Localising Strictly Proper Scoring Rules*

Ramon de Punder† Cees Diks† Roger Laeven† Dick van Dijk‡

June 23, 2023

Abstract

Forecasters are typically not equally interested in all possible realisations of a random variable under scrutiny. Financial risk managers, for instance, usually put relatively more weight on regions of extreme losses. In density forecast comparison, it is common practice to use strictly proper scoring rules to rank a collection of candidate predictive distributions. When focusing on a region of interest, however, weighted scoring rules obtained via conditioning are no longer strictly proper. We develop a general procedure for focusing, i.e., localising, scoring rules in a way that preserves their strict propriety. Our procedure provides a myriad of strictly locally proper scoring rules beyond the censored likelihood score. In particular, the focusing procedure we develop is general enough to handle both univariate and multivariate scoring rules, including the rich class of kernel scores. The one-to-one correspondence between the censored distribution and the original distribution on the region of interest preserves not only strict propriety but also the optimal power properties of the Logarithmic scoring rule. More specifically, our paper generalises the Neyman Pearson lemma, showing that the uniformly most powerful test for a localised version of this lemma's original hypotheses boils down to a censored likelihood ratio test. Based on a collection of popular scoring rules, including the Logarithmic, Spherical, Quadratic and Continuously Ranked Probability Score (CRPS), Monte Carlo simulations and the results of our empirical illustration align with the intuition that censoring bears, also in general, more desirable power properties than conditioning, especially if the number of expected tail observations is small.

*This working paper is a draft. A previous version titled 'A General Procedure for Localising Strictly Proper Scoring Rules' was first published on July 10, 2022.

†University of Amsterdam, Tinbergen Institute

‡Erasmus University Rotterdam, Tinbergen Institute

Contents

1 Introduction

Any forecasting application necessitates quantifying the relative performance of different forecasting methods. Gneiting and Raftery (2007) motivated the use of strictly proper scoring rules for this job, which has become the industry standard (Brehmer and Gneiting, 2020; Patton, 2020). The reason for this is that (strictly) proper scoring rules assign a score to the actual distribution that is (strictly) larger than the score of any other predictive distribution. Although strictly proper scoring rules admit point forecasts (e.g. mean squared error), we concentrate on their use in combination with predictive distributions and densities. Forecasts in the form of predictive distributions have gained interest in many different forecasting fields because they give a complete picture of the stochastic nature of the variable of interest (Dawid, 1984). At the same time, the specific characteristics of such applications encourage us to zoom in on certain parts of this picture, i.e. to localise the original scoring rule. In this paper, we present a general censoring-based procedure for localising scoring rules that preserves strict propriety. Our framework nests the censored likelihood (csl) scoring rule proposed by Diks et al. (2011) as a special case. We show that the uniformly most powerful test for a localised hypothesis test is based on this strictly locally proper scoring rule.

Motivating examples for local scoring rules can be found in different application areas. In risk management, for example, one is particularly interested in the left tail of the loss distribution, largely driven by regulatory capital requirements, formulated in terms of risk measures such as the Value-at-Risk (VaR) and Expected Shortfall (ES). See e.g. Diks et al. (2014), Kole et al. (2017), Opschoor et al. (2017) and Diks and Fang (2020) for applications. In macroeconomics, policymakers set – whether regulated by law or not – targets for central variables like inflation, nominal GDP and unemployment rates. For such clear targets, it is logical to zoom in on the part of the distribution around the target value. We refer to Gneiting and Ranjan (2011) and Iacopini et al. (2022) (and references therein) for interesting examples in macroeconomics.

The literature on focused scoring rules starts with the weighted likelihood score of

Amisano and Giacomini (2007), which simply multiplies the unweighted logarithmic scoring rule by a weight function. As independently observed by Diks et al. (2011) and Gneiting and Ranjan (2011), this procedure produces improper scoring rules because it favours distributions with more mass assigned to regions with higher weights, independent of the underlying distribution. As proper alternatives, Gneiting and Ranjan (2011) develop the weighted continuously ranked probability scoring (twCRPS) rule, while Diks et al. (2011) propose the conditional (cl) and csl scoring rule. Holzmann and Klar (2017, Theorem 1) observe that the procedure of the cl scoring rule can be generalised to other scoring rules than the logarithmic scoring rule. They propose a general procedure for focusing regular scoring rules that applies the regular scoring rule to a weighted transformation of the original distribution. Their approach differs from ours by the suggested transformation of the original distribution: a conditional vis-à-vis censored distribution. The impact of this difference is that our censoring-based mechanism is the only one guaranteed to deliver strictly locally proper scoring rules. Interestingly, another route leading to the conditioning mechanism of Holzmann and Klar (2017, Theorem 1) is to first generalise the weighted log-likelihood scoring rule proposed by Amisano and Giacomini (2007) and then apply a transformation coined *properisation* by Brehmer and Gneiting (2020, Theorem 1).

Our research also builds on the existing work on strictly proper scoring rules and their associated divergence measures. Although Gneiting and Raftery (2007) are responsible for the formal definition of strict propriety, scoring rules satisfying this property date back to at least the quadratic scoring (QS) rule proposed by Brier (1950). It is useful to know that this research area is dichotomous in the sense that much of the research prior to the rigorous treatment of general probability measures by Gneiting and Raftery (2007) has been conducted relative to discrete distributions on a finite outcome space, while more recent work more often follows the generality of Gneiting and Raftery (2007). For instance, the introduction of the LogS (Good, 1952; Toda, 1963) and spherical scoring (SphS) rule (Roby, 1964; Good, 1971), the initial generalisations of QS and SphS to the PowS_α and PsSphS_α families, and the axiomatic

characterisations of the LogS, PowS_α and PsSphS_α rules provided by Shuford et al. (1966), Savage (1971), Selten (1998) and Jose (2009), are all presented in a discrete context. In our analysis, we work with the generalisations of the PowS_α and PsSphS_α families advocated by Gneiting and Raftery (2007) and Ovcharov (2018).

Moreover, the expected score differences of many scoring rules are recognised as well-known divergence measures, which reduce all together to the class of Bregman divergences (Bregman, 1967) when solely considering strictly proper scoring rules (Dawid, 2007; Gneiting and Raftery, 2007; Ovcharov, 2018; Painsky and Wornell, 2019). Consequently, concentrating the score divergences of strictly proper scoring rules excludes all f-divergences except the Kullback Leibler divergence (Kullback and Leibler, 1951), which is the unique intersection of the Bregman and f-divergence families. Due to its favourable properties (Liese and Vajda, 2006) the Kullback Leibler divergence has become the cornerstone in measuring the discrepancy between densities. For example, it is the divergence that is minimised in the maximum likelihood framework (Fisher, 1922), which bears optimal properties in the context of testing and estimation. Specifically, the likelihood ratio test is the most powerful test (Neyman and Pearson, 1933) and maximum likelihood estimators are unbiased estimators reaching the Cramér–Rao lower bound.

Pivotised sample equivalents of the expected score differences are fundamental in hypothesis tests about the relative performance of two candidate predictive distributions. In line with the weighted applications we have in mind, we localise the simple versus simple hypothesis of the Neyman-Pearson lemma into statements about the underlying distribution on the region of interest. By doing so, the hypothesis about the underlying distribution becomes a multiple versus multiple hypothesis, equivalent to the hypothesis studied by Holzmann and Klar (2016). Unlike them, we are still able to derive the uniformly most powerful test for this hypothesis. The test statistic of this test is given by a localised likelihood ratio, where the localisation is performed by censoring, and necessarily not by conditioning.

Power analyses based on localised scoring rules have more frequently been studied

for the Giacomini and White (2006) test (Diks et al., 2011, 2014; Holzmann and Klar, 2016; Lerch et al., 2017). The null hypothesis of this test entails that the expected score difference between one candidate to the actual distribution is equivalent to the expected score difference between the other candidate and the actual distribution. A great advantage of this test is that all choices underlying the predictor, such as parameter uncertainty, can be seen as an integral part of the candidate, therefore also called prediction methods. For a strictly proper scoring rule, the null implies that both candidates are necessarily misspecified under the null, namely 'equally misspecified'. Yet, since which distributions are equally off from both candidates is determined by the scoring rule, this means that the null set of the GW test is a function of the selected candidates and the selected scoring rule, complicating comparisons between GW tests based on different scoring rules. To illustrate this interplay, we include a parametric example for which the conditional GW null set coincides with the full parameter space, whereas the censored GW null is a lower-dimensional subspace of the parameter space. We also compare the power properties of the GW test of the censored scoring rules with their conditional counterparts and other commonly used localised scoring rules like the twCRPS of Gneiting and Raftery (2007). In line with Diks et al. (2011), we find that censoring often leads to higher power.

