Localising Strictly Proper Scoring Rules

University of Amsterdam
Amsterdam School of Economics

tinbergen institute

Ramon de Punder, Cees Diks, Roger Laeven, Dick van Dijk

In Short

- In many applications it is desired to focus on particular regions of the distribution.
- Rules based on the conditional distribution on the region of interest are not strictly locally proper.
- Using the censored distribution instead leads to strictly locally proper scoring rules.
- Censoring maintains more information than conditioning, also beneficial for power:
  - Localised Neyman Pearson result for localised Neyman Pearson hypotheses.
  - Preferable power properties in Monte Carlo results Giacomini and White test.
  - Smaller Model Confidence Sets in an empirical risk management analysis.

Motivation

In contrast to point forecasts, density forecasts provide the full picture of the random variable of interest. For this reason, constructing and comparing them has gained much interest in the forecasting literature. In many applications, however, not all parts of the distribution are of equal importance. An example is risk management, where regulators are more interested in extreme losses (step in) than profits (send email to congratulate).

⇒ Need for a *i) localisation device* for scoring rules *ii) retaining strict propriety*.

1. Otherwise, excellent but irrelevant performance (right-tail) might overshadow poor but crucial performance (left-tail).
2. A minimal requirement for scoring rules is strict propriety. Hence preferably maintained.

Non-strictly locally proper scoring rules

1. Weighted likelihood score: $\text{wlog}(f, y) = w(y)f(y)$.
   Problem: **improper**. Assume $w(y) = \mathbf{1}_{y<r}$ and $f(y) > g(y)$ for $y < r$. Then, the expected score of f is larger than the expected score of g, even if g is the true density.
2. Conditional scoring rule: $S(f, y) = w(y)S(f_w^\sharp, y)$, $f_w^\sharp = f_w / \int f_w \mathrm{d}\mu$
   $f_w(y) = w(y)f(y)$.
   Problem: **non-strict**. Let $w(y) = \mathbf{1}_{y<r}$. Due to the normalisation in $f_w / \int f_w \mathrm{d}\mu$, any distribution proportional to f on $(-\infty, r)$ maps onto the same $f_w / \int f_w \mathrm{d}\mu = cf_w / \int cf_w \mathrm{d}\mu$, $c > 0$. See Panel (a).
3. Weighted CRPS: $\text{twCRPS}(F, y) = -\int_{r_1}^{r_2} (F(s) - 1_{y\le s})^2 \mathrm{d}s$.
   Problem: **Non-localising**. Due to the dependence on the CDF it takes into account information outside $[-r, r]$ that is not implied by the distribution on $[-r, r]$, potentially leading to a **localisation bias**. See Panel (b). Here, the twCRPS implies a score divergence indicating g to be statistically closer to p on $[-1, 1]$ than f. Since $p = f \neq g$ on $[-1, 1]$, the twCRPS is non-localising.

