Focusing on regions of interest in forecast evaluation

Hajo Holzmann & Bernhard Klar
Philipps-University Marburg & Karlsruhe Institute of Technology
holzmann@mathematik.uni-marburg.de, bernhard.klar@kit.edu

Introduction

- **Forecast evaluation** of **probability forecasts** often focuses on certain **regions of interest**
  - *Risk management:* requires appropriate loss distribution forecasts in the tails.
  - *Weather forecasts* with a focus on extreme conditions.
  - Forecasts of *environmental variables* such as ozone with a focus on concentration levels with adverse health effects.
- **Forecast ranking** according to performance within these regions.
- Show how **weighted scoring rules** can be used to this end
- allow to rank **potentially misspecified forecasts objectively** with the region of interest in mind.
- Discuss **theoretical properties** of weighted scoring rules and present **construction principles.**

Previous work

- [1]: conditional likelihood and censored likelihood score
- [2]: threshold-weighted and quantile-weighted continuous-ranked probability score
- [5]: penalized weighted likelihood score, theoretical properties of weighted scoring rules
- [4]: discuss forecaster's dilemma, cast doubts on the usefulness of weighted scoring rules

Motivating simulation

Goal: Demonstrate that weighted scoring rules useful for comparing two misspecified forecasts.

- Data: i.i.d., standard normally distributed
- F_{hlt}: piecewise defined, continuous, scaled t_4-distribution on $(-\infty, 0]$, standard normal distribution on $(0, \infty)$
- F_{hrt}: roles reversed
- Censored likelihood rule (CSL):

$$S^{CSL}(p, x; r) = \begin{cases} -\log p(x), & \text{if } x \geq r, \\ -\log\big(1 - \int_{-\infty}^{r} p(z)\mathrm{d}z\big), & \text{if } x < r. \end{cases}$$

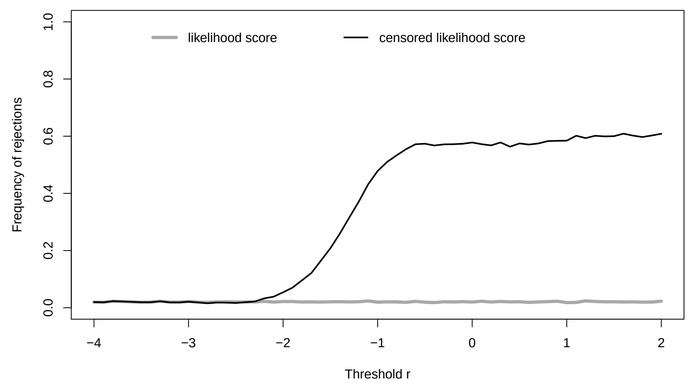

Figure 1: Frequency of rejections in two-sided Diebold-Mariano test in favor of F_{hlt} for the logarithmic (LogS) and the censored likelihood (CSL) scoring rules for sample size $n = 100$.

Conclusion: If region of interest is of form $[r, \infty)$ for some $r \geq -1$, censored likelihood score can discriminate between F_{hlt} and F_{hrt}.

Weighted scoring rules

Definitions and theoretical properties

- Observational space $(\mathcal{X}, \mathcal{F})$, family of distributions $\mathcal{M}$, family of weight functions $\mathcal{W}$ consisting of $w : \mathcal{X} \to [0, 1]$.
- **Weighted scoring rule**: a map $S : \mathcal{M} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}$ such that $S(\cdot, \cdot; w)$ is a scoring rule for each $w \in \mathcal{W}$.
- S **localizing**: if for any $P_1, P_2 \in \mathcal{M}$,

$$\forall\, F \in \mathcal{F}: \; P_1\big(\{w > 0\} \cap F\big) = P_2\big(\{w > 0\} \cap F\big) \implies S(P_1, x; w) = S(P_2, x; w) \text{ for all } x \in \mathcal{X}.$$

The condition means that the restrictions of P_i to $\{w > 0\}$ coincide, for $i = 1, 2$. Then also

$$\forall\, Q \in \mathcal{M}: \quad S(P_1, Q; w) = S(P_2, Q; w) = \int_{\mathcal{X}} S(P_2, x; w)\, \mathrm{d}Q(x).$$

- S **proper**: if $S(\cdot, \cdot; w)$ proper for each $w \in \mathcal{W}$, i.e. $S(Q, Q; w) \leq S(P, Q; w)$, $P, Q \in \mathcal{M}$.
- S **strictly locally proper**: S is localizing and proper and if $S(P, Q; w) = S(Q, Q; w)$, then the restrictions of P and Q to $\{w > 0\}$ coincide necessarily.
- S **proportionally locally proper**: if $S(P, Q; w) = S(Q, Q; w)$ is equivalent to $P\big(\{w > 0\} \cap F\big) = c\, Q\big(\{w > 0\} \cap F\big)$, for all $F \in \mathcal{F}$ and a constant $c > 0$, which depends on $P, Q \in \mathcal{M}$.

Construction

Assuming that for all $w \in \mathcal{W}$ and $P \in \mathcal{M}$ we have $\int w\, \mathrm{d}P > 0$, and set

$$\mathrm{d}P_w(x) = \frac{w(x)\, \mathrm{d}P(x)}{\int w\, \mathrm{d}P}$$

the probability distribution with density proportional to w w.r.t. P, which is assumed to belong to a family $\widetilde{\mathcal{M}}$.

Theorem 1. Let $\widetilde{S} : \widetilde{\mathcal{M}} \times \mathcal{X} \to \overline{\mathbb{R}}$ be a proper scoring rule. Then

$$S : \mathcal{M} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}, \quad S(P, x; w) = w(x)\, \widetilde{S}(P_w, x)$$

is a localizing proper weighted scoring rule. Further, if $\widetilde{S}$ is strictly proper, then S is **proportionally locally proper**.

