Forecaster's dilemma: Extreme events and forecast evaluation

Sebastian Lerch

Karlsruhe Institute of Technology

Heidelberg Institute for Theoretical Studies

9th ECB Workshop on Forecasting Techniques
Frankfurt, June 4, 2016

joint work with Thordis Thorarinsdottir, Francesco Ravazzolo and Tilmann Gneiting

Heidelberg Institute for Theoretical Studies

HITS

Karlsruhe Institute of Technology

WAVES TO WEATHER

Motivation

THE SPECTATOR

HOME | COFFEE HOUSE | ELECTION 2015 | MAGAZINE | COLUMNISTS | CULTURE HOUSE | PODCAST

The Week | Features | Columnists | Books | Arts | Life | Cartoons | Classified

Forecast failure: how the Met Office lost touch with reality

Ideology has corrupted a valuable British institution

Rupert Darwall 13 July 2013 — 118 Comments

http://www.spectator.co.uk/features/8959941/whats-wrong-with-the-met-office/

Outline

1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events
4. Case study and simulation results

Outline

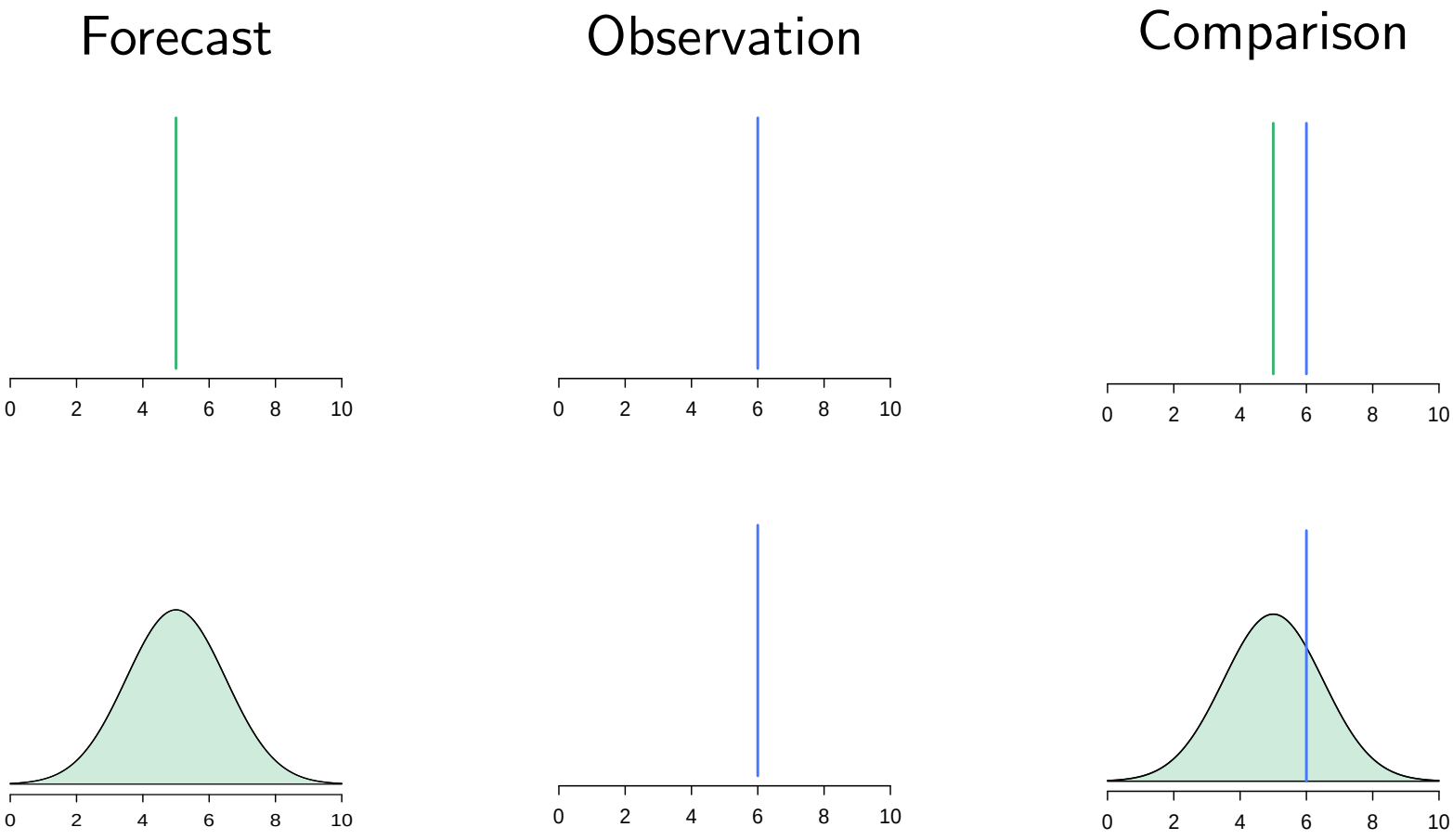
1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events
4. Case study and simulation results