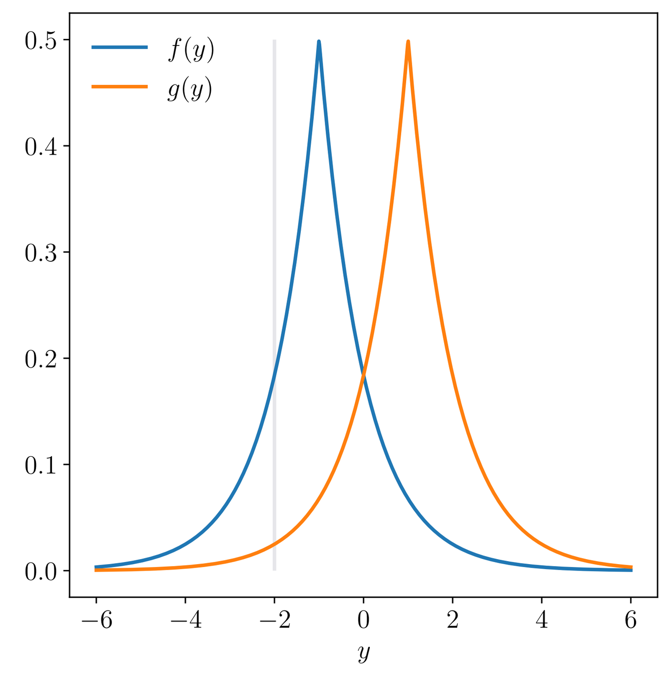

Figure 1. Proportional Laplace

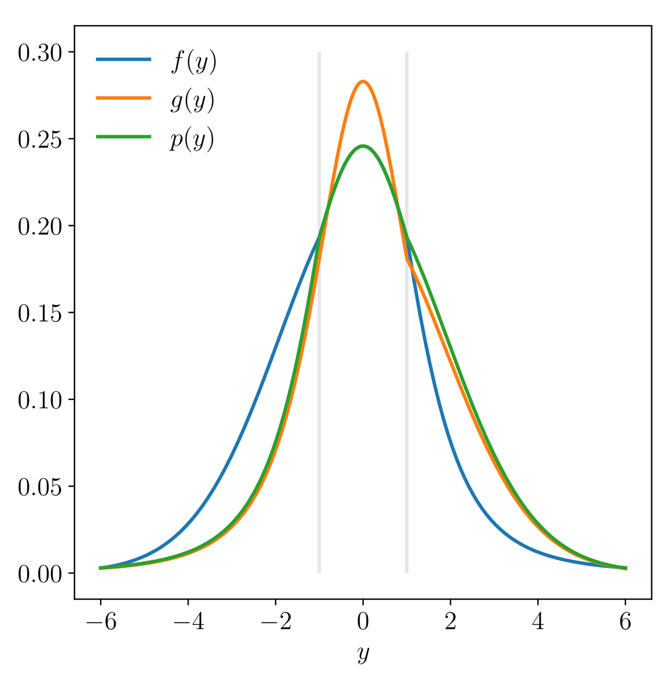

Figure 2. Localisation bias

The Censored Scoring Rule

Consider the **censored density** for an indicator weight function $w(y) = \mathbf{1}_A(y)$

$$f_w^\flat(y) = \begin{cases} f(y), & \text{if } y \in A, \\ \bar{F}_w, & \text{if } y \in A^c. \end{cases}$$

where $\bar{F}_w = \int_{A^c} f \mathrm{d}\mu$. The **censored scoring rule** applies the original scoring rule to this censored density

$$S_w^\flat(f, y) = \begin{cases} S(f_w^\flat, y), & \text{if } y \in A, \\ S(f_w^\flat, *), & \text{if } y \in A^c. \end{cases}$$

Observations outside A are censored, i.e. made indistinguishable, preserving the probability $\bar{F}_w$. By keeping this probability, censoring retains more information about the original distribution than conditioning.
More generally, for distributions $\mathbb{F}$ living on the measurable space $(\mathcal{Y}, \mathcal{G})$, like $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, the censored scoring rule reads

$$S_w^\flat(\mathbb{F}, y) = w(y)S(\mathbb{F}_w^\flat, y) + \big(1 - w(y)\big)S(\mathbb{F}_w^\flat, *), \quad \mathrm{d}\mathbb{F}_w^\flat = \mathrm{d}\mathbb{F}_w + \bar{F}_w \mathrm{d}\delta_*.$$

Theoretical Results

1. Assume that the regular scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is strictly proper and expectations relative to the censored versions of measures in $\mathcal{P}$ are also finite. Then $S_w^\flat$ is strictly locally proper.
2. $S_w^\flat$ remains strictly locally proper for other choices than $\mathrm{d}\mathbb{H} = \mathrm{d}\delta_*$, as long as $(w, \mathbb{H})$ are such that $\exists E \in \mathcal{G} : \mathbb{F}_w(E) = 0$ and $\mathbb{H}(E) > 0$, $\forall \mathbb{F} \in \mathcal{P}, \mathbb{H} \in \mathcal{H} \subseteq \mathcal{P}$, which is a regularity condition in practice.
3. Consider testing the (multiple versus multiple) hypothesis

$$H_0 : p_{0tA_t} = f_{0tA_t},\ \forall t \qquad \text{vs} \qquad H_1 : p_{1tA_t} = f_{1tA_t}, \quad \forall t,$$

where f_{0t} and f_{1t} are hypothesised conditional densities for the actual densities p_{0t} of a stochastic process $\{Y_t : \Omega \to \mathcal{Y}\}_{t=1}^T$. The Uniformly Most Powerful test is to reject H_0 if the censored likelihood ratio $\lambda(\mathbf{y}) = \frac{[f_1]_A^\flat(\mathbf{y})}{[f_0]_A^\flat(\mathbf{y})}$ is large enough.
Note: This Theorem generalises the famous Neyman Pearson Lemma.