The remainder of this paper is organised as follows. Section 2 describes the fundamental concepts on which the subsequent chapters rely. Section 3 defines the generalised censored scoring rule and includes the assumption under which it is shown to be strictly locally proper. This section also includes a rich collection of examples and a randomisation procedure, called Z-Q-randomisation, equivalent to the generalised censored scoring rule. It closes with our generalisation of the Neyman-Pearson lemma. Section 4 compares the size and power properties of a test of equal predictive ability of conditional and censored scoring rules. Section 6 concludes.

2 Scoring rules

2.1 Regular scoring rules

Consider a random variable $Y : \Omega \to \mathcal{Y}$ from a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the measurable space $(\mathcal{Y}, \mathcal{G})$. The goal of a forecaster is to choose a distribution F from a convex class of candidate distributions $\mathscr{P}$ on $(\mathcal{Y}, \mathcal{G})$ that minimises the *score divergence*

$$\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) := \mathbb{H}_S(\mathrm{P}) - \mathbb{E}_\mathrm{P} S(\mathrm{F}, \cdot)$$

over $\mathscr{P}$, where $\mathbb{H}_S(\mathrm{P}) := \mathbb{E}_\mathrm{P} S(\mathrm{P}, \cdot)$ is the *negative entropy* of P based on S. Adhering to Gneiting and Raftery (2007), the selected *scoring rule* S is restricted to be *strictly proper* to ensure that the forecaster truthfully reports the actual distribution P as the best candidate from $\mathscr{P}$, if $\mathrm{P} \in \mathscr{P}$. Definitions 1 and 2, adopted from Holzmann and Klar (2017) and Gneiting and Raftery (2007), respectively, formalise both concepts.

Definition 1 (Scoring rule)**.** *A scoring rule is any extended real-valued* ($\bar{\mathbb{R}} := [-\infty, \infty]$) *function* $S : \mathscr{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ *such that* $S(\mathrm{F}, \cdot)$ *is measurable with respect to* $\mathcal{G}$ *and quasi-integrable with respect to all* $\mathrm{P} \in \mathscr{P}$*, for all* $\mathrm{F} \in \mathscr{P}$*, and for which* $\mathbb{E}_\mathrm{P} S(\mathrm{F}, \cdot) < \infty$ *and* $\mathbb{H}_S(\mathrm{P}) \in \mathbb{R}, \forall \mathrm{P}, \mathrm{F} \in \mathscr{P}$*.*

Definition 2 ((Strictly) proper scoring rule)**.** *A scoring rule* $S : \mathscr{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ *is proper relative to* $\mathscr{P}$ *if* $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) \geq 0,\ \forall \mathrm{P}, \mathrm{F} \in \mathscr{P}$*, and strictly proper if, additionally,* $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) = 0$ *iff* $\mathrm{P} = \mathrm{F},\ \forall \mathrm{P}, \mathrm{F} \in \mathscr{P}$*.*

If a scoring rule only uses the μ-densities $f \in \mathscr{p}$ of the candidates $\mathrm{F} \in \mathscr{P}$, it is easier to work with the densities directly, i.e. to define $S : \mathscr{p} \times \mathcal{Y} \to \bar{\mathbb{R}}$ and adapt all definitions in this section accordingly. Yet, this is only possible if there exists a σ-finite measure μ on $(\mathcal{Y}, \mathcal{G})$ such that $\mathrm{F} \ll \mu, \forall \mathrm{F} \in \mathscr{P}$. Furthermore, the restrictions on S in Definition 1 guarantee a meaningful comparison of the expected score of any candidate with the negative entropy of the actual distribution, necessary for identifying (strict) propriety. Since comparisons of candidates $\mathrm{F} \in \mathscr{P}$ are in terms of P-expectations, the forecaster is,

strictly speaking, only forced to report a member from the P-a.s. equivalence of P when using a strictly proper scoring rule. For clarity, we henceforth suppress technicalities about P-a.s. equivalence. Definition 2 additionally shows that a score divergence is a *divergence measure* (see e.g. Eguchi et al. (1985)) if and only if S strictly proper. For distributions on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$, where $\mathcal{B}(\mathcal{Y})$ denotes the Borel σ-algebra on $\mathcal{Y}$, the particular form of $\mathbb{D}_S(\mathrm{P}\|\mathrm{F})$ makes it a Bregman divergence (Bregman, 1967) under the conditions listed by Ovcharov (2018).

In their review paper, Gneiting and Raftery (2007) provide an abundant list of strictly proper scoring rules, which can be divided into two categories: *local* scoring rules and *distance sensitive* scoring rules (Ehm and Gneiting, 2012). We use the same structure when discussing examples, yet allowing local scoring rules, henceforth called *semi-local*, to also depend on the density via a global norm of the density. Within the class of semi-local rules, we focus on the Logarithmic (LogS) (Good, 1952; Toda, 1963), Quadratic (QS) (Brier, 1950) and (SphS) (Roby, 1964; Good, 1971) scoring rules as well as their generalisations to the Power (PowS_α) and PseudoSpherical (PsSphS_α) families. Our selection of distance-sensitive scoring rules fits into the family of Energy Scores (ES), a subclass of the class of strictly proper scoring rules given by Theorem 5 of Gneiting and Raftery (2007), nesting the real-valued Continuously Ranked Probability Score (CRPS) (Matheson and Winkler, 1976; Hersbach, 2000) as a special case.

The inclusion of the PowS_α and PsSphS_α families, sharing LogS as a common limiting case for $\alpha \downarrow 1$, is partly due to the connection with the expected utility maximisation problems described by Jose et al. (2008). After all, the duality with specific investment problems based on the one-parameter Hyperbolic Absolute Risk Aversion (HARA) utility function family, generated by the absolute risk tolerance function $\tau_\alpha(x) = \beta + \alpha x$, with $\beta = 1$ (Merton, 1971, p. 389), gives α its interpretation as a risk tolerance parameter. Their introduction and axiomatic characterisation are found by Shuford et al. (1966), Savage (1971), Selten (1998) and Jose (2009), though we work with their continuous generalisations provided by Gneiting and Raftery (2007) and Ovcharov (2018).

2.2 Weighted scoring rules

In many applications, particular outcomes are of particular importance. To emphasise regions of the outcome space, a forecaster with scoring rule S is assumed to select a *weight function* $w \in \mathcal{W}$, that is, any $\mathcal{G}$-measurable map $w : \mathcal{Y} \to [0, 1]$. The forecaster's weight function is zero for outcomes that are of zero interest. Hence, differences in candidates expressed only on $\{w = 0\} := \{y \in \mathcal{Y} : w(y) = 0\}$ are ideally not accounted for by the scoring rule. Therefore, we restrict the analysis to the class of *localising weighted scoring rules* given by Definition 3, borrowed from Holzmann and Klar (2017).

Definition 3 (Localising weighted scoring rule)**.** *A weighted scoring rule S, that is, a map $S : \mathscr{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ such that $S_w(\cdot, \cdot)$ is a scoring rule for each $w \in \mathcal{W}$, is localising if for any* P, $\mathrm{F} \in \mathscr{P}$, $w \in \mathcal{W}$, *it holds that*

$$\forall E \in \mathcal{G} : \mathrm{P}(\{w > 0\} \cap E) = \mathrm{F}(\{w > 0\} \cap E) \implies S_w(\mathrm{P}, y) = S_w(\mathrm{F}, y), \ \forall y \in \mathcal{Y}.$$

Considering the indicator weight function $w(y) = \mathbb{1}_A(y)$, taking the value 1 if $y \in A$, and 0 otherwise, it is obvious that a localising so-weighted scoring rule cannot be strictly proper. Indeed, any distribution $\tilde{\mathrm{P}}$ equivalent to P on A is assigned the same score. Therefore, we instead aim for *strictly locally proper* weighted scoring rules, initially defined by Holzmann and Klar (2017) and included below as Definition 4.