Probabilistic forecasts

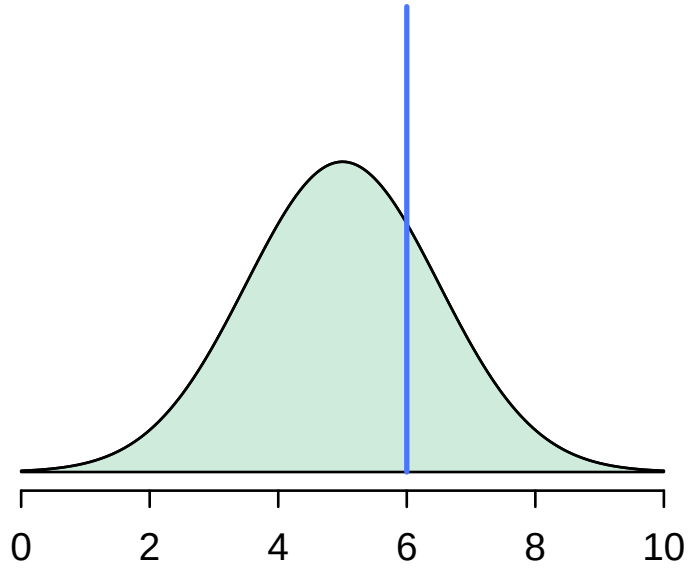
Probabilistic forecasts, i.e., forecasts in the form of probability distributions over future quantities or events,

- provide information about inherent uncertainty
- allow for optimal decision making by obtaining deterministic forecasts as target functionals (mean, quantiles, ...) of the predictive distributions
- have become increasingly popular across disciplines: meteorology, hydrology, seismology, economics, finance, demography, political science, ...

Probabilistic vs. point forecasts



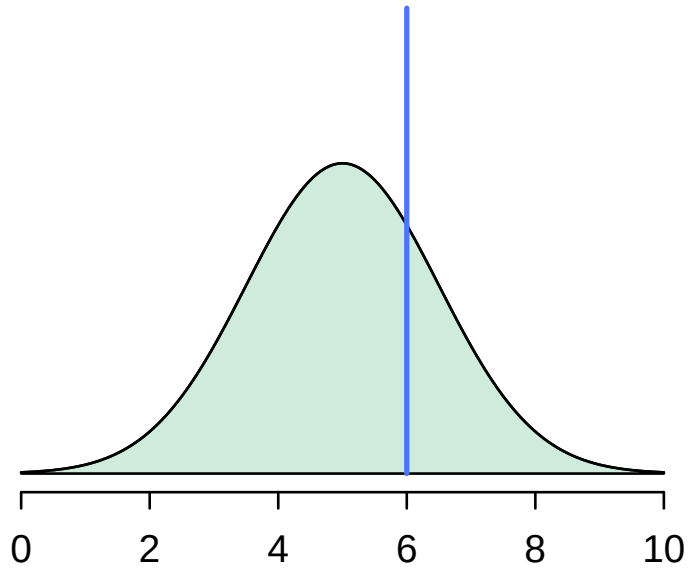
What is a good probabilistic forecast?



The goal of probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007) **Probabilistic forecasts, calibration and sharpness**. *Journal of the Royal Statistical Society Series B*, 69, 243–268.

Calibration and sharpness



Calibration: Compatibility between the forecast and the observation; joint property of the forecasts and observations

Sharpness: Concentration of the forecasts; property of the forecasts only

Evaluation of probabilistic forecasts: Proper scoring rules

A proper scoring rule is any function

$$S(F, y)$$

such that

$$\mathbb{E}_{Y \sim G} S(G, Y) \leq \mathbb{E}_{Y \sim G} S(F, Y)$$

for all $F, G \in \mathcal{F}$.

