Ranking earthquake forecasts:
On the use of proper scoring rules to discriminate forecasts

Francesco Serafini, Mark Naylor, Finn Lindgren, Maximilian Werner

- Forecasts consists of probabilities of earthquake ***activity*** for different bins (e.g. proability of observing at least one earthquake in a given space-time-magnitude bin)
- Forecasts can be ranked using ***proper*** scoring rules (e.g. Brier score, Logarithmic score)
- We prove that the Parimutuel Gambling score is ***proper only when two forecasts*** are compared
  - It ***is not proper*** when two forecasts are compared against a reference model
- We prove that the Parimituel Gambling score ***is never proper*** when comparing multiple forecasts
- We ***compare*** the Parimutuel Gambling score with the Brier and Logarithmic scores in different scenarios:

  1. Multiple bins with the same probability (analitical results)
  2. Multiple bins with different probabilities (approximated results)

- We show how to use simulation from a forecast (Temporal – ETAS) to inform decision on ***amount of data*** required to distinguish forecasts and the consequences of different ***partitioning of bins***

Contacts : francesco.serafini@ed.ac.uk

Forecast of earthquake activity

Given N space-time-magnitude bins a forecast is a collection of probabilities $\boldsymbol{p} = p_1, \ldots, p_N$ such that

$$p_i = \text{ probability of observing at least one earthquake (activity) in the i-th space-time-magnitude bin}$$

Objective: Given k forecasts $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k$ we want to be able to rank them based on the observed data $\boldsymbol{x} = x_1, \ldots, x_N$ where

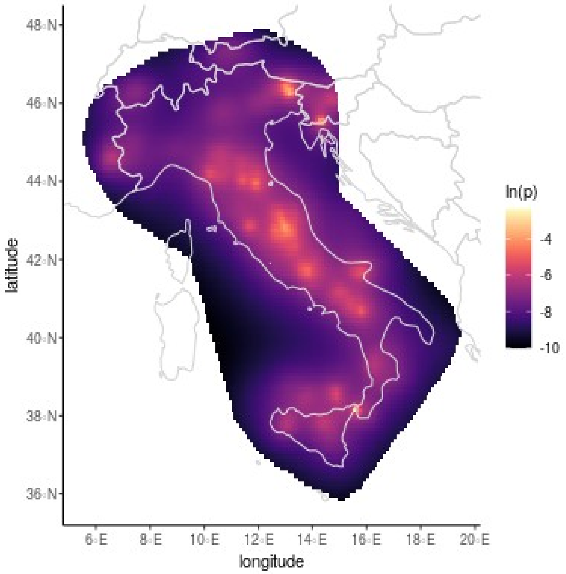
$$x_i = 1 \quad \text{activity in the i-th space-time-magnitude bin}$$
$$x_i = 0 \quad \text{no activity in the i-th space-time-magnitude bin}$$

How: using a *positively oriented* scoring rule $S(\boldsymbol{p}|\boldsymbol{x})$

forecast

data

The higher the better



5 year Italy adaptive-smoothing seismicity forecast
(Werner et. al. 2011)

Proper scoring rules

Before observing the data $\boldsymbol{x}$ the score value can be seen as a random variable $S(\boldsymbol{p}|\boldsymbol{X})$

The distribution of $S(\boldsymbol{p}|\boldsymbol{X})$ depends only on the distribution of $\boldsymbol{X}$

$\boldsymbol{X}=X_1,\ldots,X_N$ is a collection of binary random variables $X_i \sim \mathrm{Ber}(p_i^*)$

Bernoulli random variable

Considering only the i-th bin, the *expected* score is given by

$$E[S(p_i|X_i)]=S(p_i|1)\,p_i^*+S(p_i|0)(1-p_i^*)$$

Definition : A scoring rule is proper if for any $p_i \neq p_i^*$ the *expected* score is such that

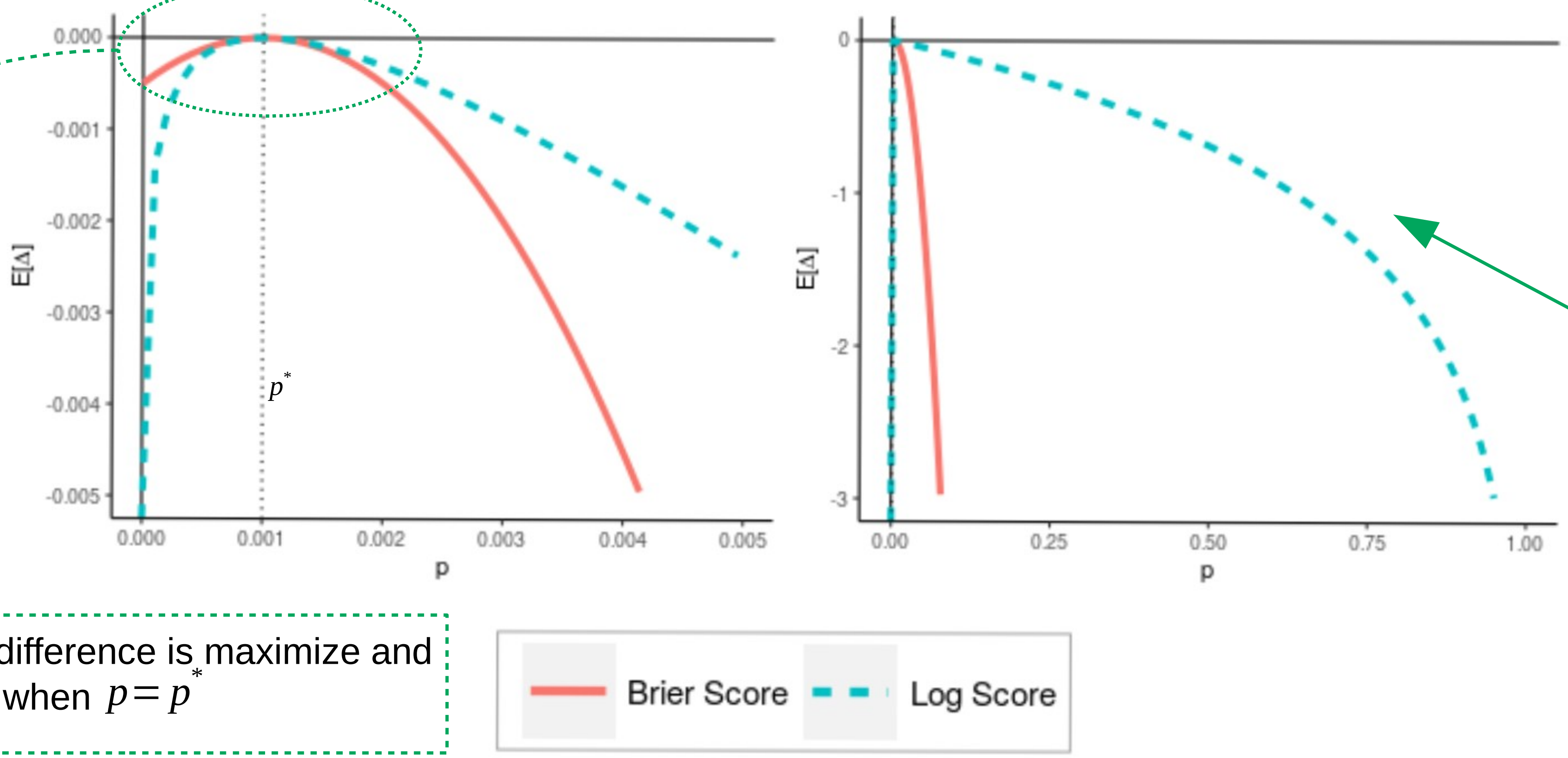
$$E[S(p_i|X_i)] \leqslant E[S(p_i^*|X_i)]$$

The expected score is maximized by the value of p which generates the data.