Definition 4 ((Strictly) locally proper scoring rule)**.** *A weighted scoring rule $S : \mathscr{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ is locally proper relative to $(\mathscr{P}, \mathcal{W})$ if it is localising and $S_w(\cdot, \cdot)$ is proper for each $w \in \mathcal{W}$. Furthermore, it is strictly locally proper relative to$(\mathscr{P}, \mathcal{W})$ if, additionally,*

$$\mathrm{P}(\{w > 0\} \cap E) = \mathrm{F}(\{w > 0\} \cap E), \forall E \in \mathcal{G} \iff \mathbb{E}_{\mathrm{P}} S_w(\mathrm{P}, \cdot) = \mathbb{E}_{\mathrm{P}} S_w(\mathrm{F}, \cdot), \ \forall w \in \mathcal{W}.$$

Before turning to our solution to weighting scoring rules, we first recall two weighting procedures contained in the literature. First, the recipe $S_w(\mathrm{F}, y; w) = w(y)S(\mathrm{F}, y)$ proposed by Amisano and Giacomini (2007) for the Logarithmic scoring rule is clearly not strictly locally proper. Indeed, as shown by Example 1 of Diks et al. (2011), it does

not rule out scoring rules that are completely determined by a region where one candidate density dominates the other, yielding a higher expected score for the dominating one, irrespective of the actual distribution.

A second recipe failing to deliver strictly locally proper scoring rules is the *conditional scoring rule*

$$S_w^\sharp(\mathrm{F}, y) := w(y) S(\mathrm{F}_w^\sharp, y), \qquad \mathrm{dF}_w^\sharp := \frac{1}{\mathrm{F}_w(\mathcal{Y})} \mathrm{dF}_w,$$

proposed by Holzmann and Klar (2017). This rule applies the regular scoring rule to the *conditional distribution* $\mathrm{F}_w^\sharp$, here defined as the *weighted kernel* $\mathrm{dF}_w := w\mathrm{dF}$ scaled (sharpened) by a factor $1/\mathrm{F}_w(\mathcal{Y})$. Again, this scoring rule completely ignores outcomes in $\{w = 0\}$. Though now, the distribution is adjusted accordingly, making the scoring rule locally proper. However, since it cannot discriminate between distributions that are proportional to each other on $\{w > 0\}$, it is not strictly locally proper (Holzmann and Klar, 2017). Figure 1a illustrates the potential consequences of this lack of discriminative ability.

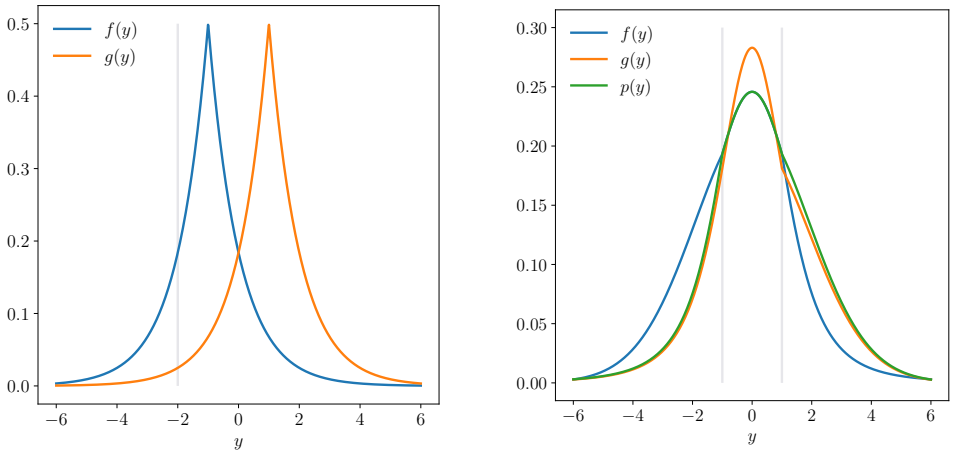
A hint from these examples is that we should not completely forget about $\{w = 0\}$ when focusing on $\{w > 0\}$. Yet, to stay localising, we can only use information about a candidate's distribution on $\{w = 0\}$ that is implied by the information on $\{w > 0\}$. A clear example of a non-localising weighted scoring rule is the twCRPS,

$$\mathrm{twCRPS}(F, y) = \int_{-\infty}^{\infty} w(s) \Big(F(s) - \Delta_y(s) \Big)^2 \mathrm{d}s,$$

for weight functions of the kind $w(y) = \mathbb{1}_{[r_1, r_2]}(y)$, where $r_1 < r_2$ and $r_1, r_2 \in \mathbb{R}$. The piecewise Student-$t(\nu_1, \nu_2, \nu_3)$ example displayed in Figure 1b shows that the cumulative character of the CDF destroys the localisation to $A = [r_1, r_2] = [-1, 1]$. Here in particular, we selected a distracting candidate G that is similar to P outside A in the sense that $\nu_{1G} = \nu_{1P} = 3$ and $\nu_{3G} = \nu_{3P} = 40$, but different on A, with $3 = \nu_{2G} < \nu_{2P} = 5$. In contrast, $\nu_{2F} = \nu_{2P}$, while $\nu_{1F} = \nu_{3P}$ and $\nu_{1F} = \nu_{3P}$. The fact that $\mathbb{D}_{\mathrm{twCRPS}}(\mathrm{P} \| \mathrm{F}) > \mathbb{D}_{\mathrm{twCRPS}}(\mathrm{P} \| \mathrm{G})$, while F and P coincide on A, re-

veals that the twCRPS is distracted by the good fit of G outside A. Since the bias $\mathbb{D}_{\text{twCRPS}}(\text{P}\|\text{F}) - \mathbb{D}_{\text{twCRPS}}(\text{P}\|\text{G}) \approx 0.028$ is the sole consequence of the weighted scoring rule being non-localising, it is henceforth referred to as a *localisation bias*.

Figure 1: Non-strictly locally proper scoring rules



(a) Laplace (b) Piecewise Student-t

Note: (a) The densities f and g are both from the Laplace family with common scale parameter $\theta = 1$, but different location parameter $\mu_f = -1$ and $\mu_g = 1$. Since Laplace tails are known to be proportional for members of equivalent scale, it follows that $S_w^\sharp(f, y) = S_w^\sharp(g, y)$ on $A = (-\infty, -2)$, while $f \neq g$ on A. Therefore, $S_w^\sharp$ cannot be strictly locally proper, e.g. consider $p = f$. (b) f, g and p are all piece-wise Student-t, constructed such that $f = p \neq g$ on $A = [-1, 1]$. More specifically, the density $f(y; \boldsymbol{\nu}_F)$ is the normalisation of the kernel $\tilde{f}(y; \boldsymbol{\nu}_F) = q(y; \nu_{1F}) \frac{q(-1;\nu_{2F})}{q(-1;\nu_{1F})} \mathbb{1}_{y<-1} + q(y; \nu_{2F}) \mathbb{1}_{-1 \leq y \leq 1} + q(y; \nu_{3F}) \frac{q(1;\nu_{2F})}{q(1;\nu_{3F})} \mathbb{1}_{y>1}$. As a result of its non-locality, the twCRPS implies a score divergence indicating g to be statistically closer to p on A than f. Since $p = f \neq g$ on A, the twCRPS is therefore not strictly locally proper.