We consider scores to be negatively oriented penalties that forecasters aim to minimize.

Proper scoring rules prevent hedging strategies.

Gneiting, T. and Raftery, A. E. (2007) **Strictly proper scoring rules, prediction, and estimation**. *Journal of the American Statistical Association*, 102, 359–378.

Examples

Popular examples of proper scoring rules include

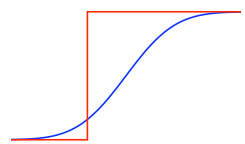
- the logarithmic score

$$\mathrm{LogS}(F, y) = -\log(f(y)),$$

  where f is the density of F,

- the continuous ranked probability score

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 dz,$$



  where the probabilistic forecast F is represented as a CDF.

Outline

1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events
4. Case study and simulation results

Media attention often exclusively falls on prediction performance in the case of extreme events

Bad Data Failed To Predict Nashville Flood	NBC, 2011
Weather Service Faulted for Sandy Storm Surge Warnings	NBC, 2013
How Did Economists Get It So Wrong?	NY Times, 2009
Nouriel Roubini: The economist who predicted worldwide recession	Guardian, 2009
An exclusive interview with Med Yones - The expert who predicted the financial crisis	CEOQ Mag, 2010
A Seer on Banks Raises a Furor on Bonds	NY Times, 2011

Toy example

We compare Alice's and Bob's forecasts for $Y \sim \mathcal{N}(0, 1)$,

$$F_{\text{Alice}} = \mathcal{N}(0, 1), \qquad F_{\text{Bob}} = \mathcal{N}(4, 1)$$

Based on all 10 000 replicates,

Forecaster	CRPS	LogS
Alice	**0.56**	**1.42**
Bob	3.53	9.36

When the evaluation is restricted to the largest ten observations,

Forecaster	CRPS	LogS
Alice	2.70	6.29
Bob	**0.46**	**1.21**

Verifying only the extremes erases propriety

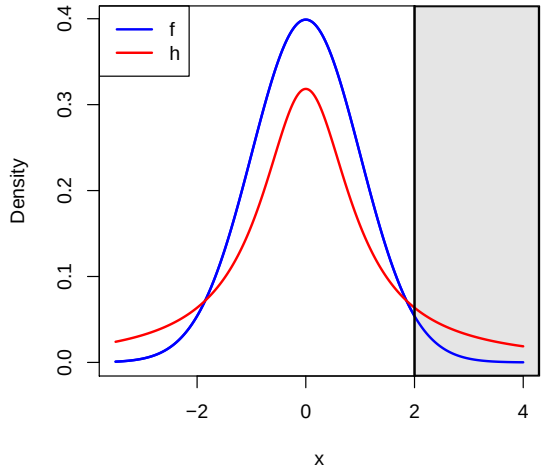
Some econometric papers use the restricted logarithmic score

$$\text{R-LogS}_{\geq r}(F, y) = -\mathbb{1}\{y \geq r\} \log f(y).$$

However, if $h(x) > f(x)$ for all $x \geq r$, then

$$\mathbb{E}\,\text{R-LogS}_{\geq r}(H, Y) < \mathbb{E}\,\text{R-LogS}_{\geq r}(F, Y)$$

independent of the true density.



Indeed, if the forecaster's belief is F, her best prediction under $\text{R-LogS}_{\geq r}$ is

$$f^*(z) = \frac{\mathbb{1}(z \geq r) f(z)}{\int_r^\infty f(x)dx}.$$

The forecaster's dilemma

Given any (non-trivial) proper scoring rule S and any non-constant weight function w, any scoring rule of the form

$$S^*(F, y) = w(y)S(F, y)$$

is improper.

The expected value $\mathbb{E}_{Y\sim G}S^*(F, y)$ is minimized for

$$f^*(z) = \frac{w(z)g(z)}{\int w(x)g(x)dx}.$$

Forecaster's dilemma: Forecast evaluation based on a subset of extreme observations only corresponds to the use of an improper scoring rule and is bound to discredit skillful forecasters.

