Comparing Sequential Forecasters

Yo Joong (YJ) Choe and Aaditya Ramdas

Department of Statistics and Data Science, Machine Learning Department, Carnegie Mellon University

Motivation

You're watching the World Series, and you find two forecasters, *FiveThirtyEight* and *Vegas*, sequentially making predictions on each game's outcome.

How can you tell if one forecaster has made (significantly) better forecasts than the other?

2019 World Series (WSN vs. HOU)	Game 1	2	3	4	5	6	7
FiveThirtyEight	38%	41%	53%	59%	37%	41%	48%
Vegas Betting Odds	35%	38%	41%	51%	34%	37%	43%
Difference	*3%*	*3%*	***12%***	***8%***	*3%*	*4%*	*5%*
WSN Result	Win	Win	Loss	Loss	Loss	Win	Win

Figure: Probability forecasters on the outcome of the 2019 World Series games (WSN vs. HOU). Green means that their forecasts were correct; Red means they were incorrect (threshold at 50%).

This problem extends far beyond sports: properly evaluating and comparing sequential forecasters can be crucial in meteorology, epidemiology, elections, economics, finance, and more.

A Game-Theoretic Setup

Inspired by *game-theoretic statistics* [1], we define a **forecast comparison game** with which we develop our sequential inference methods:

Game (Comparing Forecasters)

Let $\mathscr{P}$ denote the space of probability distributions on an outcome space $\mathscr{Y}$. For rounds $t = 1, 2, \ldots$:

1. *Forecaster 1 makes their forecast $p_t \in \mathscr{P}$.*
2. *Forecaster 2 makes their forecast $q_t \in \mathscr{P}$.*
   (Steps 1 and 2 are in an arbitrary order.)
3. *Reality chooses $r_t \in \mathscr{P}$.*
   (r_t is hidden from the forecasters.)
4. *$y_t \sim r_t$ is sampled and revealed.*

This game is *sequential* in nature, and *no* assumptions are placed on the behaviors (dynamics) of Reality or Forecasters.

Objectives

The game-theoretic setup helps us develop forecast comparison methods that are:

1. **Time-Uniform a.k.a. Anytime-Valid:** validity under continuous monitoring and at all (data-dependent) stopping times;
2. **"Distribution-Free":** no assumptions on the time-varying dynamics of $(r_t)_{t\geq 1}$;
3. **Model-Free:** no assumptions on the forecasts $(p_t)_{t\geq 1}$ and $(q_t)_{t\geq 1}$; and
4. Reflective of the ***average* predictive ability over time**, as opposed to uniform dominance over time.

Our methods also apply to general types of forecasts (probability, functional, and distribution) and outcomes (binary, multiclass, and continuous), as long as the evaluation metric (scoring rule) is bounded.

Parameters of Interest: Average Score Differentials $(\Delta_t)_{t\geq 1}$.

Let $S : \mathscr{P} \times \mathscr{Y} \to \mathbb{R}$ be any **scoring rule** that evaluates probabilistic forecasts. Higher scores imply better forecasts.
Given a scoring rule S, we estimate the time-varying **average score differentials** $(\Delta_t)_{t\geq 1}$ between the two forecasters:

$$\Delta_t := \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}_{i-1}\left[S(p_i, y_i) - S(q_i, y_i)\right],$$

where $\mathbb{E}_{i-1}$ is the conditional expectation w.r.t. the past (filtration of the game up to round $i-1$).
In the following, $(\hat{\Delta}_t)_{t\geq 1}$ denotes its empirical estimate without the conditional expectation.

Examples of scoring rules. Let $y \in \{0, 1\}$ be a binary outcome and $p \in [0, 1]$ a probability forecast on y.

- *Brier Score*: $S(p, y) = 1 - (p - y)^2$.
- *Zero-One Score (i.e., Accuracy)*:
  $S(p, y) = \mathbf{1}(p \geq 1/2)y + \mathbf{1}(p < 1/2)(1 - y)$.

Other (bounded) scoring rules can also be used, depending on the types of forecasts and outcomes.

Confidence Sequences & E-Processes

A **confidence sequence (CS)** [2] $(C_t)_{t\geq 1}$ is a sequence of confidence intervals (CI) that *uniformly* cover a time-varying parameter *at all times* ("time-uniform"):

$$\mathbb{P}(\forall t \geq 1 : \Delta_t \in C_t) \geq 1 - \alpha,$$

given a significance level $\alpha \in (0, 1)$. This differentiates a CS from a fixed-time CI, whose coverage guarantee is limited to a pre-specified sample size.

An **e-process** [3] measures the amount of evidence accumulated against the null; an **e-value** is simply a realization of an e-process at any given time. More formally, an e-process $(E_t)_{t\geq 0}$ for a (composite) null H_0 is a nonnegative adapted process with $E_0 = 1$ and

$$\mathbb{E}_P[E_\tau] \leq 1, \quad \forall \text{stopping time } \tau, \ \forall P \in H_0.$$

Under H_0, an e-process is bounded by 1 at arbitrary stopping times ("anytime-valid"), and it will only grow large if we obtain evidence against the null.