3 Localising scoring rules by censoring

3.1 Censored scoring rule

To overcome issues like the non-strictness and non-locality of the weighted scoring rules discussed above, we propose to use censoring as focusing mechanism. Censoring (Bernoulli, 1760) is a statistical concept that is used in econometrics to model a dependent variable whose value is only partially known (Tobin, 1958). More specifically, for realisations in A^c, the complement of A, it is only known that they are not in A. Events in A^c are hence indistinguishable after censoring and 'A^c' could therefore be

viewed as a single outcome of the censored random variable. To avoid confusion, we label observations in A^c by '$*$' rather than 'A^c' itself, which is nothing but an abstract event for which one can alternatively read 'NaN'. The censored random variable

$$Y_A^\flat = \begin{cases} Y, & Y \in A, \\ *, & Y \in A^c, \end{cases}$$

is defined relative to the extended measurable space $(\mathcal{Y}^*, \mathcal{G}^*)$, where $\mathcal{Y}^* = \mathcal{Y} \cup \{*\}$ and $\mathcal{G}^* = \sigma(\{\mathcal{G}, *\})$, that is, the smallest σ-algebra containing the collection $\{\mathcal{G}, *\}$. Similar to the conditional distribution, we extend the definition of the distribution of $Y_A^\flat$ to general weight functions $w \in \mathcal{W}$. In particular, we define the *censored distribution* as

$$\mathrm{dF}_w^\flat = \mathrm{dF}_w + \bar{\mathrm{F}}_w \mathrm{d}\delta_*, \qquad \bar{F}_w := \int_{\mathcal{Y}} (1-w)\mathrm{dF}, \qquad w \in \mathcal{W},\ \mathrm{F} \in \mathscr{P}, \tag{1}$$

where δ_* denotes the Dirac measure at $*$, i.e. $\delta_*(E) = \mathbb{1}_E(*)$. To make this change of measure well-defined, we consider the original measures $\mathrm{F} \in \mathscr{P}$ relative to the extended measurable space $(\mathcal{Y}^*, \mathcal{G}^*)$, by defining $\mathrm{F}(*) = 0$ and taking some value for $w(*)$. In case $\mathrm{F} \ll \mu, \forall \mathrm{F} \in \mathscr{P}$, we are invited to work with the μ-densities $f \in \mathscr{p}$ instead, and their associated $(\mu + \delta_*)$-densities

$$f_w^\flat = w f \mathbb{1}_{y \neq *} + \bar{F}_w \mathbb{1}_{y=*}, \qquad w \in \mathcal{W},\ f \in \mathscr{p}. \tag{2}$$

A detailed proof of this result is deferred to the Online Appendix. Albeit restricted to $w(y) = \mathbb{1}_A(y)$, Borowska et al. (2020) also work with an explicit formulation of the censored density, coinciding with $f_A^\flat$ in the context of maximum likelihood. Here $f_A^\flat$ is preferred notation for $f_{\mathbb{1}_A}^\flat$.

Ideally, the censored scoring rule would be given by $S_A^\flat(\mathrm{F}, y) = S(\mathrm{F}_A^\flat, y_A^\flat)$, as this would fully respect the forecaster's specific choice of the regular scoring rule S. The censored scoring rule given by Definition 5 reduces to this definition for the indicator weight function $w(y) = \mathbb{1}_A(y)$. The censored scoring rule is also attractive for general

weight functions, but this will be particularly clear from the randomisation perspective taken in Section 3.3, which yields a similar identity for general weight functions; see Equation (4). According to Theorem 1, the censored scoring rule is strictly locally proper. Since Theorem 1 is a corollary of Theorem 2, we have sustainably omitted a proof for this result.

Definition 5 (Censored scoring rule)**.** *Let* $S : \mathscr{P}^\flat \times \mathcal{Y} \to \bar{\mathbb{R}}$, $\mathscr{P}^\flat = \{\mathrm{F}_w^\flat, \mathrm{F} \in \mathscr{P}, w \in \mathcal{W}\}$, *denote a scoring rule. Then, the corresponding censored scoring rule is given by the map* $S^\flat : \mathscr{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$,

$$S_w^\flat(\mathrm{F}, y) := w(y)S(\mathrm{F}_w^\flat, y) + \big(1 - w(y)\big)S(\mathrm{F}_w^\flat, *),$$

where the censored distribution $\mathrm{F}_w^\flat$ *is defined in Equation* (1)*.*

Theorem 1. *Suppose that the regular scoring rule* S *is strictly proper relative to* $\mathscr{P}^\flat$*. Then, the censored scoring rule* $S^\flat$ *in Definition 5 is strictly locally proper relative to* $(\mathscr{P}, \mathcal{W})$*.*

The assumption in Theorem 1 ensures that the scoring rule is well-defined with respect to mixed continuous-discrete distributions on $(\mathcal{Y}^*, \mathcal{G}^*)$. We will verify that this assumption holds in the examples discussed in Subsection 3.4.

Let us conclude this section by providing some intuition for the result of Theorem 1. Given some weight function $w \in \mathcal{W}$, it should be clear that censoring maintains a one-to-one connection with the original distribution on $\{w > 0\}$. This relation can be harmed by conditioning due to the additional normalisation of the weighted kernel. This difference is even clearer for indicator weight functions since $\mathrm{F}_A^\flat = \mathrm{F}$, while $\mathrm{F}_A^\sharp \neq \mathrm{F}$, on A. Because of this, only the censored scoring rule allows for identifying the original distributions on $\{w > 0\}$ when comparing two candidates F and G. This additionally requires disentanglabity of the weighted kernels and discrete probabilities in the censored measures, implied by $\mathrm{F}_w(*) = \mathrm{G}_w(*) = 0$. Consequently, the assumed strict propriety of the original rule localises to $\{w > 0\}$ for the censored scoring rule.

3.2 Generalised censored scoring rule

Given the intuition at the end of the previous section, it is not entirely surprising that one can perform other transformations to the distribution on $\{w > 0\}$ as long as the transformation is independent of the distribution and traceable when comparing two candidate distributions. The latter requirement is formalised by Assumption 1, under which the *generalised censored scoring rule* in Definition 6 is still strictly locally proper. Appendix A.1 details a proof for this result, summarised by Theorem 2.

Definition 6 (Generalised censored scoring rule)**.** *Let* $S : \mathscr{P}^\flat \times \mathcal{Y} \to \bar{\mathbb{R}}$ *denote a scoring rule. The associated generalised censored scoring rule is given by the map* $S^\flat : \mathscr{P} \times \mathcal{Y} \times \mathcal{W} \times \mathscr{H} \to \bar{\mathbb{R}}$*,*

$$S^\flat_{w,\mathrm{H}}(\mathrm{F}, y) = w(y)S(\mathrm{F}^\flat_{w,\mathrm{H}}, y) + \big(1 - w(y)\big)\mathbb{E}_{\mathrm{H}} S(\mathrm{F}^\flat_{w,\mathrm{H}}, \cdot), \quad \mathrm{dF}^\flat_{w,\mathrm{H}} = \mathrm{dF}_w + \bar{F}_w \mathrm{dH},$$

where $\mathrm{F}^\flat_{w,\mathrm{H}}$ *is referred to as the generalised censored distribution of* F*.*

Assumption 1. *A weight function* $w \in \mathcal{W}$ *and nuisance distribution* $\mathrm{H} \in \mathscr{H} \subseteq \mathscr{P}$ *is such that* $\exists E \in \mathcal{G} : \mathrm{F}_w(E) = 0$ *and* $\mathrm{H}(E) > 0,\ \forall \mathrm{F} \in \mathscr{P}, \mathrm{H} \in \mathscr{H}$*.*

Theorem 2. *Suppose that (i) the regular scoring rule* S *in Definition 6 is strictly proper relative to* $\mathscr{P}$*, and (ii)* $\mathcal{W}$ *and* $\mathscr{H}$ *are such that Assumption 1 is satisfied. Then, the generalised censored scoring rule* $S^\flat$ *in Definition 6 is strictly locally proper relative to* $(\mathscr{P}, \mathcal{W}, \mathscr{H})$*.*

Finally, a corollary of Lemma 3 in the proof of Theorem 2 in Appendix A.1 is that

$$\mathbb{D}_{S^\flat_{w,\mathrm{H}}}(\mathrm{F} \| \mathrm{G}) = \mathbb{D}_S(\mathrm{F}^\flat_{w,\mathrm{H}} \| \mathrm{G}^\flat_{w,\mathrm{H}}), \tag{3}$$

i.e. the censored score divergence from F to G is the score divergence of the corresponding censored distributions. In particular, this means that we have identified a family of so-called *localised divergence measures*, satisfying the properties of a divergence measure (see Subsection 2.1) on $\{w > 0\}$. Indeed, if S is strictly proper, such that $\mathbb{D}_S$ is a

divergence measure, it follows that $\mathbb{D}_{S^\flat_{w,\mathrm{H}}}(\mathrm{F}\|\mathrm{G}) \geq 0$, with strict equality if and only if $\mathrm{F}(E \cap \{w > 0\}) = \mathrm{G}(E \cap \{w > 0\})$, $\forall E \in \mathcal{G}$.