If a score is proper for a single bin, the mean across different bins is also proper.

Proper scoring rules – Example

The plot shows the expected score difference between p (varying) and $p^*=0.001$, $E[\Delta]=E[S(p|X)]-E[S(p^*|X)]$



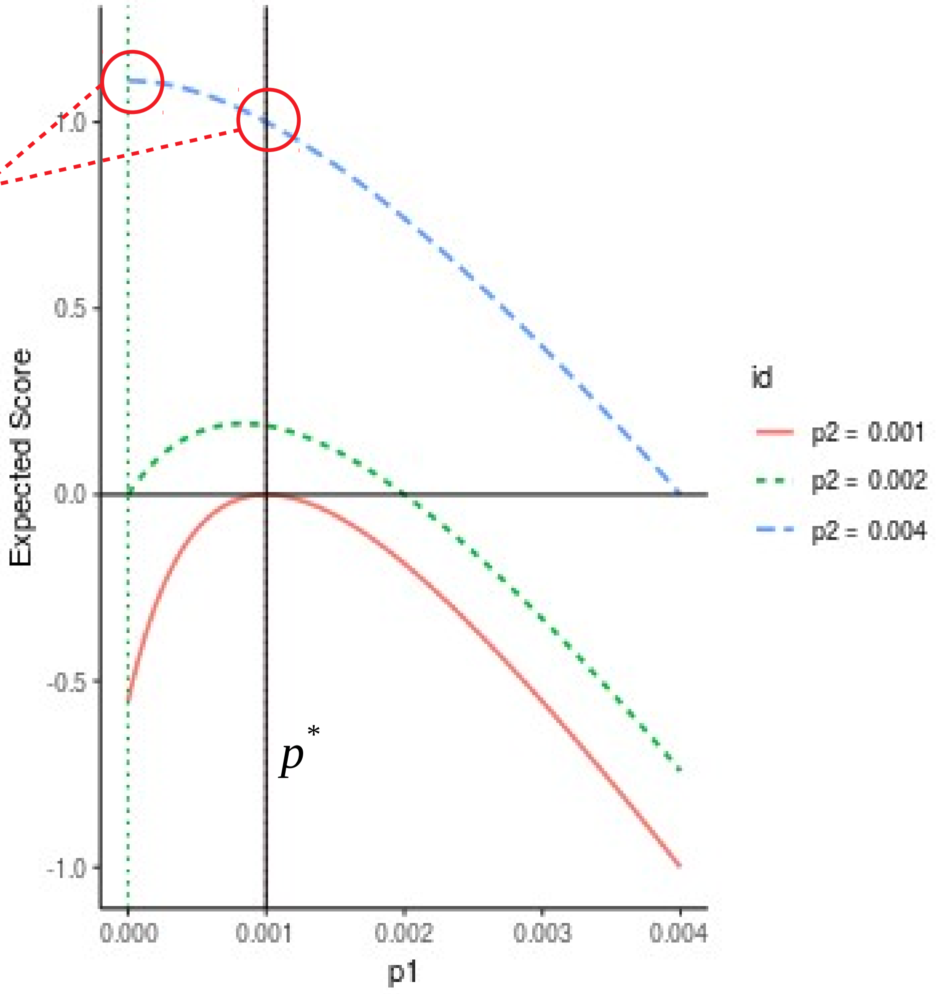
The score difference is maximize and equal zero when $p=p^*$

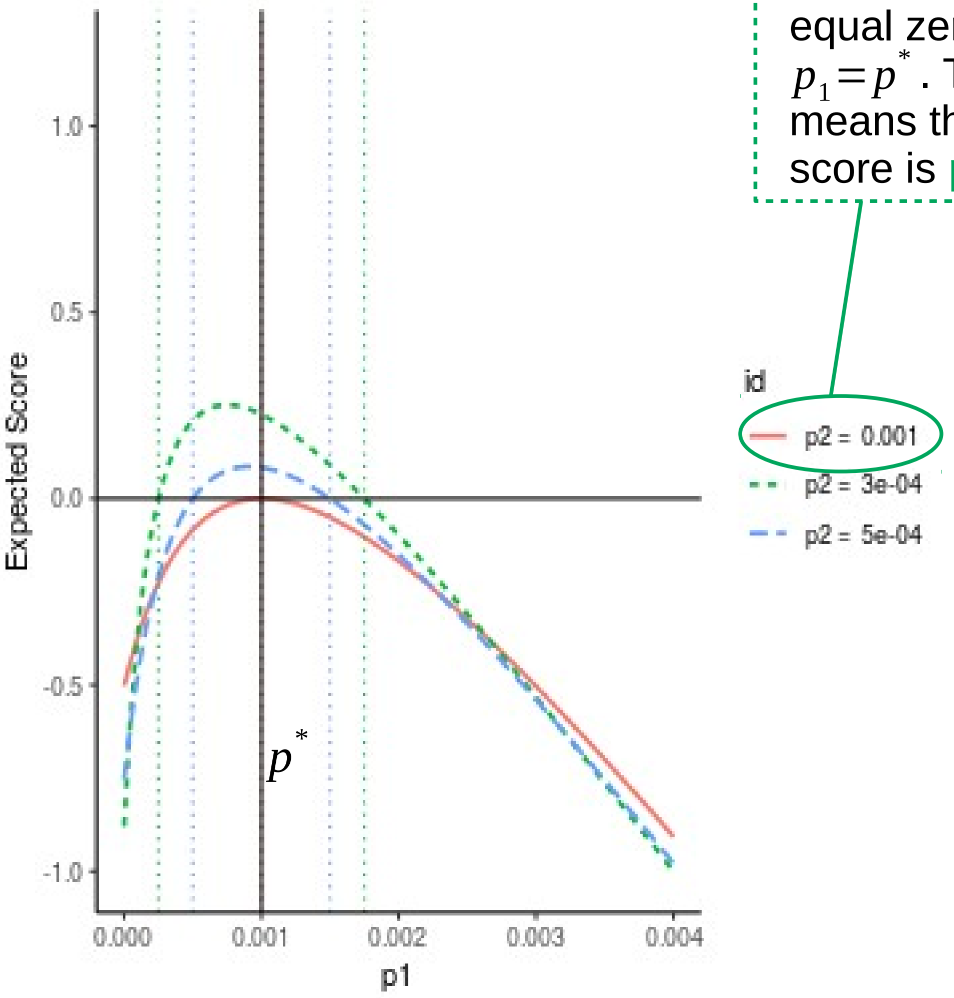
The score difference is never positive, which means that p^* is always preferred to p

The Parimituel Gambling – Comparing two forecasts

The parimutuel gambling score needs **at least** two forecast to be defined.
When comparing two forecasts, when $S(p_1, p_2|x) > 0$ then the first forecast is preferred

When $p_2 \neq p^*$ the score for p_1 **is not** maximized at $p_1 = p^*$. Using $p_2 = 0.004$ as reference model, $p_1 = 0$ has an higher score than $p_1 = p^*$. This means that the score is **not proper** when two forecasts are compared against a reference model.



