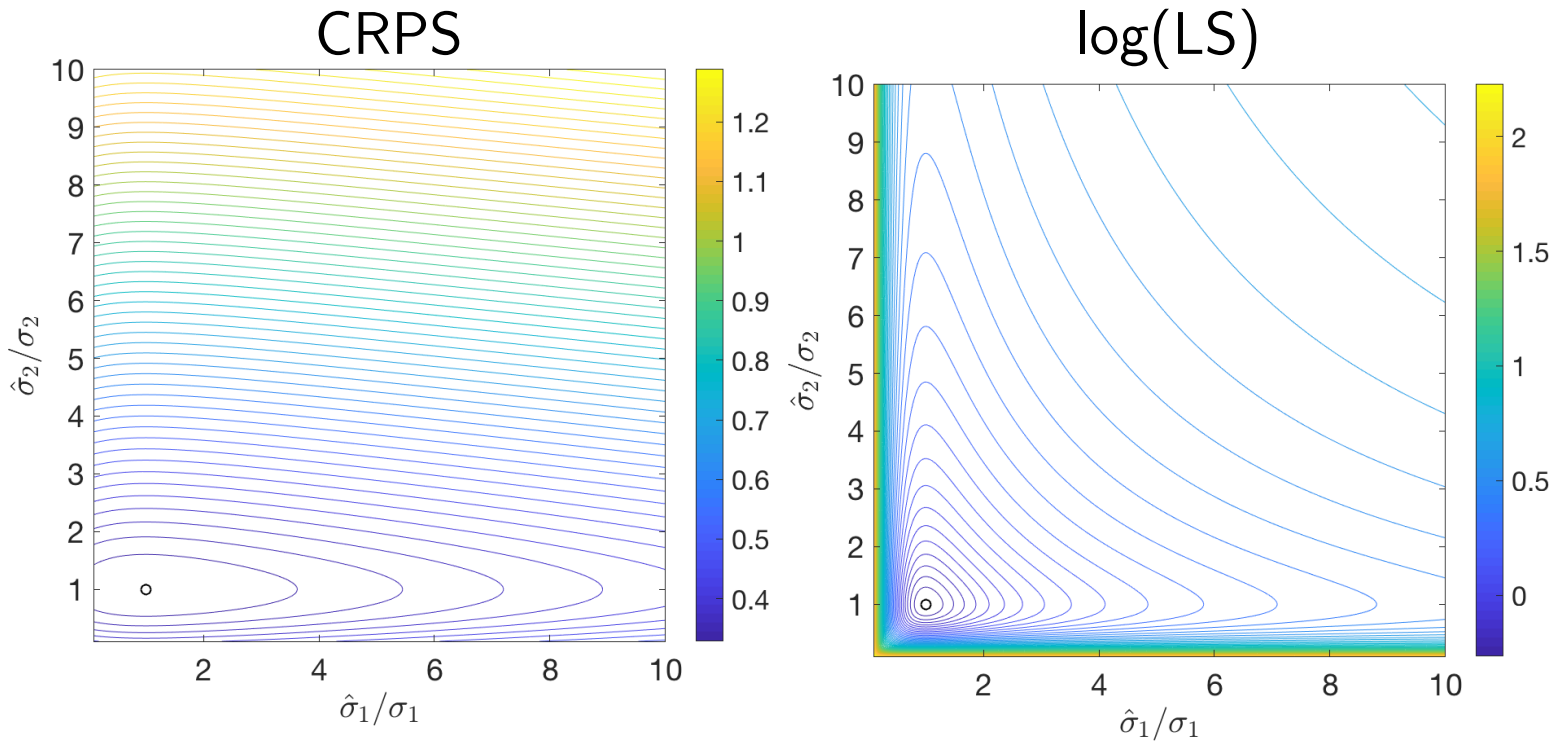


Figure: Estimate of π

Point distribution and π



Recall the issue



locale scale

Definition

Let S be a proper scoring rule and let $\mathbb{Q}_\theta = \mathbb{Q}_{[\mu,\sigma]}$ be a probability measure with location μ and scale σ. Assume that there exist a constant $p \in \mathbb{R}$ and a function $s(\mathbb{Q}_\theta, r) : \mathcal{F} \times \mathbb{R}^2 \to \mathbb{R}^+$, such that for each $r \in \mathbb{R} \times \mathbb{R}$

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

Then s is the scale function of S, which is locally scale invariant if $s(\mathbb{Q}_\theta, r) \equiv s(\mathbb{Q}, r)$.

locale scale function

- The scale function of the log score, $S(\mathbb{P}, y) = \log(f(y))$, is locally scale invariant.

locale scale function

- The scale function of the log score, $S(\mathbb{P}, y) = \log(f(y))$, is locally scale invariant.
- The scale function of the CRPS is

$$s(\mathbb{Q}_\sigma, r) = \sigma S(\mathbb{Q}_1, r),$$

  i.e. the scale function is not locally scale invariant.