3.3 Z, Q-Randomisation

The (generalised) censored scoring rule in Definition 5 (6) of the previous section can alternatively be formulated in terms of a randomisation procedure. This procedure relies on an auxiliary random variable Z_w, indicating, conditional on the realisation y, whether the observation is censored or not. More specifically, we let

$$y^\flat_{Z_w} = \varphi(y, Z_w), \qquad \varphi(y, Z_w) := \begin{cases} y, & Z_w = 1, \\ *, & Z_w = 0, \end{cases}$$

where $Z_w|(Y = y) \sim \mathrm{BIN}\big(1, w(y)\big)$. By working out the conditional expectation, it is obvious that $Y^\flat_w = \mathbb{E}_{Z_w|(Y=Y)}\varphi(Y, Z_w)$ coincides with the specification of the censored random variable in Subsection 3.1. For $w(y) = \mathbb{1}_A(y)$, the random variable Z_A degenerates to being one if $y \in A$ and zero otherwise, so that $Y^\flat_{Z_A} = Y^\flat_A$ with probability one. Correspondingly, the *Z-randomisation* definition of the censored scoring rule reads

$$S^\flat_w(\mathrm{F}, y) = \mathbb{E}_{Z_w|Y=y} S(\mathrm{F}^\flat_w, y^\flat_{Z_w}), \tag{4}$$

undeniably equivalent to the censored scoring rule defined by Definition 5.

A similar line of reasoning holds for the generalised censored scoring rule. In addition to the auxiliary random variable Z_w, we introduce an independent random variable Q with distribution H. Rather than labelling the observation as censored, we now take a random draw from Q if $Z_w = 1$, i.e. we define

$$y^\flat_{\mathrm{H},w} := \varphi_{w,\mathrm{H}}(y, Z_w, Q), \qquad \varphi_{w,\mathrm{H}}(y, Z_w, Q) := \begin{cases} Y, & \text{if } Z_w = 1, \\ Q, & \text{if } Z_w = 0. \end{cases}$$

As anticipated, the distribution of $Y^\flat_{\mathrm{H},w} = \mathbb{E}_{Z_w|(Y=Y),\mathrm{H}}\varphi(Y, Z_w)$ coincides with the

specification of $\mathrm{F}^{\flat}_{w,\mathrm{H}}$ in Equation (1). Additionally, the generalised censored scoring rule of Definition 6 admits the Z, Q-*randomisation* representation

$$S^{\flat}_{\mathrm{H},w}(\mathrm{F}, y) = \mathbb{E}_{Z_w|(Y=y),\mathrm{H}} S\big(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y^{\flat}_{\mathrm{H},w}\big).$$

The randomisation perspective further clarifies why $S^{\flat}_{\mathrm{H},w}$ generalises $S^{\flat}_{w}(\mathrm{F}, y)$. Indeed, by choosing a degenerate distribution for Q at $*$, each 'random draw' from Q will be precisely the censoring label $*$ of the Z-randomisation procedure. Put differently, $S^{\flat}_{\mathrm{H},w} = S^{\flat}_{w}(\mathrm{F}, y)$ for $\mathrm{H} = \delta_*$.

3.4 Examples

3.4.1 Semi-local scoring rules

We will now apply our censoring framework to the regular scoring rules introduced in Subsection 2.1. Following the classification into semi-local and distance-sensitive scoring rules, we start with localising the former class. Together with the main characteristics of the LogS, PowS_α and PsSphS_α families, Table 1 presents the localised versions of these families based on conditioning, censoring and generalised censoring. Given the strict propriety classes in Table 1, one can easily verify their strict propriety with respect to $\mathscr{f}^{\flat}_{\alpha}$ since $\|f^{\flat}_{w}\|^{\alpha}_{\alpha} \leq 1 + \|f\|^{\alpha}_{\alpha} < \infty$, $\forall f \in \mathscr{f}_\alpha$, $\forall w \in \mathcal{W}$, where $\alpha = 1$ for LogS. Furthermore, the Bregman generator functions $\zeta(t)$ refer to the well-known subclass of *separable Bregman divergences*, consisting of the score divergences based on strictly proper scoring rules $S_\zeta : \mathscr{p}(\mathcal{Y}, \mathcal{G}, \mu) \times \mathcal{Y} \to \mathbb{R}$ of the form

$$S_\zeta(p, y) = \zeta'(p(y)) - \int_{\mathcal{Y}} \zeta'(p(y))p(y) - \zeta(p(y))\mu(\mathrm{d}y).$$

Comparing the censored and conditioned versions of the rules, we notice that the censored variants bear an isolated $\bar{F}_w$-dependent term, preserving the coverage probability of $\{w = 0\}$. While preserving the likelihood $\bar{F}_w$ of being censored, Table 1 also shows that the censored scoring rules are independent of $*$, the label of a censored

observation. Hence, for this selection of scoring rules, one could alternatively work with an actual number like r for the location of the residual probability $\bar{F}_w$. Strictly speaking, we need to require $\mathrm{F}_w(r) = 0$ in that case, to keep the censored scoring rule strictly locally proper (see Assumption 1), but this is trivially met by restricting to either continuous measures or weight functions satisfying $w(r) = 0$, or both. The generalised censored scoring rules in Table 1 show that the invariance with respect to the location of the discrete probability mass holds more generally. In particular, the generalised censored scoring rules turn out to be entirely invariant to the choice of the nuisance density on $\{w = 0\}$ upon normalisation by the α-norm of h, i.e. to the class of densities $\tilde{h} = h/\|h\|_\alpha$, where $\alpha = 1$ for LogS. Since $\|h\|_1 = 1$, the latter means that LogS is invariant to the unnormalised choice of h, as can be seen from Table 1. Finally, Table 1 includes the localised divergence measures $\mathbb{D}_{S_w^\flat}$, which are all localised Bregman divergences since all regular divergences $\mathbb{D}_S$ in this table are Bregman.

3.4.2 Distance sensitive scoring rules

A rich class of distance-sensitive scoring rules is the Energy Score family

$$\mathrm{ES}_\beta(\mathrm{F}, y) = \frac{1}{2}\mathbb{E}_\mathrm{F}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|_2^\beta - \mathbb{E}_\mathrm{F}\|\mathbf{Y} - \mathbf{y}\|_2^\beta, \qquad \beta \in (0, 2),$$

known to be strictly proper to the class of Borel probability measures on $\mathbb{R}^d$ such that $\mathbb{E}_F\|\mathbf{Y}\|_2^\beta < \infty$ (Gneiting and Raftery, 2007). From this expression, it is immediately clear that the corresponding censored ES family depends, in contrast to the semi-local scoring rules, on $*$, or more particularly, the distance $d(\mathbf{y}) = \|\mathbf{y} - *\|_2$. Specifically,

$$S_{w,d}^\flat(\mathrm{F}, \mathbf{y}) = \frac{1}{2}\mathbb{E}_{\mathrm{F}_w^\flat}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|_2^\beta - \mathbb{E}_{\mathrm{F}_w^\flat}\Big(w(\mathbf{y})\|\mathbf{Y} - \mathbf{y}\|_2^\beta + \big(1 - w(\mathbf{y})\big)d(\mathbf{Y})^\beta\Big).$$

Of course, it is not surprising that distance sensitive scoring rules are sensitive to the location of the discrete probability $\bar{F}_w$. An easy way to define $d(\mathbf{y})$ is to simply add the location of $\bar{F}_w$ by choosing $* \in \mathbb{R}^d$. It is important, however, to keep in mind that the censored scoring rule is not invariant with respect to this additional piece of

information. More precisely, the selected value for $*$, say $\mathbf{r}$, is now not only representing the event of being censored but also the value an observation gets after being censored.

Assuming that the weight function at hand has a finite number of pivotal points $\mathcal{A} := \{\mathbf{a}_i\}_{i=1}^{n_a}$, e.g. the edge(s) of an indicator function, the centre of a kernel, etc., we consider the following two choices for the censored distance

(i) $d_{\text{rand}}(\mathbf{y}) = \left\| \mathbf{y} - \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{a}_i \right\|$

(ii) $d_{\min}(\mathbf{y}) = \min_{\mathcal{A}} \|\mathbf{y} - \mathbf{a}_i\|$

The first suggestion is equivalent to taking $\mathbf{r} = \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{a}_i$, making it straightforward to show that Theorem 1 applies. The second suggestion does not necessarily reduce to a choice for $\mathbf{r}$. Though, since this choice is in line with the assumptions of Theorem 1 of Székely and Rizzo (2005), one can still easily verify the assumption of Theorem 1.