Outline

1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events
4. Case study and simulation results

Proper weighted scoring rules I

Proper weighted scoring rules provide suitable alternatives.

Gneiting and Ranjan (2011) propose the threshold-weighted CRPS

$$\text{twCRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 w(z)\, \mathrm{d}z$$

$w(z)$ is a weight function on the real line.

Gneiting, T. and Ranjan, R. (2011) **Comparing density forecasts using threshold- and quantile-weighted scoring rules**. *Journal of Business and Economic Statistics*, 29, 411–422.

Proper weighted scoring rules II

Diks et al. (2011) propose the conditional likelihood score,

$$\mathrm{CL}(F, y) = -w(y) \log \left(\frac{f(y)}{\int w(z) f(z)\, \mathrm{d}z} \right),$$

and the censored likelihood score,

$$\mathrm{CSL}(F, y) = -w(y) \log f(y) - (1 - w(y)) \log \left(1 - \int w(z) f(z)\, \mathrm{d}z \right).$$

$w(z)$ is a weight function on the real line.

Diks, C., Panchenko, V. and van Dijk, D. (2011) **Likelihood-based scoring rules for comparing density forecasts in tails**. *Journal of Econometrics*, 163, 215–233.

Role of the weight function

The weight function w can be tailored to the situation of interest.

For example, if interest focuses on the predictive performance in the right tail,

$$w_{\text{indicator}}(z) = \mathbb{1}\{z \geq r\}, \text{or}$$
$$w_{\text{Gaussian}}(z) = \Phi(z|\mu_r, \sigma_r^2)$$

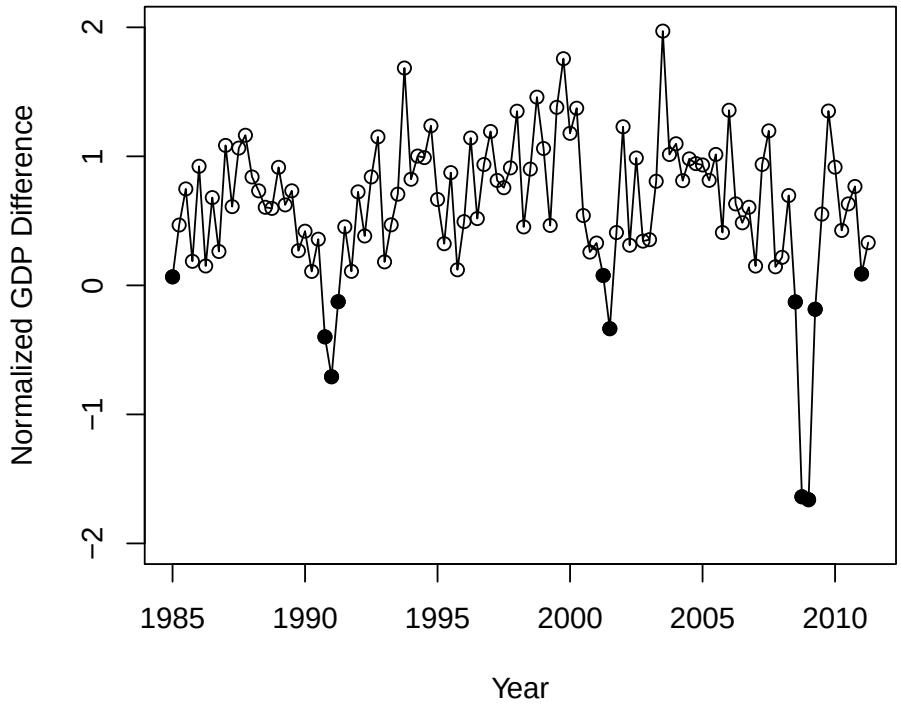
Choices for the parameters r, μ_r, σ_r can be motivated and justified by applications at hand.

Outline

1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events
4. Case study and simulation results

Case study: Macroeconomic forecasting

- Probabilistic forecasts of quarterly GDP growth for the U.S.
- Evaluation period 1985 – 2011.
- Prediction horizon of 1 and 4 quarters ahead.



Clark, T. E. and Ravazzolo, F. (2015) **Macroeconomic forecasting performance under alternative specifications of time-varying volatility.** *Journal of Applied Econometrics*, 30, 551–575.