Known issue

- The issue of unbalanced predictive distribution is not unknown. A lot of work has been put of standardizing observations

$$S(\mathbb{P}, y) = \frac{|med(\mathbb{P}) - y|}{\sqrt{\mathsf{V}_{\mathbb{P}}[Y]}}$$

not a proper score.

Known issue

- The issue of unbalanced predictive distribution is not unknown. A lot of work has been put of standardizing observations

$$S(\mathbb{P}, y) = \frac{|med(\mathbb{P}) - y|}{\sqrt{\mathsf{V}_{\mathbb{P}}[Y]}}$$

  not a proper score.
- Previous solutions is to use a reference prediction, using a so called skill score

$$S^{skill}(\mathbb{P}, y) = \frac{S(\mathbb{P}, y)}{S(\mathbb{P}^{ref}, y)}$$

  here $\mathbb{P}^{ref}$ is the reference predictor. However the results will be determined by the reference measure

Known issue

- The issue of unbalanced predictive distribution is not unknown. A lot of work has been put of standardizing observations

$$S(\mathbb{P}, y) = \frac{|med(\mathbb{P}) - y|}{\sqrt{\mathsf{V}_{\mathbb{P}}[Y]}}$$

  not a proper score.
- Previous solutions is to use a reference prediction, using a so called skill score

$$S^{skill}(\mathbb{P}, y) = \frac{S(\mathbb{P}, y)}{S(\mathbb{P}^{ref}, y)}$$

  here $\mathbb{P}^{ref}$ is the reference predictor. However the results will be determined by the reference measure
- An other alternative is to use a weighted CRPS

$$S(\mathbb{P}, y) = \int \left(\mathbb{P}(X \leq x) - \mathbb{I}(y \leq x)\right)^2 \omega(x) dx$$

  see for Gneiting and Ranjan, 2011.

Our idea

- Recall the **continuous ranked probability score (CRPS)** is given by

$$\mathrm{S}(\mathbb{P}, y) = -\mathsf{E}_{\mathbb{P}}|X - y| + \frac{1}{2}\mathsf{E}_{\mathbb{P}}\mathsf{E}_{\mathbb{P}}|X - X'|$$

Our idea

- Recall the **continuous ranked probability score (CRPS)** is given by

$$\mathrm{S}(\mathbb{P}, y) = -\mathsf{E}_{\mathbb{P}}|X - y| + \frac{1}{2}\mathsf{E}_{\mathbb{P}}\mathsf{E}_{\mathbb{P}}|X - X'|$$

- We introduce a different scoring rule, which we denote **standardized continuous ranked probability score (SCRPS)**:

$$S(\mathbb{P}, y) = -\frac{\mathsf{E}_{\mathbb{P}}|X - y|}{\mathsf{E}_{\mathbb{P}}\mathsf{E}_{\mathbb{P}}|X - X'|} - \frac{1}{2}\log\left(\mathsf{E}_{\mathbb{P}}\mathsf{E}_{\mathbb{P}}|X - X'|\right)$$

Kernel scores

Theorem (Gneiting and Raftery (2007))

Let $\mathbb{P}$ be a Borel probability measure on a Hausdorff space Ω. Assume that g is a non-negative, continuous negative definite kernel on $\Omega \times \Omega$ and let $\mathcal{P}$ denote the class of Borel probability measures on Ω such that $\mathsf{E}_{\mathbb{P},\mathbb{P}}\left[g(X,Y)\right] < \infty$. Then the scoring rule

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

is proper on $\mathcal{P}$.

- CRPS is obtained by noting that $g(x,y) = |x-y|$ is a negative definite kernel.
- In fact $g(x,y) = |x-y|^{\alpha}$, $\alpha \in (0,2]$ is a negative definite kernel.

h-function Kernel scores

Theorem

Let g be a non-negative, continuous negative definite kernel on $\Omega \times \Omega$, and let $\mathbb{P}$ be Borel probability measure on Ω. Let h be a monotonically increasing concave differentiable function on $\mathbb{R}^+$. Further let $\mathcal{P}$ denote the class of Borel probability measures on Ω s.t $\mathsf{E}_{\mathbb{P}}\mathsf{E}_{\mathbb{P}}\left[g(X, X')\right] < \infty$. Then the scoring rule

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

is proper on $\mathcal{P}$.