We illustrate the role of the censored distance by two concrete examples. For the left-tail indicator function (the same holds for the right-tail indicator), Example 1 shows that both choices for d lead to a censored scoring rule coinciding with the twCRPS. This is an interesting result since the twCRPS is strictly locally proper for precisely these two types of weight functions (Holzmann and Klar, 2017, Theorem 5). Indeed, for the centre indicator function $w(y) = \mathbb{1}_{[-r,r]}(y)$, the twCRPS is knowingly failing to be strictly locally proper since it is non-localising. The corresponding censored scoring rules following from the censored distances in Example 2 are strictly locally proper and hence also still localising. In sharp contrast to the twCRPS, the censoring procedure enforces the weighted scoring rule to be localising by considering anything outside $[-r, r]$ as the same event. In this way, the censored rules prevent for the localisation bias introduced in Subsection 2.2, illustrated by Figure 1b.

Example 1. *Consider the CRPS, i.e. the ES_1 family for $d = 1$ and take $w(y) = \mathbb{1}_{(-\infty,r)}(y)$ as weight function. Following the examples of pivotal points of weight functions, we let $n_a = 1$ and $a_1 = r$. The associated censored distances become $d_{min}(y) = |y - r|$ and $d_{rand}(y) = |y - r|$. Hence, the choice of both censored distances*

reduce to replacing $*$ by r, leading to the scoring rule

$$\begin{aligned} CRPS^{\flat}_{w,d_{min}}(\mathrm{F}, y) &= CRPS^{\flat}_{w,d_{rand}}(\mathrm{F}, y) \\ &= \frac{1}{2}\mathbb{E}_{\mathrm{F}^{\flat}_w}|Y - \tilde{Y}| - \mathbb{E}_{\mathrm{F}^{\flat}_w}\Big(w(y)|Y - y| + \big(1 - w(y)\big)|Y - r|\Big) \\ &= twCRPS(F, y). \end{aligned}$$

Example 2. *Consider the CRPS and the centre indicator function $w(y) = \mathbb{1}_{[-r,r]}(y)$, for which $a_1 = -r$ and $a_2 = r$. The censored distances read $d_{min}(y) = |y - r| \wedge |y + r|$ and $d_{rand}(y) = |y|$. Both censored scoring rules do not coincide with the twCRPS, at which one arrives if we would put observations below $-r$ equal to $-r$ and observations above r to r. The latter is clearly not an example of a censored scoring rule, since it uses information outside the region of interest that is not implied by the information within the region of interest. The use of this additional information makes the twCRPS non-localising and hence prone to the localisation bias illustrated by Figure 1b. The verification of the strict propriety of the CRPS on the extended outcome space $\mathbb{R}^*$ with censored distance $d_{min}(y) = |y - r| \wedge |y + r|$ is deferred to the Online Appendix.*

Expanding upon the centre indicator example discussed in Example 2, it should be noted that the distances $d_{\min}$ and d_{rand} sometimes result in distances $|y - *|$ that are significantly off. For example, $d_{\min}$ can differ greatly from the non-censored distance between an observation $y \in A$ near $-r$ and a censored observation located far into the right tail of the distribution before censoring. Although these errors to some extent cancel each other out, this observation can also serve as an inspiration for improvement. In particular, we suggest to alternatively use the generalised censored distribution

$$\mathrm{dF}^{\flat}_w = \mathrm{dF}_w + \bar{F}_w\big(\gamma \mathrm{d}\delta_{a_1} + (1 - \gamma)\mathrm{d}\delta_{a_2}\big), \qquad a_1, a_2 \in \mathbb{R}, \gamma \in [0, 1], \tag{5}$$

distributing the residual probability $\bar{F}_w$ over the pivotal points a_1 and a_2 with proportions γ and $1 - \gamma$, respectively.

Unlike the twCRPS, this measure does not rely on the location of observations

outside the region of interest, except for the fact that they are not in A. The measure is a generalisation of the censored measure at a single critical point, and can in turn easily be verified to n_a pivotal points by taking $\mathrm{dH} = \sum_{i=1}^{n_a} \gamma_i \mathrm{d}\delta_i$ as reference distribution, with γ_i restricted to the unit simplex $\Delta(n)$ to maintain H as a probability measure. The choice of parameter γ depends on the specific application. For the indicator function $w(y) = \mathbb{1}_{[-r,r]}$ in Example 2, it is appropriate to select $\gamma = \frac{1}{2}$ when comparing the predictive performance of two candidates that are both symmetric around zero.

The parameter γ depends on the application at hand. For the indicator function $w(y) = \mathbb{1}_{[-r,r]}$ in Example 2 it makes sense to choose $\gamma = \frac{1}{2}$ if one aims to compare the predictive ability of two candidates that are both symmetric around zero. In these types of applications, data is typically available to estimate the proportion of residual probabilities of the candidates based on the DGP. It is important to note that using the data instead of the candidates to estimate γ, sets a level playing field for the candidates in terms of their performance on A. After all, this approach ensures that the relative performance of the candidates on A is not obscured by the performance outside A (for the part that is not entirely implied by the distribution on A).

Mathematically, we can illustrate the difference between the generalised censored scoring rule based on the censored measure in Equation (5) and the twCRPS as follows. For the centre indicator function $w(y) = \mathbb{1}_A(y)$, where $A = [a_1, a_2]$, we have the following equality

$$\mathrm{twCRPS}(\mathrm{F}, y) = \mathrm{CRPS}(\mathrm{F}_w^\dagger, y_w^\dagger), \quad \mathrm{dF}_w^\dagger = \mathrm{dF}_w + \bar{F}_w\big(\gamma_\mathrm{F}\mathrm{d}\delta_{a_1} + (1-\gamma_\mathrm{F})\mathrm{d}\delta_{a_2}\big)$$

where $\gamma_\mathrm{F} = F_{w\mathrm{L}}/\bar{F}_w$, $F_{w\mathrm{L}} = \mathrm{F}(A_\mathrm{L})$, $A_\mathrm{L}^c = (-\infty, a_1)$. Furthermore, $y_w^\dagger = y\mathbb{1}_A(y) + a_1\mathbb{1}_{A_\mathrm{L}^c}(y) + a_2\mathbb{1}_{A_\mathrm{R}^c}(y)$, with $A_\mathrm{R}^c = (a_2, \infty)$, allowing the twCRPS to assign different scores to observations in A_L^c and A_R^c. One crucial distinction between the generalised censored measure and $\mathrm{F}_w^\dagger$ is that the latter candidate's reference distribution depends on the candidate itself, namely through the dependence of the proportion parameter on F. In expectation, the difference between the twCRPS and the generalised censored

scoring rule reduces to precisely this difference between $\gamma = \mathrm{P}_{w\mathrm{L}}/\bar{P}_w$, where P denotes the underlying distribution of Y, and γ_{F}. Specifically,

$$\mathbb{E}_{\mathrm{P}}\mathrm{twCRPS}(\mathrm{F}, Y) = \mathbb{E}_{\mathrm{P}}\mathrm{CRPS}^{\dagger}_{w}(\mathrm{F}, Y),$$

where the only difference between $\mathrm{CRPS}^{\dagger}_{w}$ and $\mathrm{CRPS}^{\flat}_{w}$ is the dependence on $\mathrm{F}^{\dagger}_{w}$ rather than $\mathrm{F}^{\flat}_{w}$, i.e.

$$\mathrm{CRPS}^{\dagger}_{w}(\mathrm{F}, y) = \begin{cases} \mathrm{CRPS}(\mathrm{F}^{\dagger}_{w}, y), & \text{if } y \in A \\ \gamma\mathrm{CRPS}(\mathrm{F}^{\dagger}_{w}, a_1) + (1-\gamma)(\mathrm{F}^{\dagger}_{w}, a_2), & \text{if } y \in A^c, \end{cases}$$

which, contrary to the twCRPS, does not depend on whether an observation is in A_{L} or A_{R}.