In our setting, we consider the one-sided null $H_0^{\mathrm{w}} : \Delta_t \leq 0, \forall t$, so higher e-values imply that there is more evidence favoring Forecaster 1 over 2.

Key Result: Anytime-Valid and Distribution-Free Sequential Forecast Comparison.

Suppose that $|S(p_i, y_i) - S(q_i, y_i)|$ are bounded a.s. (e.g., Brier and zero-one scores). Then,

$$C_t := \left(\hat{\Delta}_t \pm t^{-1} u_{\alpha/2}(\hat{V}_t)\right) \text{ forms a } \mathbf{(1-\alpha)\text{-CS}} \text{ for } \Delta_t, \quad \forall \alpha \in (0, 1),$$

where $u_{\alpha/2}$ is a sub-exponential uniform boundary [2] and $\hat{V}_t$ is an empirical estimate of the variance process. This is an example of a *variance-adaptive* CS, and its width shrinks to zero at a $O(1/\sqrt{t})$ rate, up to log factors.

Furthermore, given the null hypothesis $H_0^{\mathrm{w}} : \Delta_t \leq 0, \forall t$ (saying "p is no better than q on average"),

$$E_t(\lambda) := \exp\left\{\lambda t \hat{\Delta}_t - \psi_E(\lambda)\hat{V}_t\right\} \text{ is an } \mathbf{e\text{-process}} \text{ for } H_0^{\mathrm{w}}, \quad \forall \lambda \in [0, \lambda_{\max}),$$

where $\psi_E(\lambda) = -\log(1-\lambda) - \lambda$ is the exponential cumulant generating function (CGF).

Applications: Comparing Baseball & Weather Forecasters

FiveThirtyEight vs. Betting Odds on MLB Games

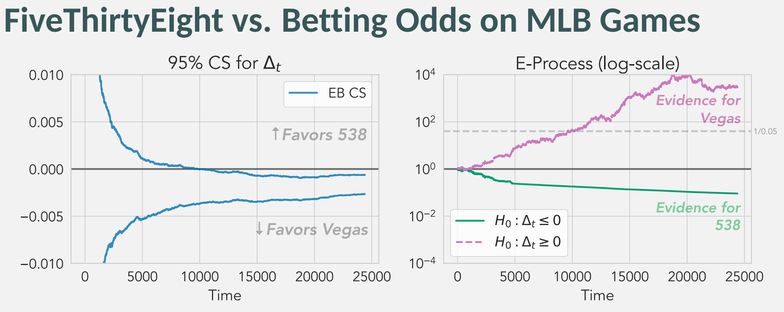

Figure: 95% CS and e-processes for comparing FiveThirtyEight's win probability forecasts and Vegas betting odds on Major League Baseball (MLB) games. Scoring rule is the Brier score. Data includes all regular and post-season games from 2010 to 2019.

Comparing Ensemble Weather Forecasters

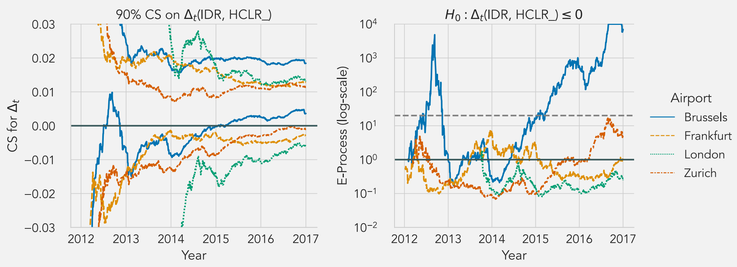

Figure: 90% CSs and e-processes for comparing pairs of 1-day probability of precipitation (PoP) forecasters from 2012 to 2017 at four locations. Scoring rule is the Brier score. IDR and HCLR_ are two statistical postprocessing methods for predicting the PoP of the next day. Data & forecasts from [4].

References

[1] G. Shafer and V. Vovk, "Game-theoretic foundations for probability and finance," *Wiley*, 2019.

[2] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon, "Time-uniform, nonparametric, nonasymptotic confidence sequences," *The Annals of Statistics*, 2021.

[3] A. Ramdas, J. Ruf, M. Larsson, and W. M. Koolen, "Testing exchangeability: Fork-convexity, supermartingales and e-processes," *Int. J. Approx. Reason.*, 2021.

[4] A. Henzi and J. F. Ziegel, "Valid sequential inference on probability forecast performance," *Biometrika*, 2021.

[5] T. L. Lai, S. T. Gross, and D. B. Shen, "Evaluating probability forecasts," *The Annals of Statistics*, 2011.

Find Us Online!

- Paper: arxiv.org/abs/2110.00115
- Code: github.com/yjchoe/ComparingForecasters
- Email: {yjchoe,aramdas}@cmu.edu