Examples. *Conditional likelihood score* from [1].
Weighted version of the *Hyvärinen score* (also multivariate)

$$S(p, x; w) = 2\frac{p''(x)}{p(x)}\, w(x) - \Big(\frac{p'(x)}{p(x)}\Big)^2 w(x) + 2\frac{p'(x)}{p(x)}\, w'(x).$$

Weighted version of the *CRPS* and *multivariate energy scores*

$$\mathrm{wCRPS}(F, x; r) = 1\{x > r\} \int_r^{\infty} \Big(\frac{F(z) - F(r)}{1 - F(r)} - 1\{x \leq z\}\Big)^2 \mathrm{d}z, \qquad w(x) = 1\{x > r\}.$$

Theorem 2. Let $\mathbf{s}(\alpha, z)$ be a strictly proper scoring rule for the success probability $\alpha \in (0, 1)$ of a binary outcome variable $z \in \{0, 1\}$. Then

$$S_{\mathbf{s}}(P, x; w) = w(x)\, \mathbf{s}\Big(\int w\, \mathrm{d}P, 1\Big) + \big(1 - w(x)\big)\, \mathbf{s}\Big(\int w\, \mathrm{d}P, 0\Big)$$

is a localizing proper weighted scoring rule for the probability forecast P.
If additionally $S(P, x; w)$ is a proportionally locally proper weighted scoring rule, then

$$\widehat{S}(P, x; w) = S_{\mathbf{s}}(P, x; w) + S(P, x; w)$$

is **strictly locally proper**.

Examples. *Censored likelihood score* (CSL) from [1].
Penalized weighted likelihood score (PWL) from [5].
Strictly locally proper weighted version of CRPS (wsCRPS):

$$\mathrm{wsCRPS}(F, x; r) = 1\{x > r\}\Big[F(r)^2 + \int_r^{\infty} \Big(\frac{F(z) - F(r)}{1 - F(r)} - 1\{x \leq z\}\Big)^2 \mathrm{d}z\Big] + 1\{x \leq r\}\,\big(1 - F(r)\big)^2.$$

Relation to hypothesis testing

- P_0, P_1: two competing (forecast) distributions for i.i.d. observations
- *Region of interest* A, assuming $0 < P_0(A), P_1(A) < 1$.
- Test composite hypothesis and alternative

$$H_0 : P = P_0 \quad \text{on } A \quad \text{vs.} \quad H_1 : P = P_1 \quad \text{on } A$$

using score-differences (Diebold-Mariano test) with localizing weighted scoring rule.

- Forecast P is only relevant for the hypotheses through observations $x \in A$. For $x \notin A$ only the total probability $1 - P(A)$ matters.
- *Censored likelihood rule:* optimal localizing weighted scoring rule in terms of power.

Empirical illustration

- Daily Deutsche Bank log returns $y_t = \ln(P_t/P_{t-1})$, from January 1, 2009 until December 31, 2016.
- *GARCH(1,1) model*, using normal, t and skew-t distributions for the *innovations*, one-step-ahead density forecasts with a rolling window scheme .

		w(x) = $1\{x \leq r\}$			w(x) = $1\{x \geq r\}$		
		$r = -3$	$r = -1$	$r = 0$	$r = 0$	$r = 1$	$r = 3$
	proportion	0.096	0.30	0.50	0.50	0.32	0.092
normal GARCH vs. t-GARCH	LogS	2.43	2.43	2.43	2.43	2.43	2.43
	CRPS	1.51	1.51	1.51	1.51	1.51	1.51
	CSL	1.89	1.71	1.96	1.63	1.73	0.95
	PWL	1.85	1.69	1.99	1.66	1.78	0.94
	wsCRPS	1.91	0.38	0.51	1.32	1.89	0.70
normal GARCH vs. skew-t-GARCH	LogS	2.18	2.18	2.18	2.18	2.18	2.18
	CRPS	1.22	1.22	1.22	1.22	1.22	1.22
	CSL	2.01	1.97	2.06	0.74	1.12	0.23
	PWL	1.96	1.94	2.13	0.83	1.18	0.24
	wsCRPS	1.67	1.26	0.63	0.44	0.80	-0.25
t-GARCH vs. skew-t-GARCH	LogS	-0.61	-0.61	-0.61	-0.61	-0.61	-0.61
	CRPS	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
	CSL	1.65	2.30	1.31	-2.10	-1.49	-1.76
	PWL	1.66	2.20	1.60	-2.03	-1.46	-1.72
	wsCRPS	0.53	1.45	0.07	-1.79	-0.96	-0.91

Table 1: t-statistics for Diebold-Mariano test: Positive values indicate superiority of forecasts from the second method.

Conclusions

- A weighted scoring rule allows to objectively decide in favor of a misspecified forecast which is
  - **superior** to a competing forecast on a **region of interest**,
  - even though it may be **inferior outside** this region.
- **General construction principle**, also multivariate and *without assuming densities.*
- **Optimal rule** for testing: **censored likelihood rule**.
- Poster based on [3].

References

[1] C. Diks, V. Panchenko, and D. van Dijk. Likelihood-based scoring rules for com- paring density forecasts in tails. *Journal of Econometrics*, 2011.

[2] T. Gneiting and R. Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 2011.

[3] H. Holzmann and B. Klar. Focusing on regions of interest in forecast evaluation. *Annals of Applied Statistics, to appear*.

[4] S. Lerch, T. La Thorarinsdottir, F. Ravazzolo, and T. Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 2016.

[5] J. Pelenis. Weighted scoring rules for comparison of density forecasts on subsets of interest. *Preprint*, 2014.