Models for GDP growth

- Baseline autoregressive (AR) model:

$$Y_t \,|\, y_{<t}, b_0, \ldots, b_p, \sigma \sim \mathcal{N}\left(b_0 + \sum_{i=1}^{p} b_i y_{t-i}, \sigma^2 \right)$$

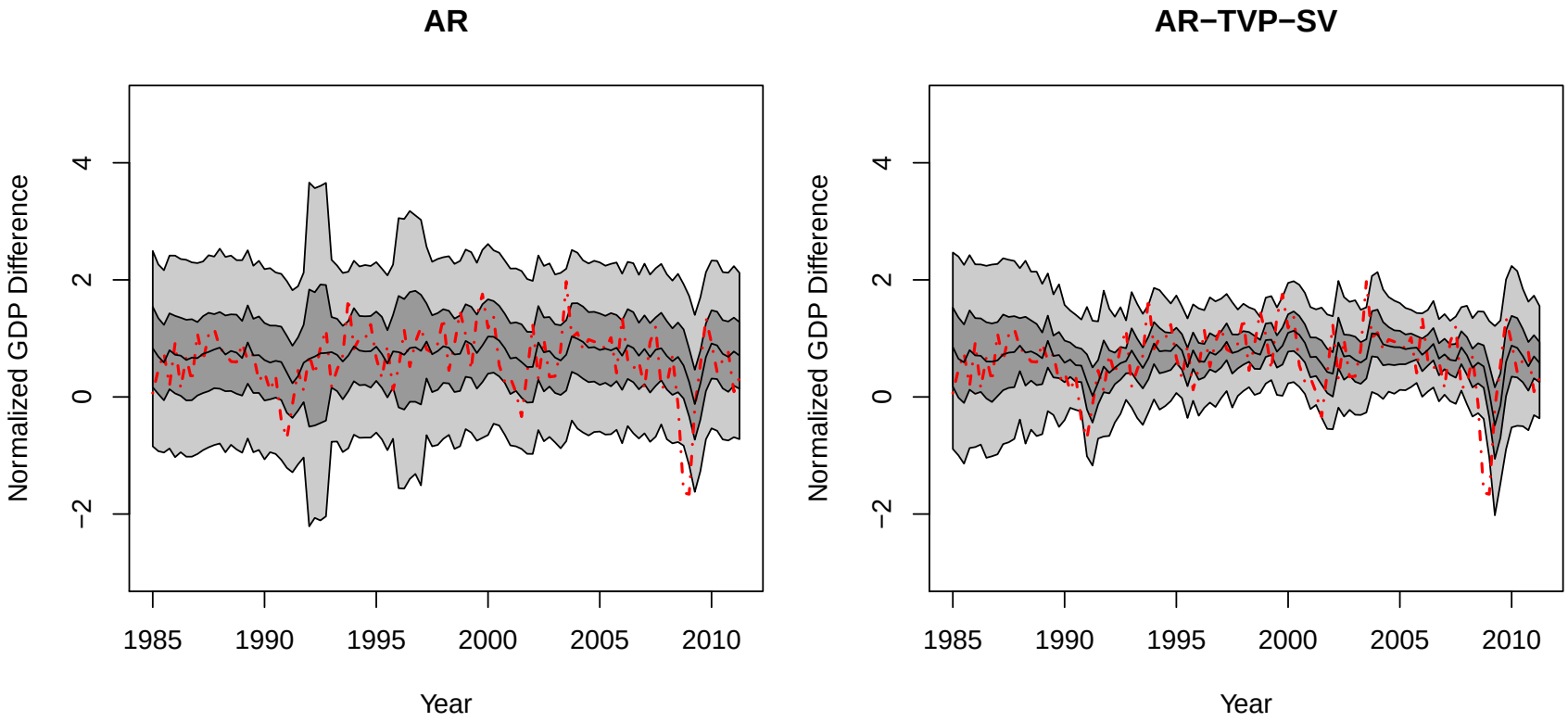
- AR-TVP-SV model of Cogley and Sargent (2005)

$$Y_t \,|\, y_{<t}, b_{0,t}, \ldots, b_{p,t}, \lambda_t \sim \mathcal{N}\left(b_{0,t} + \sum_{i=1}^{p} b_{i,t} y_{t-i},\ \lambda_t \right),$$

$$b_{i,t} \,|\, b_{i,t-1}, \tau \sim \mathcal{N}\big(b_{i,t-1}, \tau^2\big), \quad i = 0, \ldots, p,$$

$$\log \lambda_t \,|\, \lambda_{t-1}, \sigma \sim \mathcal{N}\big(\log \lambda_{t-1}, \sigma^2\big).$$