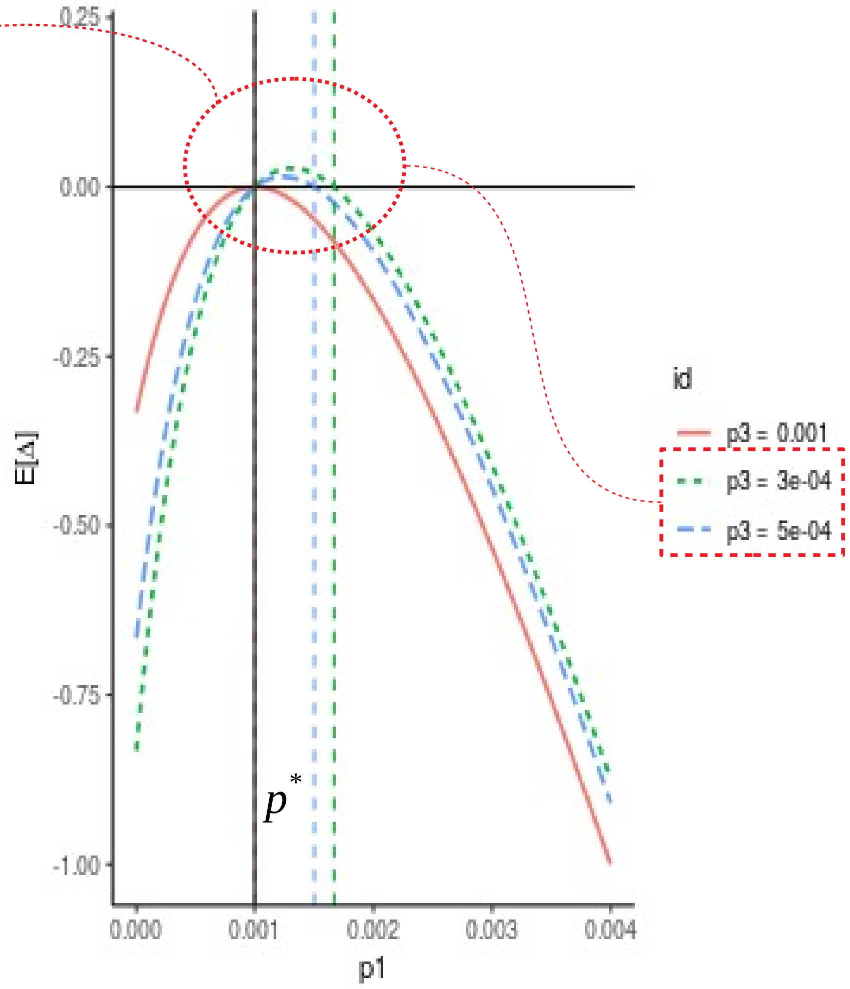


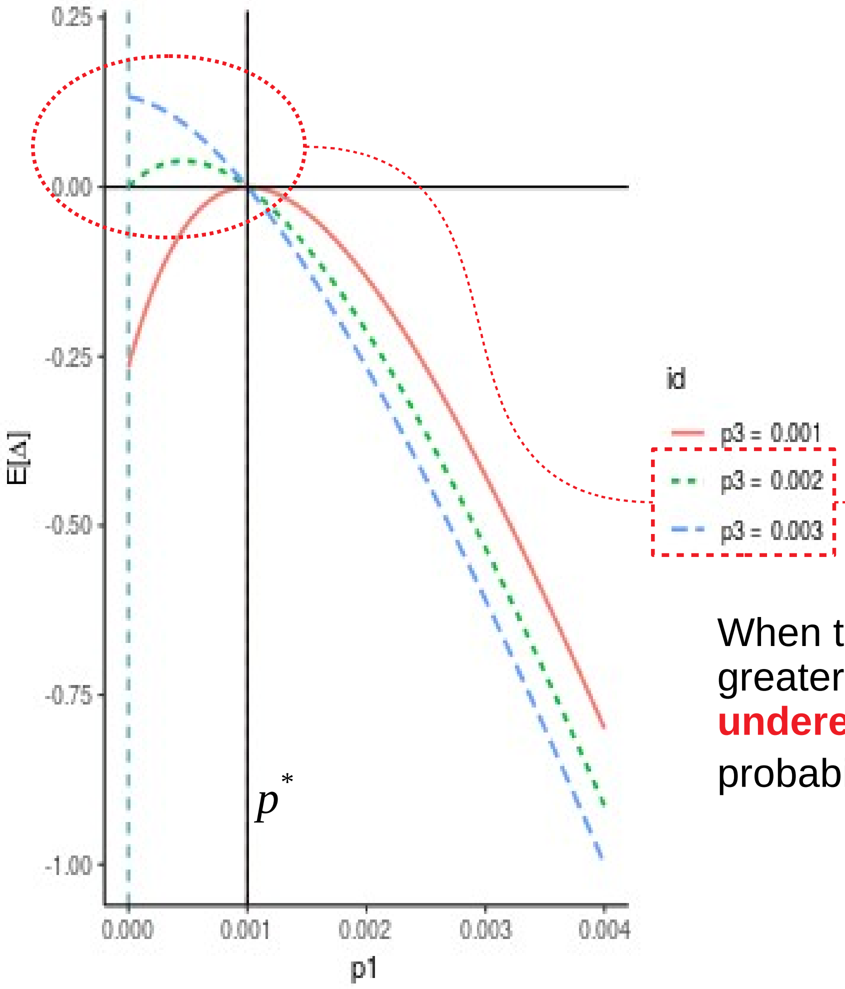
When $p_2 = p^*$ the score for p_1 is always negative (never preferred) and it is maximized and equal zero when $p_1 = p^*$. This means that the score is **proper**

The Parimituel Gambling – Comparing multiple forecasts

The parimutuel gambling score uses the average forecast as reference model.
Below, three forecasts comparison: the figures shows the score difference between p_1 (varying) and $p_2 = p^*$

When the average forecast is smaller than p^* forecasts **overestimating** the true probability are preferred





To be proper the difference should be always negative.