Monte Carlo

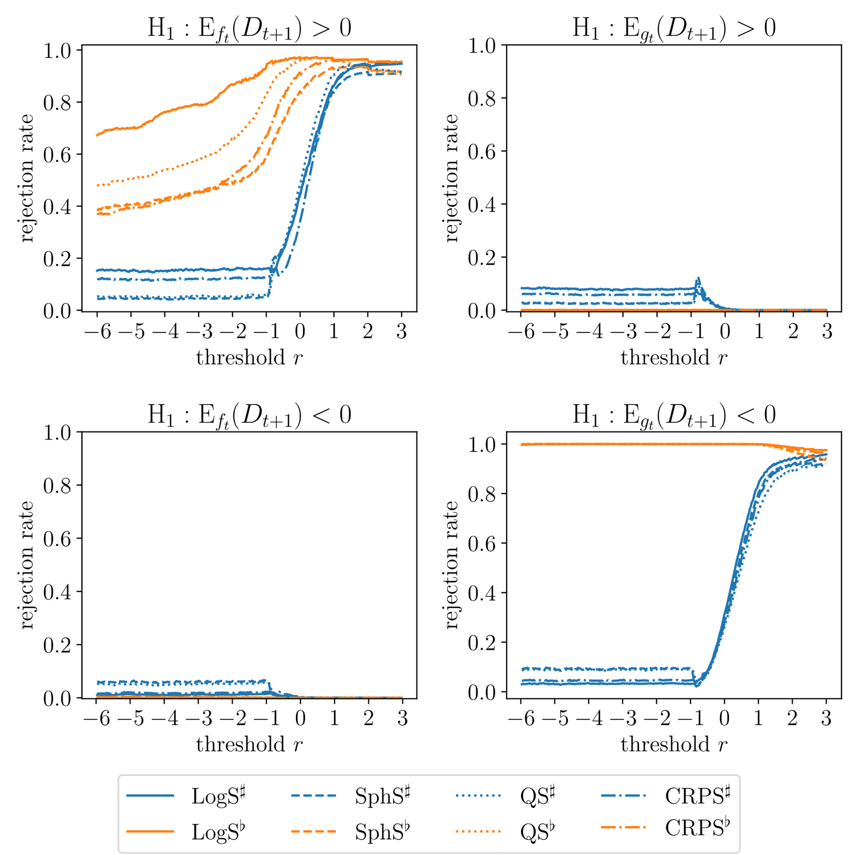

Figure 3. Laplace$(-1, 1)$ vs Laplace$(1, 1.1)$

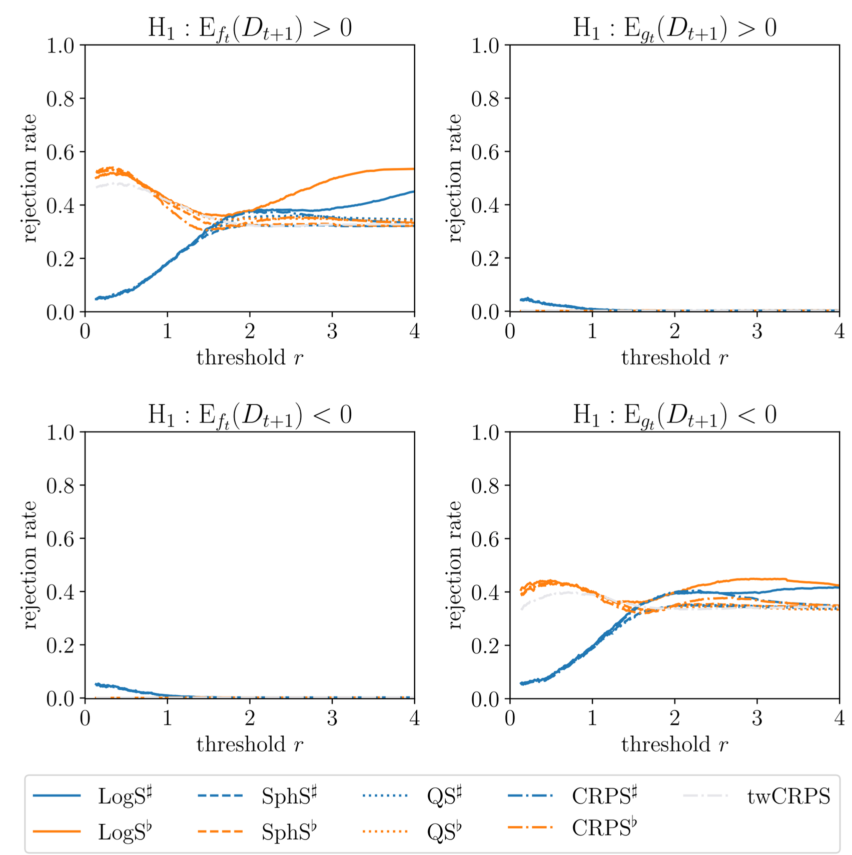

Figure 4. Centre: $\mathcal{N}(0, 1)$ vs $t(5)$

One-sided rejection rates of the GW-test of equal predictive ability of the candidates f_t and g_t at a nominal significance level of 0.05 based on 10,000 simulations. The DGP is either f_t (left-hand side) or g_t (right-hand side). Rejections in the top panels are in favour of f_t, while rejections in the bottom panels are in favour of g_t. The incorporated weight function is $w(y) = \mathbf{1}_{y<r}$ in Figure 3 and $w(y) = \mathbf{1}_{-r\le y\le r}$ in Figure 4. The number of expected observations in the region of interest is kept constant at $c = 20$ in Figure 3 and $c = 200$ in Figure 4. The implemented scoring rules are the Logarithmic, Spherical, Quadratic scoring rules and the Continuously Ranked Probability Score.

Empirical Application

	LogS		QS		SphS		CRPS	
	$\flat$	$\sharp$	$\flat$	$\sharp$	$\flat$	$\sharp$	$\flat$	$\sharp$
RGARCH-t	$\underline{*}$	$\underline{*}$	$\underline{*}$	$\underline{*}$	$\underline{*}$	$\underline{*}$	$*$	$\underline{*}$
TGARCH-t		$\underline{*}$	$*$	$\underline{*}$		$\underline{*}$		$\underline{*}$
GARCH-t		$\underline{*}$		$\underline{*}$		$\underline{*}$		$*$
RGARCH-$\mathcal{N}$			$\underline{*}$	$\underline{*}$	$\underline{*}$	$*$	$\underline{*}$	$\underline{*}$
TGARCH-$\mathcal{N}$				$\underline{*}$			$*$	
GARCH-$\mathcal{N}$								

$\text{MCS}_{0.75}$ ($\underline{*}$) and $\text{MCS}_{0.90}$ ($*$ and $\underline{*}$) based on censored ($\flat$) and conditional ($\sharp$) scoring rules. The weight function is the left-tail indicator function based on a rolling empirical quantile at level $q = 0.10$. As data we use the log-returns of the S&P500, that is, $y_t = \log(P_t/P_{t-1})$, where P_t is the adjusted closing price on day t. This time series consists of 6,777 observations in total, spanning from January 2, 1996, to December 30, 2022, and is obtained from Yahoo Finance. The realised measure is downloaded from the Risklab page of Dacheng Xiu's website.

References

[1] C Diks, V Panchenko, and D van Dijk. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230, 2011.

[2] R Giacomini and H White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, 2006.

[3] T Gneiting and AE Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[4] T Gneiting and R Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011.

[5] P R Hansen, A Lunde, and J M Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

[6] H Holzmann and B Klar. Focusing on regions of interest in forecast evaluation. *Annals of Applied Statistics*, 11(4):2404–2431, 2017.

✉ R.F.A.depunder@uva.nl