Probabilistic 1-quarter ahead forecasts of GDP growth



Verification

Based on all observations,

Model	CRPS		LogS	
	$h = 1$	$h = 4$	$h = 1$	$h = 4$
AR	0.330	0.359	1.044	1.120
AR-TVP-SV	**0.292**	**0.329**	**0.833**	**1.019**

When the evaluation is restricted to observations ≤ 0.1,

Model	R-CRPS$_{\leq 0.1}$		R-LogS$_{\leq 0.1}$	
	$h = 1$	$h = 4$	$h = 1$	$h = 4$
AR	**0.654**	**0.870**	**1.626**	**2.010**
AR-TVP-SV	0.659	0.970	2.016	3.323

Verification with proper weighted scoring rules

$$\text{twCRPS} = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 w(z)\, \mathrm{d}z$$

$$w_{\text{indicator}}(z) = \mathbb{1}\{z \leq 0.1\}$$
$$w_{\text{Gaussian}}(z) = 1 - \Phi(z \,|\, \mu_r = 0.1, \sigma_r^2 = 1)$$

Model	twCRPS			
	$w_{\text{indicator}}$		w_{Gaussian}	
	$h = 1$	$h = 4$	$h = 1$	$h = 4$
AR	0.062	0.068	0.111	0.120
AR-TVP-SV	**0.052**	**0.062**	**0.101**	**0.115**

Diebold-Mariano tests

Formal test of equal predictive performance of F_t and G_t for an observation y_{t+k} that lies k time steps ahead.

Denote the mean scores on a test set ranging from $t = 1, \ldots, n$ by

$$\bar{S}_n^F = \frac{1}{n} \sum_{t=1}^{n} S(F_t, y_{t+k}) \qquad \text{and} \qquad \bar{S}_n^G = \frac{1}{n} \sum_{t=1}^{n} S(G_t, y_{t+k}),$$

Diebold-Mariano test: Under the null hypothesis of a vanishing expected score difference and standard regularity conditions

$$t_n = \sqrt{n} \, \frac{\bar{S}_n^F - \bar{S}_n^G}{\hat{\sigma}_n}$$

is asymptotically standard normal.

$\hat{\sigma}_n^2$ is an estimator of the asymptotic variance of the score difference.

Diebold, F. X. and Mariano, R. S. (1995) **Comparing predictive accuracy**. *Journal of Business and Economics Statistics*, 13, 253–263.

Simulation study: Setting

Motivation: Neyman-Pearson lemma suggests superiority of tests of equal predictive performance based on unweighted LogS.

Simulation setting tailored to benefit proper weighted scoring rules

- compare two forecast distributions neither of which corresponds to the true sampling distribution
- forecast distributions only differ on the positive half-axis
- sample size is fixed at $n = 100$