When the average forecast is greater than p^* forecasts **underestimating** the true probability are preferred

Ranking Forecasts is an **estimation problem**

It is *impossible* to observe directly the expected value of a score

We need to **estimate it** using the **data**

When using a limited amount of data, our estimate may be **far** from the quantity we wish to estimate

It means we could end up expressing a preference for the model ***more distant*** from the data generating model

When comparing two models, a way to avoid this problem is to look at the **confidence interval** of the expected score difference

If the interval contains zero **we do not express a preference**

If the interval lies entirely above zero **we express a preference for the first model**

If the interval lies entirely below zero **we express a preference for the second model**

Multiple Bins Single Probability – Confidence Interval

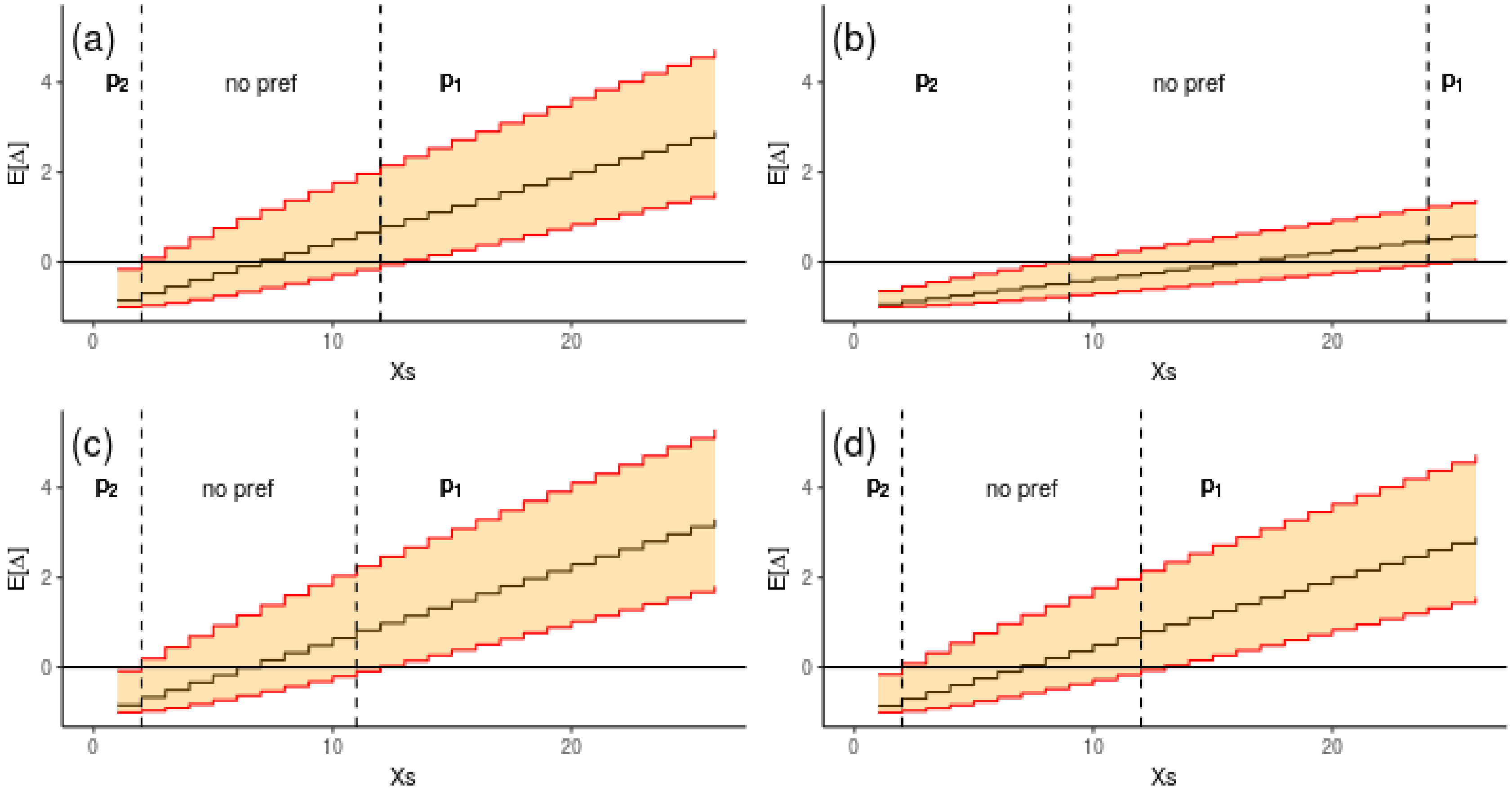
We consider $N = 10000$ independent bins all with the same probability $p^* = 0.001$

We compare two forecasts, $p_1 = p^*$ and $p_2 = p^*/3$

Given that, the observed score difference and the confidence interval **depends only** on the number of observed active bins X_s

X_s is a Binomial random variable of size N and probability p^*

We can calculate **analitically** the confidence intervals and the probability of each outcome



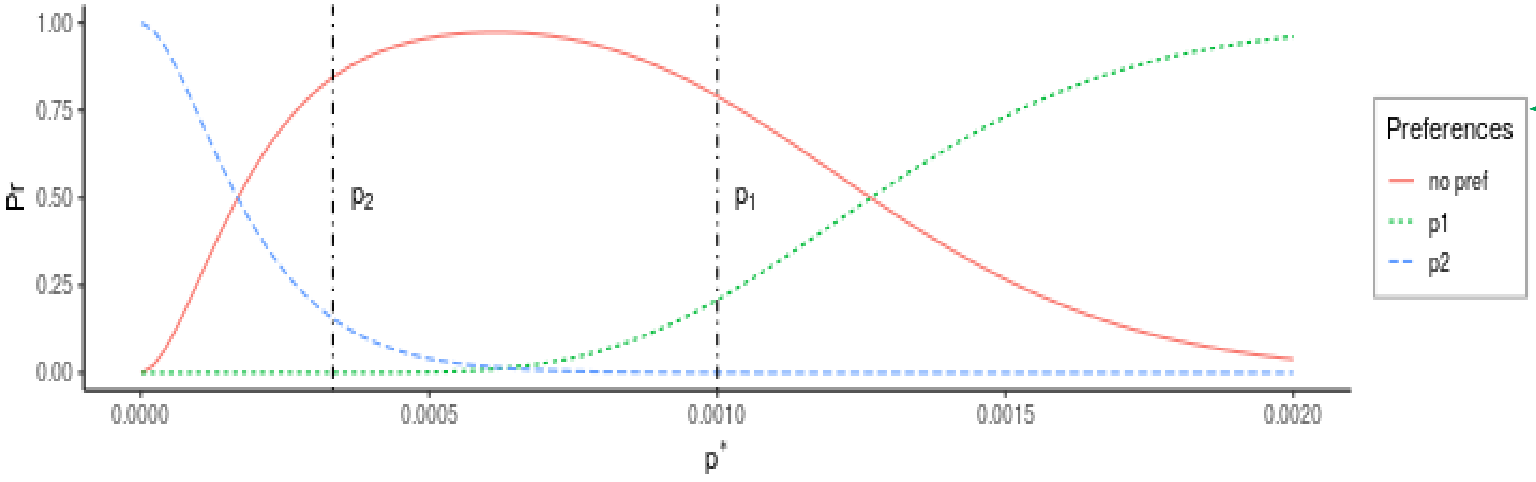
(a) Brier score, (b) Pairwise Gambling score, (c) Log score, (d) Full Gambling score

The PG score (b) is **improper** and favors the second forecast, this is because we have used as reference model $p_0 = 5\,p^*$.