For centre indicator case, for which the twCRPS is not strictly locally proper and hence not a generalised censored scoring rule, we have now derived the alternative (close to censoring) procedure, which is helpful in two ways. (i) By revealing the recipe for obtaining the twCRPS, we uncovered the multivariate twCPRS for practioners that are despite the localisation-bias still willing to use the twCRPS in a multivariate setting. (ii) We have uncovered precisely the difference between the twCRPS and the generalised censored scoring rule, i.e. γ versus γ_{F} in the definition of the focused measure.

3.5 Localised Neyman–Pearson

In anticipation of our favourite applications, we now switch to an explicit time series context. In particular, consider a stochastic process $\{Y_t : \Omega \to \mathcal{Y}\}_{t=1}^{T}$ from a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\mathcal{Y}^T, \mathcal{G}^T)$, where $\mathcal{Y}^T$ and $\mathcal{G}^T$ denote the product outcome space and σ-algebra of the individual outcome spaces $\mathcal{Y}$ and σ-algebras $\mathcal{G}$, respectively. The process generates the filtration $\{\mathcal{F}_t\}_{t=1}^{T}$, in which $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$ is the information set at time t, satisfying $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$, $\forall t$. We denote

predictive distributions of Y_{t+1} based on $\mathcal{F}_t$ by F_t, predictive distribution functions by F_t and predictive μ_t-densities by f_t. The existence of the sequence of densities f_t is implied by the existence of a sequence of measures $\{\mu_t\}$ such that $\mathrm{F}_t \ll \mu_t, \forall t$. Furthermore, the regions of interest $A_t \subseteq \mathcal{Y}$ are always assumed to be $\mathcal{F}_t$-measurable.

The aim of this section is to derive a uniformly most powerful (UMP) test for the following null and alternative hypothesis

$$\mathbb{H}_0 : p_{0t}\mathbb{1}_{A_t} = f_{0t}\mathbb{1}_{A_t},\ \forall t \qquad \text{vs} \qquad \mathbb{H}_1 : p_{1t}\mathbb{1}_{A_t} = f_{1t}\mathbb{1}_{A_t}, \quad \forall t. \tag{6}$$

Although the predictive densities $f_{jt} = \frac{\mathrm{F}_{jt}}{\mathrm{d}\mu_t}$, $j \in \{0,1\}$, are assumed to be known, the testing problem remains a multiple versus multiple hypothesis test due to the lacking specification of the density outside the regions of interest A_t. Yet, since the densities p_{jt} must integrate to one on $A_t \cup A_t^c$, the null hypothesis does imply that these densities integrate to $\mathrm{F}_{jt}(A_t^c)$ on A_t^c. Therefore, the implied specification on A_t^c can be summarised as

$$\frac{\mathrm{F}_{jt}(A^c)}{\mathrm{H}_{jt}(A^c)} h_{jt}\mathbb{1}_{A_t^c} = \mathrm{F}_{jt}(A^c)[h_{jt}]^{\sharp}_{A_t^c}\mathbb{1}_{A_t^c}, \quad j \in \{0,1\},$$

where the unknown densities $h_{jt} = \frac{\mathrm{H}_{jt}}{\mathrm{d}\mu_t}$ can be seen as infinite dimensional nuisance parameters.

Explicitising the implied assumption on A_t^c in the hypotheses and phrasing them in terms of a statement about the whole sample distribution leads to the following equivalent hypotheses

$$\mathbb{H}_j : p_j(\mathbf{y}) = \prod_{t=0}^{T-1} \left(f_{jt}(y_{t+1})\mathbb{1}_{A_t}(y_{t+1}) + \mathrm{F}_{jt}(A^c)[h_{jt}]^{\sharp}_{A_t^c}(y_{t+1})\mathbb{1}_{A_t^c}(y_{t+1}) \right), \quad j \in \{0,1\}.$$

Since the densities f_{jt} are fixed, and the densities h_{jt} are unrestricted under both hypothesis, the class of densities satisfying hypothesis $\mathbb{H}_j$ can alternatively be written

as

$$\mathscr{p}_j = \left\{ \prod_{t=0}^{T-1} \left(f_j(y_{t+1}) \mathbb{1}_{A_t}(y_{t+1}) + \mathrm{F}_{jt}(A^c)[h_{jt}]^\sharp_{A_t^c}(y_{t+1}) \mathbb{1}_{A_t^c}(y_{t+1}) \right), h_j \in \mathscr{h} \right\}, \quad j \in \{0,1\},$$

in which $\mathscr{h}$ denotes the space of all densities on $A^c = \prod_{t=0}^{T-1} A_t^c$.

Let $\phi : \mathcal{Y}^T \to [0,1]$ denote a test function determining which values should be included in the critical region. In terms of the index set of all observations $\mathcal{I} = \{1, \ldots, T\}$, this space can also be denoted as $\mathcal{Y}(\mathcal{I}) = \prod_{t \in \mathcal{I}} \mathcal{Y}_t$. The aim of this section is to find a uniformly most powerful (UMP) test ϕ^* of size α for testing problem (6), i.e. a solution to the maximisation problem

$$\max_{\phi \in \Phi(\alpha)} \mathbb{E}_{p_1} \phi, \qquad \Phi(\alpha) = \{\phi : \sup_{p_0 \in \mathscr{p}_0} \mathbb{E}_{p_0} \phi \leq \alpha\}. \tag{7}$$

As a first step toward the solution, given by Theorem 3, let us fix an $h_1 \in \mathscr{h}$ so that the distribution under the alternative is completely known. Given the fact that the hypotheses are, in the end, silent about the shape of the density on A^c, we conjecture that a UMP test neglects the information about the shape of the density on A^c. If $T = 2$, for example, and we consider the optimal test on $A_1 \times A_2^c$, our intuition is that an optimal test does not care about the shape of $[h_2]^\sharp_{A_2{}^c}$, that is, the specific values $[h_2]^\sharp_{A_2{}^c}(y_2)$ for all $y_2 \in A_2{}^c$, but just about the total probability of an outcome falling into A_2^c. In other words, we expect that a solution to problem (7) has integrated out the dependence on the nuisance densities.

Although it is obvious that marginalising out the still assumed to be fixed density $h_1 \in \mathscr{h}$ is harmless in terms of power, it is non-trivial that this is an affordable strategy in terms of size for all $h_0 \in \mathscr{h}$. Lemma 1 and its proof show that the subclass of tests disregarding information about the shape of h_1 is guaranteed to be size correct. In our search for the UMP test, Corollary 1 then formalises the idea that we can restrict our attention to tests of the conjectured kind.

Lemma 1. *Consider testing problem* (6) *and suppose that the outcomes* $(y_t)_{t \in \mathcal{I}_A}$ *are*

in A_t, and the remaining $n-k$, with $k=|\mathcal{I}_A|$, observations $(y_t)_{t\in\mathcal{I}_{A^c}}$ in A_t^c. For an arbitrary but fixed density $h_1\in\mathscr{h}$, the test

$$\psi_{h_1}:\mathcal{Y}^T\to[0,1],\quad \psi_{h_1}=\int_{\mathcal{Y}(\mathcal{I}_{A^c})}\phi^*_{h_1}\prod_{t\in\mathcal{I}_{A^c}}[h_{1t}]^\sharp_{A_t^c}\mathbb{1}_{A_t^c}\mathrm{d}\mu_t$$

where $\phi^*_{h_1}$ denotes a solution to problem (7), is such that $\psi_{h_1}\in\Phi(\alpha)$.

Corollary 1. *Consider testing problem (6) and suppose that the outcomes $(y_t)_{t\in\mathcal{I}_A}$ are in A_t, and the remaining $T-k$, with $k=|\mathcal{I}_A|$, observations $(y_t)_{t\in\mathcal{I}_{A^c}}$ in A_t^c. Let $\Psi(\alpha)\subseteq\Phi(\alpha)$ denote the class of size α tests on $\mathcal{Y}^T$ that are constant in arguments varying in $\mathcal{Y}(\mathcal{I}_{A^c})$. Then,*

$$\max_{\phi\in\Phi(\alpha)}\mathbb{E}_{p_1}\phi=\max_{\psi\in\Psi(\alpha)}\mathbb{E}_{p_1}\psi,\qquad\forall h_1\in\mathscr{h}.$$

For any fixed $h_1\in\mathscr{h}$, the reduced optimisation problem resulting from Corollary 1, simplifies to a simple versus simple hypothesis in terms of the censored measures $\mathrm{d}[\mathrm{F}_{jt}]^\flat_{A_t}=\mathbb{1}_{A_t}\mathrm{d}\mathrm{F}_{jt}+\mathrm{F}_{jt}(A_t^c)\mathrm{d}\delta_*$, enabling us to formalise a localised version of the Fundamental Lemma of Neyman and Pearson (1933), included below as Theorem 3.