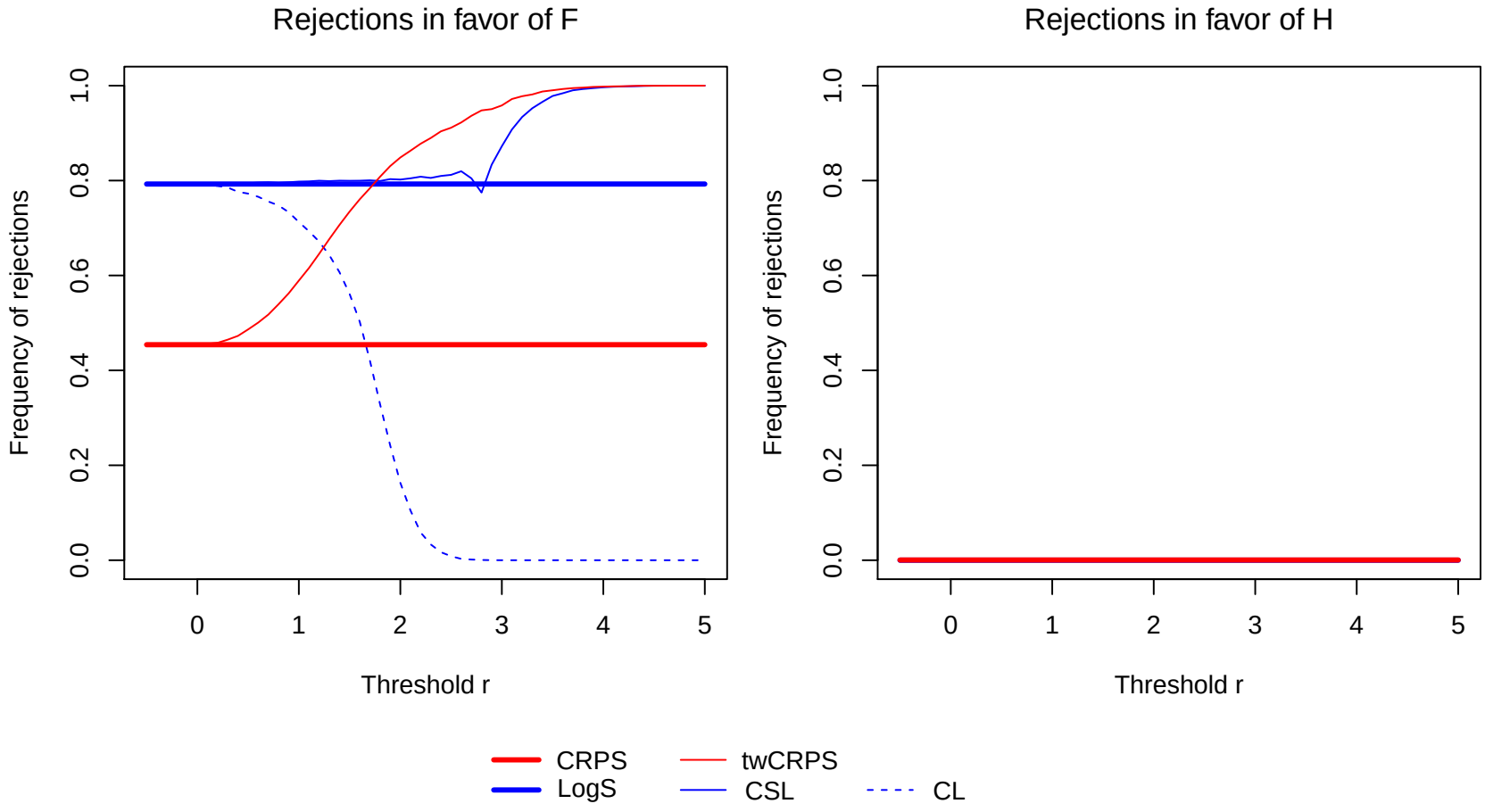
Compare three forecast distributions with densities:

- $\phi(x)$, standard normal density,
- $h(x) = \mathbb{1}\{x \leq 0\}\,\phi(x) + \mathbb{1}\{x > 0\}\,t_4(x)$,
- $f(x) = 0.5\,\phi(x) + 0.5\,h(x)$

using two-sided DM tests.