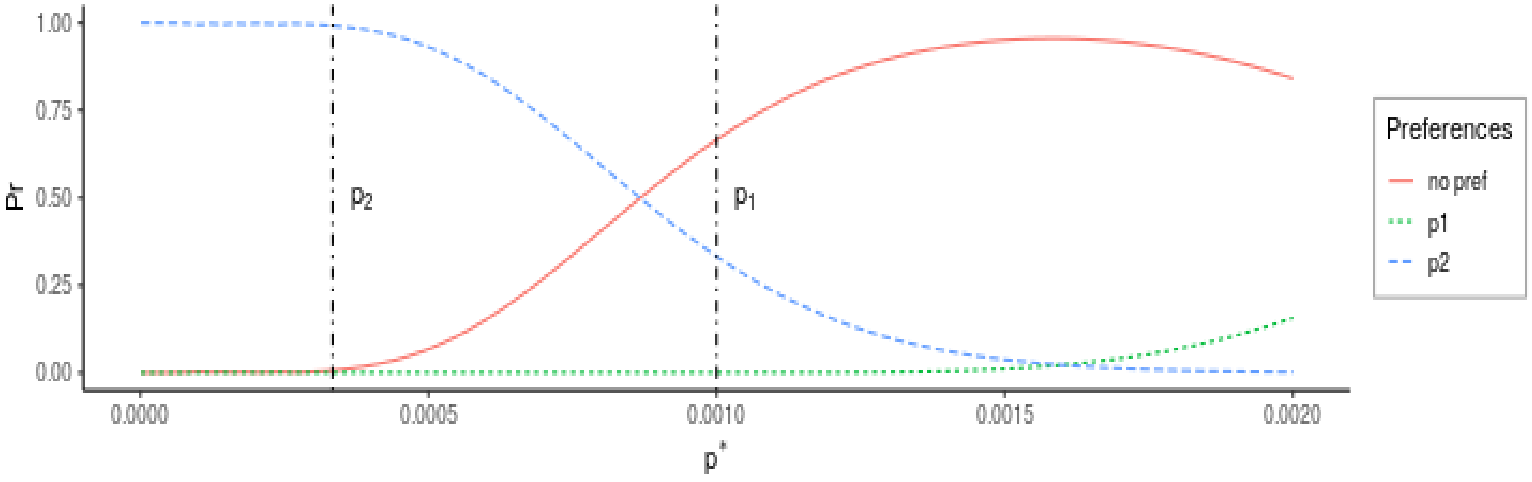
The FG score (d) is **proper** and favors the first forecast which is equal to the data generating model.

Multiple Bins Single Probability – Outcome Probabilities

Outcome probabilities varying the value of p^*



Brier score: It is proper and presents high probability of not expressing a preference when p^* is between p_1 and p_2 ; probability of preferring p_1 grows increasing p^* ; probability of preferring p_2 grows decreasing p^*

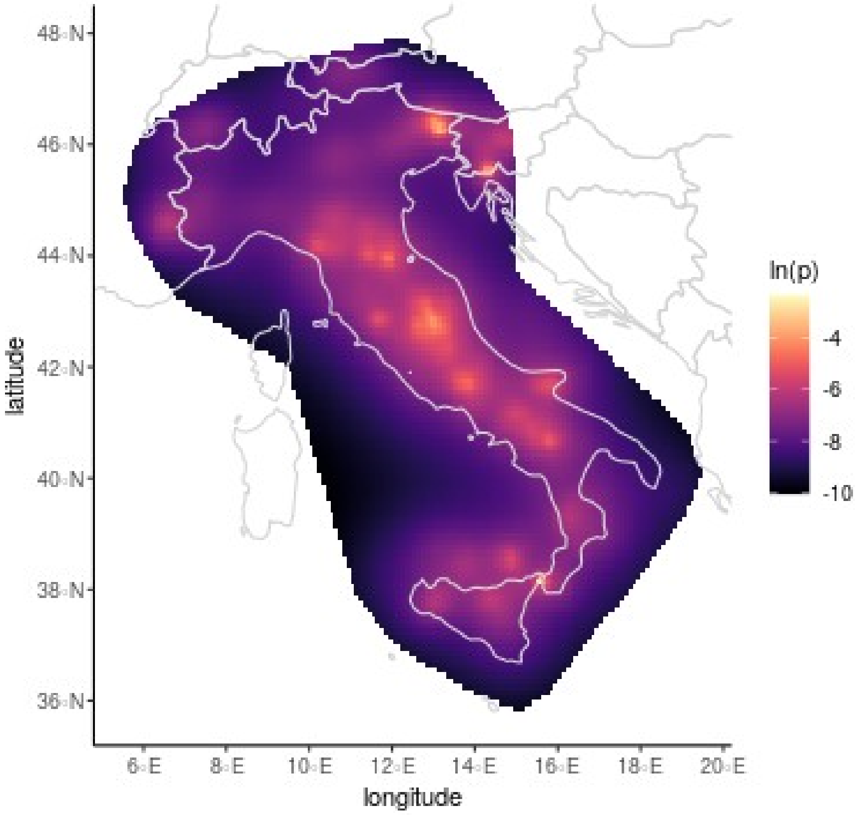


Pairwise Gambling score: It is improper and presents an higher probability of preferring p_2 even when p_1 is closer to p^* .
This is because we have used as reference model $p_0=5p^*$ for which the score tends to favor models underestimating the probability p^*

Multiple Bins Multiple Probabilities

The probabilities p_i^* are different for any bin (e.g depends on the location of the bin)

5 year Italy adaptive-smoothing forecast

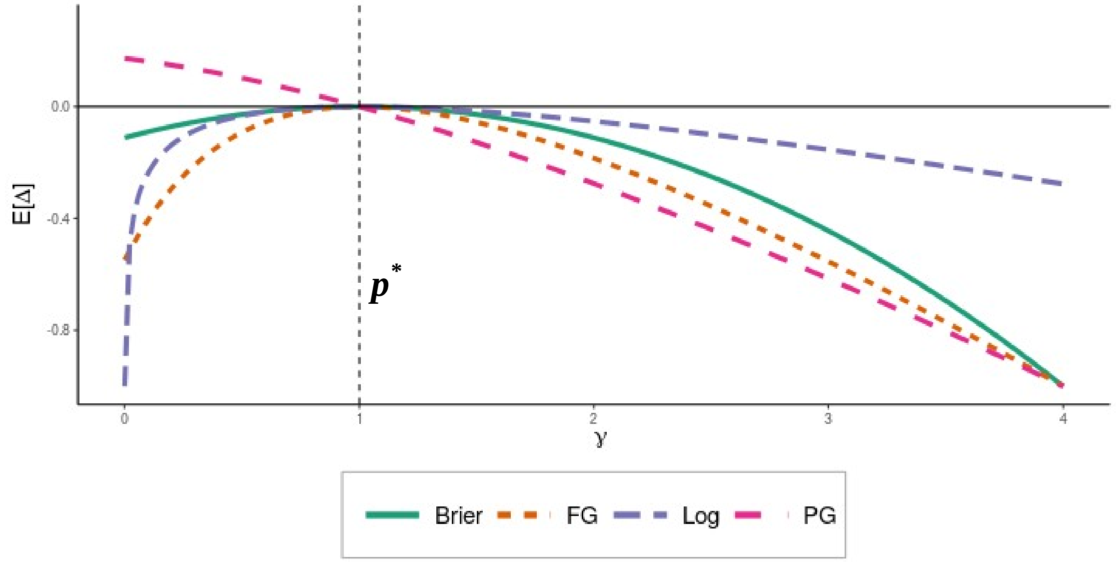


The observed score and the confidence interval now, depends on **where** we observe activity and the bins are **not independent.**