- sCRPS is obtained by noting that $g(x, y) = |x - y|$ is a negative definite kernel, and $h(x) = \frac{1}{2}\log(x)$ is a monotonically increasing concave differentiable function.

Scale function

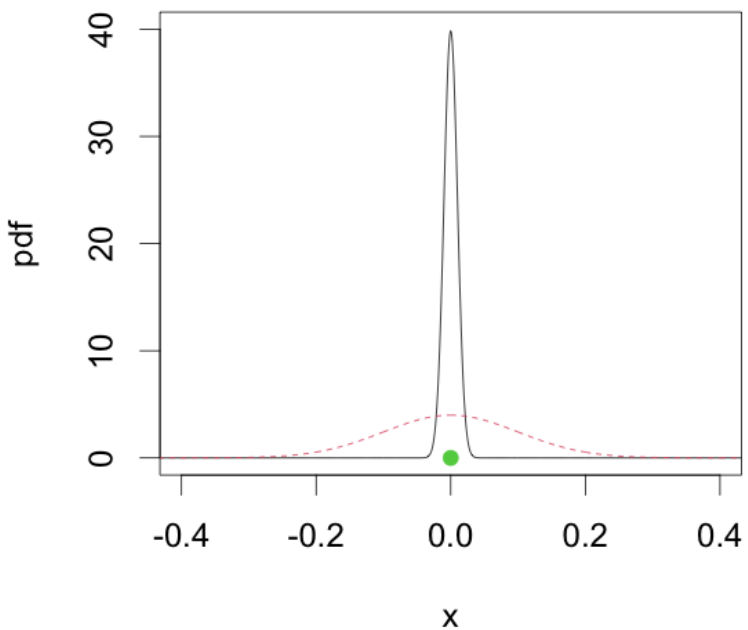
- The scale function of the log score, $S(\mathbb{P}, y) = -\log(f(y))$, is locally scale invariant.
- The scale function of the CRPS is

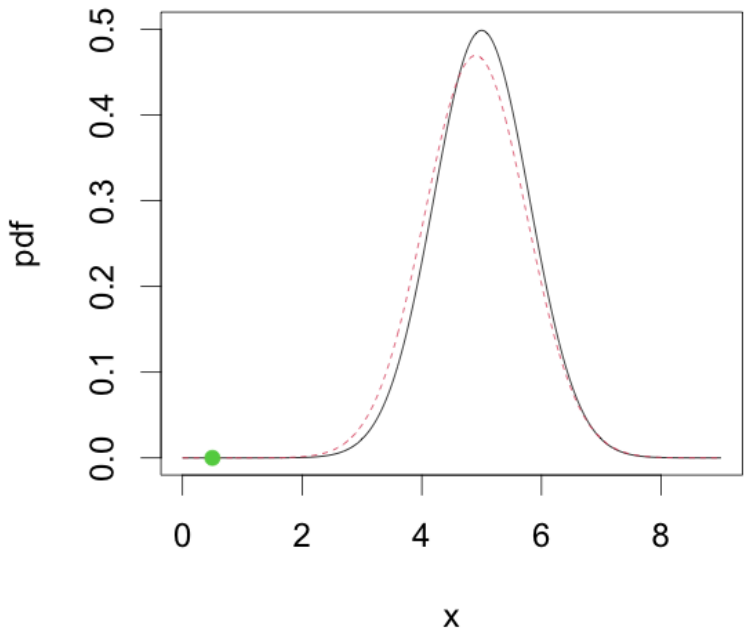
$$s(\mathbb{Q}_\sigma, r) = \sigma S(\mathbb{Q}_1, r),$$

  i.e. the scale function is not locally scale invariant.
- The scale function SCRPS is locally scale invariant!

Two observations, two models, result

	Model 1			Model 2		
	CRPS	log-score	sCRPS	CRPS	log-score	sCRPS
Y_1	0.0023	-3.6862	-1.5351	0.0234	-1.3836	-0.3838
Y_2	4.0486	16.516	4.9338	3.9204	14.154	4.5666
mean	2.0255	6.4149	1.6994	1.9719	6.3853	2.0914





Other example