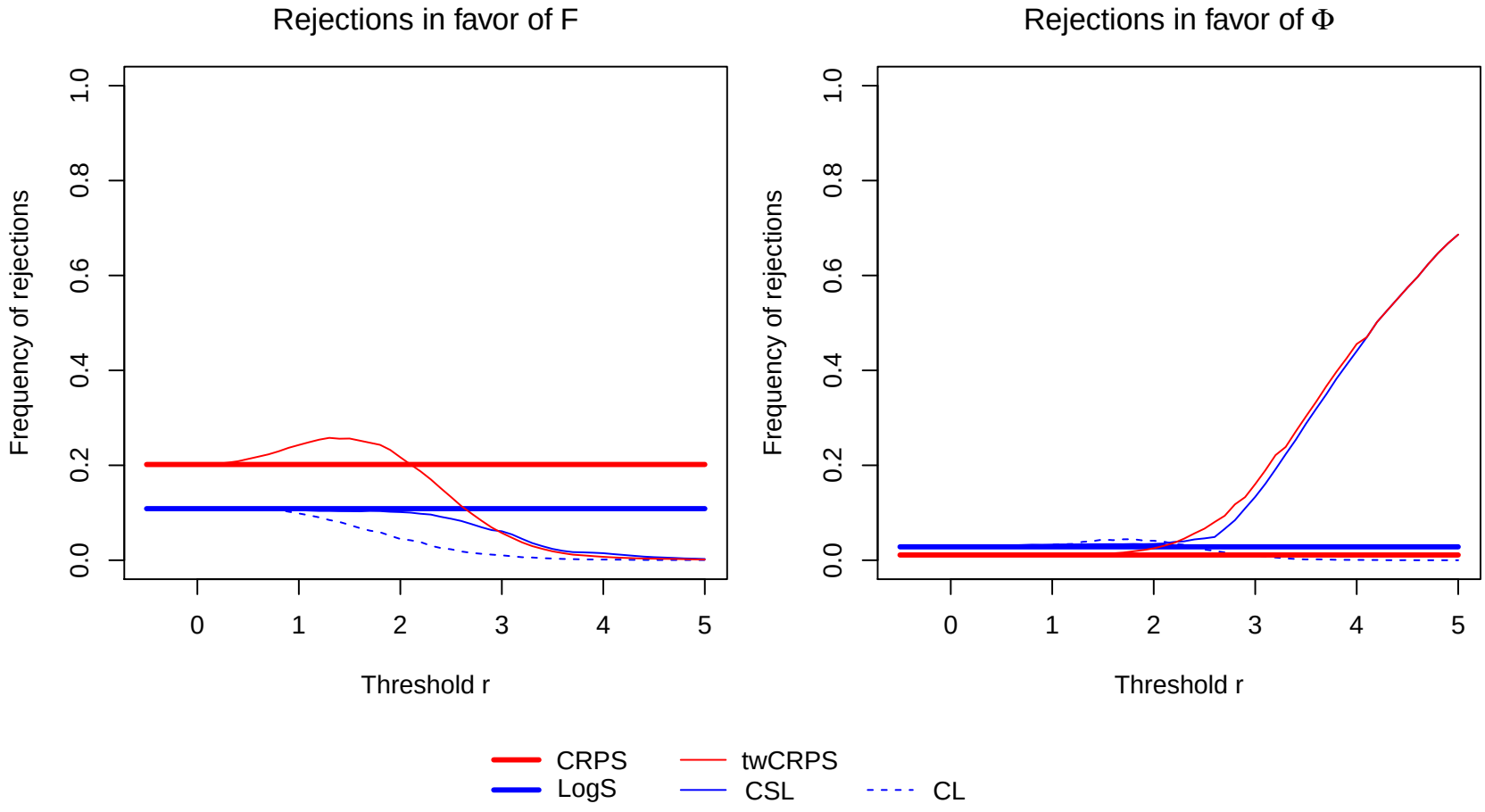
Simulation study: Variant 1

Truth $= \Phi$, compare F and H (F should be preferred)



Simulation study: Variant 2

Truth $= H$, compare F and Φ (F should be preferred)



Tail dependence of proper weighted scoring rules

Consider $w_r(z) = \mathbb{1}\{z \geq r\}$ and a threshold r such that $y_i < r$ for all $i = 1, \ldots, n$.

Then all proper weighted scoring rules do not depend on the observations and are solely determined by the tail probabilities.

$$\begin{aligned}
\overline{\mathrm{CL}}_n^F &= 0 \\
\overline{\mathrm{CSL}}_n^F &= -\log F(r) \\
\overline{\mathrm{twCRPS}}_n^F &= \int_r^\infty (F(z) - 1)^2 \, \mathrm{d}z
\end{aligned}$$

The forecast distribution with the lighter tail then receives the better score, irrespectively of the true distribution.

Summary and conclusions

- Forecaster's dilemma: Verification on extreme events only is bound to discredit skillful forecasters.
- The only remedy is to consider all available cases when evaluating predictive performance.
- Proper weighted scoring rules emphasize specific regions of interest, such as tails, and facilitate interpretation, while avoiding the forecaster's dilemma.
- In particular, the weighted versions of the CRPS share (almost all of) the desirable properties of the unweighted CRPS.
- Practical benefits of using proper weighted scoring rules in terms of power may be limited.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2015) **Forecaster's dilemma: Extreme events and forecast evaluation**. Preprint available at http://arxiv.org/abs/1512.09244.

Thank you for your attention!