We have to resort to **approximation methods** for the confidence intervals and to **simulations** to calculate the probability

Knowing p_i^* we can always **calculate exactly** the expected value of the score difference

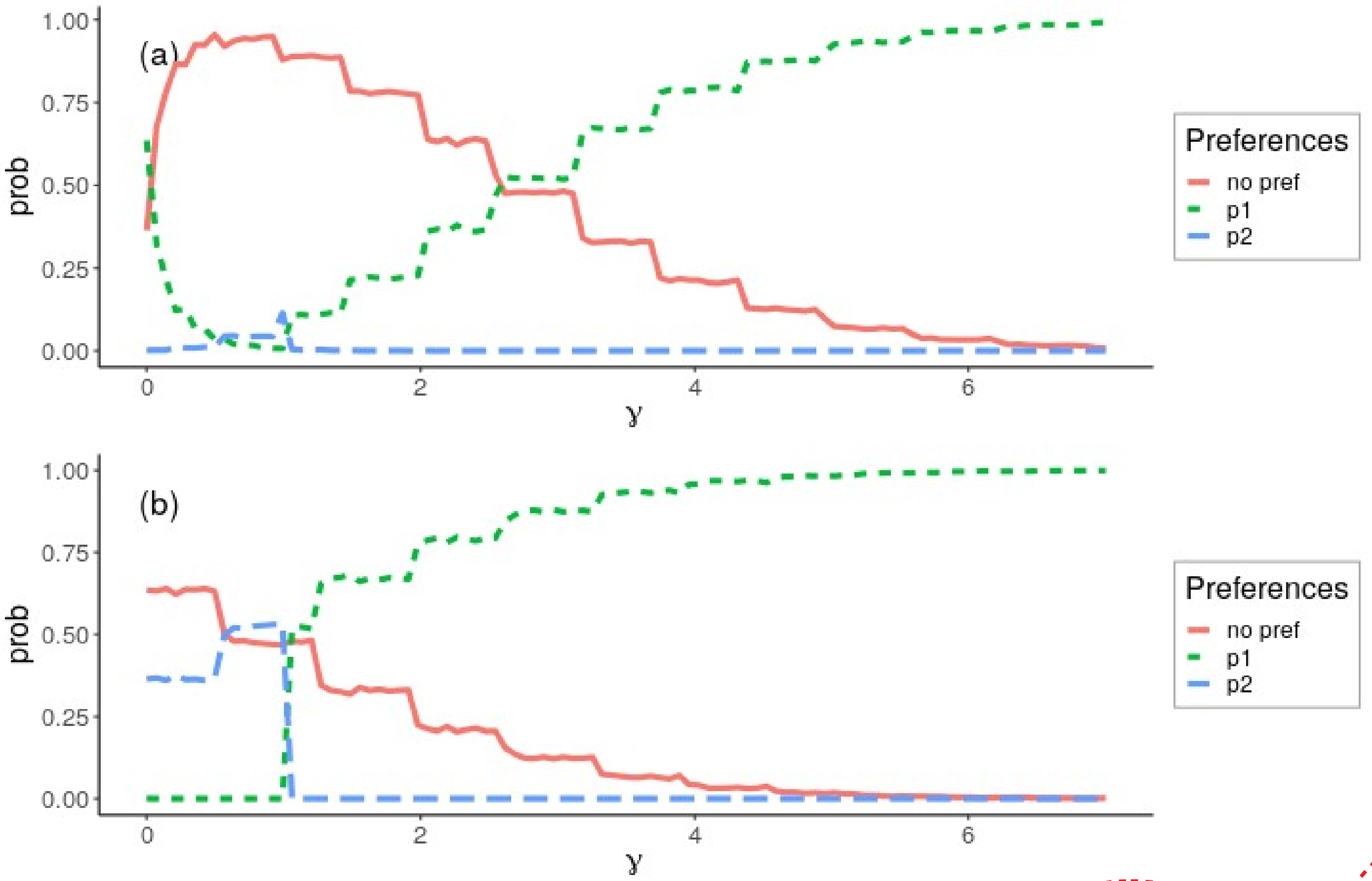
Expected score difference between $\boldsymbol{p_2} = \gamma \boldsymbol{p}^*$ and $\boldsymbol{p_1} = \boldsymbol{p}^*$



Pairwise Gambling score, using $\boldsymbol{p_0} = 5\,\boldsymbol{p}^*$ as reference, favors forecast underestimating $\boldsymbol{p}^*(\gamma < 1)$

Multiple Bins Multiple Probabilities-Outcome probabilities

We used **Gaussianly approximated** confidence intervals and **10000 simulations** to obtain an approximation of the probability of each outcome.



(a) Log score; (b) Pairwise Gambling score with reference model $p_0 = 5\,p^*$

Using 5 year Italy forecast as data generating model p^*, $p_1 = p^*$ and $p_2 = \gamma\, p^*$

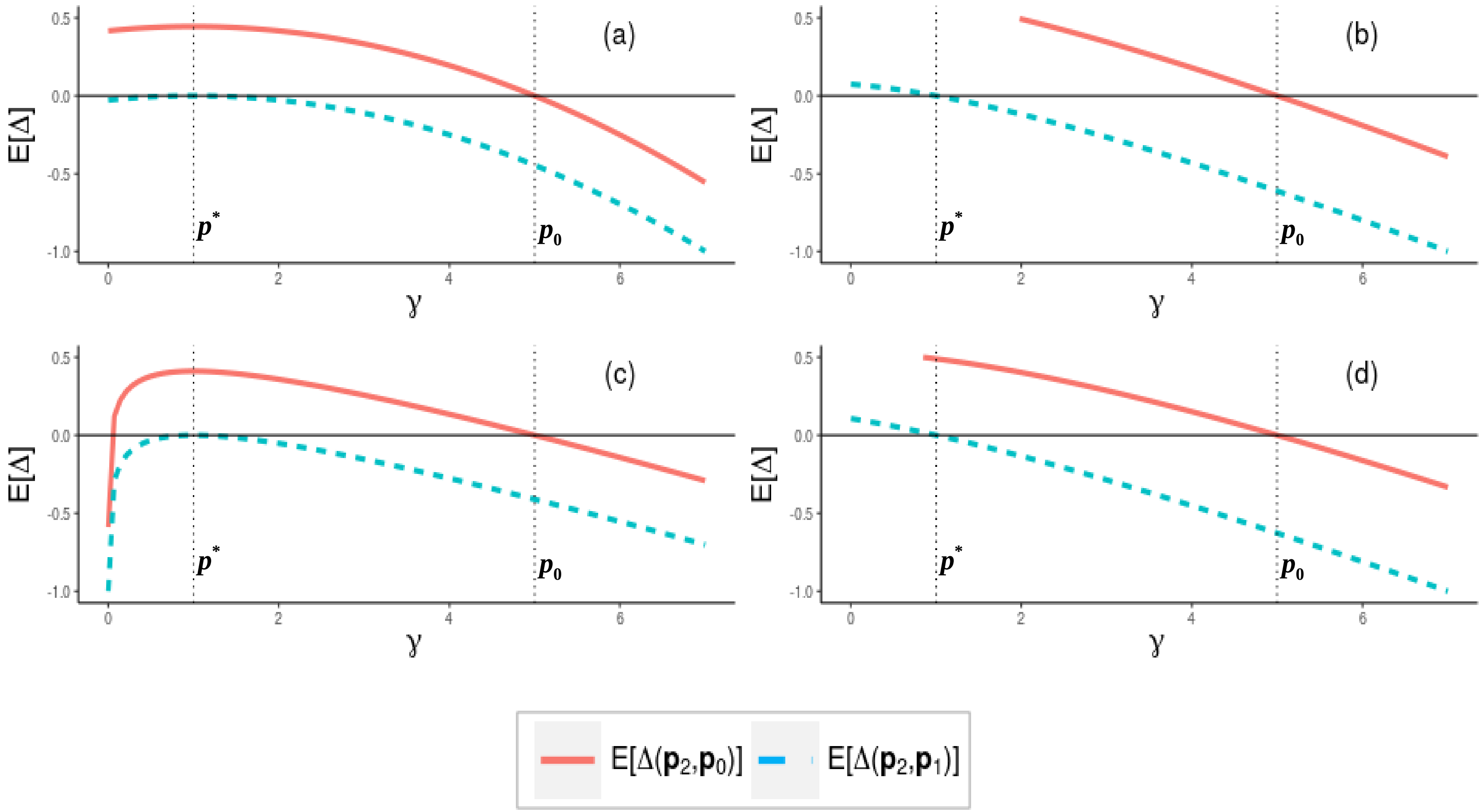
The Log score is **proper** and do not distinguish between models when they are similar $\gamma \approx 1$. As γ moves from 1 the probability of preferring p_1 increases

The PG score is **improper** and it presents an high probability of preferring p_1 when $\gamma > 1$. However this probability drops to almost zero when $\gamma < 1$

This is because the average forecast is greater than p^* which means that the reward is higher for forecast underestimating p^*

Multiple Bins Multiple Probabilities-Three forecasts comparison

Expected score difference between $\boldsymbol{p}_2 = \gamma\, \boldsymbol{p}^*$ and $\boldsymbol{p}_0 = 5\, \boldsymbol{p}^*$ and between $\boldsymbol{p}_2$ and $\boldsymbol{p}_1 = \boldsymbol{p}^*$



(a) Brier score; (b) Pairwise Gambling (PG) score; (c) Log score; (d) Full Gambling (FG) score

The Brier and Log score are **proper** and always favor the data generating model. However, they penalize differently forecasts close to zero. The **Log score** is more strict and prefers $\boldsymbol{p}_0$ over $\boldsymbol{p}_2$ when $\boldsymbol{p}_2$ approaches zero. The **Brier score** instead prefers $\boldsymbol{p}_2$ for any value of $\gamma < 5$

Both the PG and FG score are **improper** in this case and favor forecasts underestimating $\boldsymbol{p}^*$

Experimental Design

A way to conduct earthquake forecasting experiments is to select a region W, divide it in N equally sized bins $b_1, \dots, b_N$ and ask modelers to provide their probabilities of activity $\boldsymbol{p_i} = p_{i1}, \dots, p_{iN}$.

The results will be influenced by the choise of W and the binning $b_1, \dots, b_N$. Choosing W represents choosing the **amount of data** needed. Choosing the bins $b_1, \dots, b_N$ represents choosing **how we partition** (use) the data at hand.

The choice of the amount of data and the how to partion it influences the **probability of distinguishing** (expressing a preference) between different models. Having a **region too small** as well as **bins too wide** lower the probability of expressing a preference. However, we cannot calculate this probability directly, because we do not know $\boldsymbol{p}^*$

If we have the forecast in a ***catalog-based*** format we can use the simulations to estimate, for each model $\boldsymbol{p_i}$, *the probability of expressing a preference for* $\boldsymbol{p_i}$ ***when*** $\boldsymbol{p_i}$ ***is the data generating model.***

Studying this probabilities provides information about **the ability of the score of distinguishing** between models in terms of amount of data needed and how to partition it.

Experimental Design – Example Temporal-ETAS

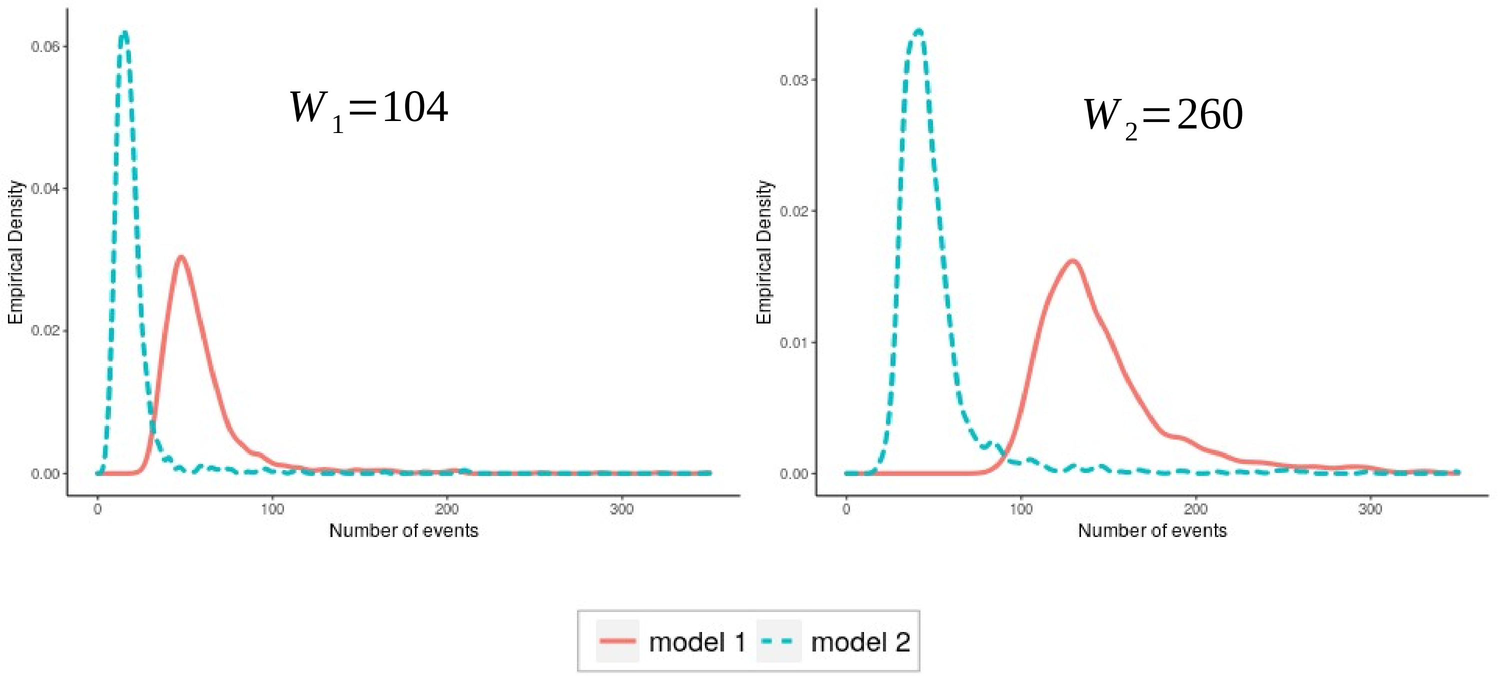
Take **two Temporal-ETAS** model with intensity given by:

$$\lambda(t|H_t)=\mu+K\sum_{i:t_i<t}\exp(\alpha(m_i-M_0))\frac{c^{p-1}(p-1)}{(t-t_i+c)^p}$$

The **first model** has parameters $\mu_1, K_1, \alpha_1, c_1, p_1$ equal to the Maximum Likelihood estimates calculated using data from the *Hauksson relocated catalogue* for California between 2000 and 2010 and magnitude greater than $M_0=3.95$. Time is expressed in weeks

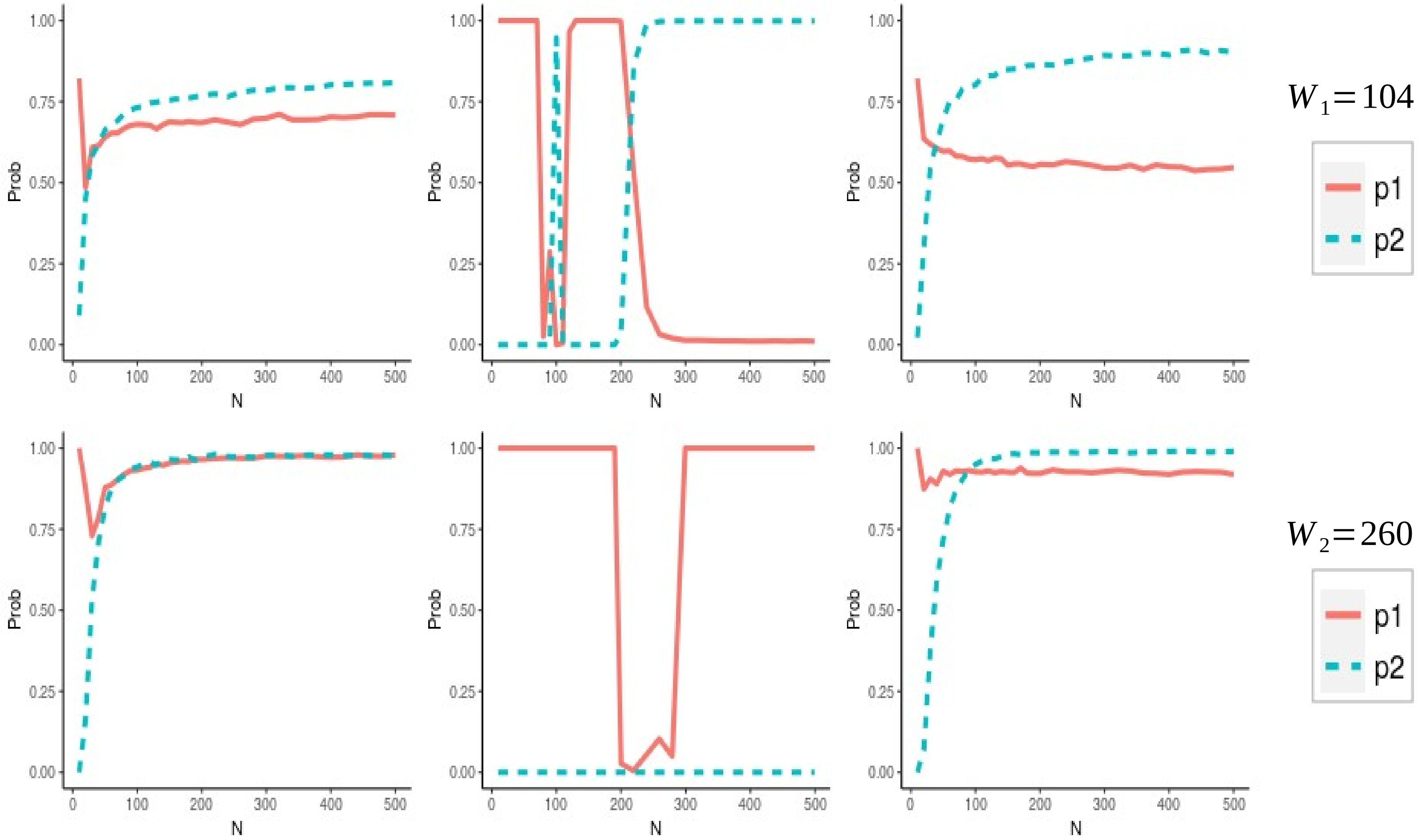
The second differes from the first only by the value of one parameter. The **second model** has a smaller background rate $\mu_2=\mu_1/3$

We have simulated data for two different regions: $W_1=104$ weeks corresponding to 2 years and $W_2=260$ weeks corresponding to 5 years



Experimental Design – Example Temporal-ETAS

Log score | Pairwise Gambling score | Full Gambling score



Considering **more bins** seems to have a **positive effect** (up to certain degree) on the probability of distinguishing between models. This is because we differentiate only between *active* and *non-active* bins. Everytime there is more than one event in a bin we are **throwing away information**.

The total amount of information provided by the data depends on the **extent** of the region. Considering **larger regions** has a **positive effect** on the probability of distinguishing between models and reduce the difference between cases.

Summary

- It is important for a score to be proper. Different proper scores provides different way to penalize the forecast but they always favor the data generating model. Improper scores may biased towards under/over
- The parimutuel gambling score for k > 3 forecast and the parimutuel gambling score for k = 2 with a reference model are improper. The bias depends on the average forecast.
- Ranking earthquakes can be seen as an estimation problem and confidence interval for the score difference between two models can be used to express (or not) a preference.
- Using simulated data from the models we can explore the probability of expressing a preference for a model when it is the data generating one.
- Studying this probability may provide insights into the amount of data needed and how to partition it in order to maximize the probability of distinguishing between models.

Contacts:

Email : francesco.serafini@ed.ac.uk
Preprint : https://arxiv.org/abs/2105.12065
Code : https://github.com/Serra314/Serra314.github.io/tree/master/Ranking_earthquake_forecast