Theorem 3 (Localised Neyman-Pearson)**.** *The UMP test for testing problem (6) is given by*

$$\phi^\flat_A(\mathbf{y})=\begin{cases}1, & \text{if } \lambda(\mathbf{y})>c\\ \gamma & \text{if } \lambda(\mathbf{y})=c\\ 0, & \text{if } \lambda(\mathbf{y})<c\end{cases}\qquad\lambda(\mathbf{y})=\frac{[f_1]^\flat_A(\mathbf{y})}{[f_0]^\flat_A(\mathbf{y})},\qquad[f_j]^\flat_A(\mathbf{y})=\prod_{t=0}^{T-1}[f_{jt}]^\flat_{A_t}(y_{t+1}),$$

where $j\in\{0,1\}$ and c is the largest constant such that $[\mathrm{F}_0]^\flat_A\big(\lambda(\mathbf{y})\geq c\big)\geq\alpha$ and $[\mathrm{F}_0]^\flat_A\big(\lambda(\mathbf{y})\leq c\big)\geq 1-\alpha$, and $\gamma\in[0,1]$ is such that $\alpha=[\mathrm{F}_0]^\flat_A\big(\lambda(\mathbf{y})>c\big)+\gamma[\mathrm{F}_0]^\flat_A\big(\lambda(\mathbf{y})=c\big)$.

It is worth emphasising that the obtained equivalence between testing problem (6)

and $\mathbb{H}_j:\ p_j = [f_j]^{\flat}_{A_t}$, $j \in \{0,1\}$, is a priori unobvious, since

$$p_j = \Big(f_j \mathbb{1}_A + \mathrm{F}_j(A^c)[h_j]^{\sharp}_{A^c}\mathbb{1}_{A^c}\Big) \implies p_j = [f_j]^{\flat}_A,$$

but not the other way around. Formulated differently,

$$\mathbb{H}_j : [p_j]^{\flat}_A = [f_j]^{\flat}_A$$

is a multiple versus multiple hypothesis about p_j (for example satisfied if $p_j = [f_j]^{\flat}_A$), but a simple versus simple hypothesis about $[p_j]^{\flat}_A$. Furthermore, we have included an example for the special case that $T = 1$ in the Online Appendix, showing that we arrive at the same solution as Holzmann and Klar (2016) for this special case.

We close this section with two corollaries of Theorem 3, the proofs of which are deferred to the Online Appendix. Corollary 2 reveals that, unsurprisingly, the localised NP test given by Theorem 3 can alternatively be formulated by the censored likelihood score of Diks et al. (2011). Corollary 3 ensures that the conditional operator does not bear a UMP test too, making the censored operator strictly preferred over the conditional one in the current setting.

Corollary 2. *Another formulation of the UMP test for testing problem (6) is given by the test defined in Theorem 3, with $\lambda(\mathbf{y})$ replaced by $\tilde{\lambda}(\mathbf{y}) = \sum_{t=0}^{T-1}\big(S^{csl}_{A_t}(f_{1t}, y_{t+1}) - S^{csl}_{A_t}(f_{0t}, y_{t+1})\big)$, where $S^{csl}_{A_t}$ denotes the censored likelihood score (csl) proposed by Diks et al. (2011).*

Corollary 3. *For testing problem (6), the test*

$$\phi^{\sharp}_A(\mathbf{y}) = \begin{cases} 1, & \text{if } \lambda^{\sharp}(\mathbf{y}) > c \\ \gamma & \text{if } \lambda^{\sharp}(\mathbf{y}) = c \\ 0, & \text{if } \lambda^{\sharp}(\mathbf{y}) < c \end{cases} \qquad \lambda^{\sharp}_A(\mathbf{y}) = \frac{[f_1]^{\sharp}_A(\mathbf{y})}{[f_0]^{\sharp}_A(\mathbf{y})}\mathbb{1}_A(\mathbf{y}), \qquad [f_j]^{\sharp}_A(\mathbf{y}) = \prod_{t=1}^{T}[f_{jt}]^{\sharp}_{A_t}(y_t),$$

where $j \in \{0,1\}$ and c is the largest constant such that $[\mathrm{F}_0]^{\flat}_A\big(\lambda(\mathbf{y}) \geq c\big) \geq \alpha$ and

$[\mathrm{F}_0]_A^{\flat}\big(\lambda(\mathbf{y}) \leq c\big) \geq 1-\alpha$, *and* $\gamma \in [0,1]$ *is such that* $\alpha = [\mathrm{F}_0]_A^{\flat}\big(\lambda(\mathbf{y}) > c\big) + \gamma[\mathrm{F}_0]_A^{\flat}\big(\lambda(\mathbf{y}) = c\big)$, *is not UMP.*

Table 1: Examples semi-local scoring rules

Name	Logarithmic	Power family	PseudoSpherical family
		Regular	
$S(f,y)$	$\mathrm{LogS}(f,y) = \log f(y)$	$\mathrm{PowS}_\alpha(f,y) = \alpha f(y)^{\alpha-1} - (\alpha-1)\|f\|_\alpha^\alpha$	$\mathrm{PsSphS}_\alpha(f,y) = \frac{f(y)^{\alpha-1}}{\|f\|_\alpha^{\alpha-1}}$
Special cases		$\mathrm{QS}(f,y) = \mathrm{PowS}_2(f,y)$	$\mathrm{SphS}(f,y) = \mathrm{PsSphS}_2(f,y)$
		$\mathrm{LogS}(f,y) = \lim_{\alpha\downarrow 1}\mathrm{PowS}_\alpha(f,y)$	$\mathrm{LogS}(f,y) = \lim_{\alpha\downarrow 1}\mathrm{PsSphS}_\alpha(f,y)$
$\mathbb{H}_S(f)$	$\mathbb{E}_f \log f$	$\|f\|_\alpha^\alpha$	$\|f\|_\alpha$
$\mathbb{D}_S(f\|g)$	$\mathrm{KL}(f\|g) = \mathbb{E}_f \log\left(\frac{f}{g}\right)$	$\|f\|_\alpha^\alpha - \alpha\int g^{\alpha-1}(f-g)\mathrm{d}\mu - \|g\|_\alpha^\alpha$	$\|f\|_\alpha - \frac{\int f g^{\alpha-1}\mathrm{d}\mu}{\|g\|_\alpha^{\alpha-1}}$
$\alpha = 2$	–	$\|f-g\|_2^2$	$\|f\|_2\,(1 - C(f,g))$
SP class	$\not\ell_1$	$\not\ell_\alpha$	$\not\ell_\alpha$
$\zeta(t)$	$t\log t$	t^α	–
$S(\tilde{f},\tilde{y})$	$\log f(y) - \log\|b\|$	$\left(\frac{1}{\|b\|}\right)^{\alpha-1}\mathrm{PowS}_\alpha(f,y)$	$\left(\frac{1}{\|b\|}\right)^{\frac{\alpha-1}{\alpha}}\mathrm{PsSphS}_\alpha(f,y)$
		Focused	
$S_w^\sharp(f,y)$	$w(y)\log\left(\frac{f(y)}{1-F_w}\right)$	$w(y)\left(\alpha\left(\frac{f_w(y)}{1-F_w}\right)^{\alpha-1} - (\alpha-1)\left\|\frac{f_w(y)}{1-F_w}\right\|_\alpha^\alpha\right)$	$w(y)\frac{f_w(y)^{\alpha-1}}{\|f_w\|_\alpha^{\alpha-1}}$
$S_w^\flat(f,y)$	$w(y)\log f(y) + \big(1-w(y)\big)\log\bar{F}_w$	$w(y)\alpha f_w(y)^{\alpha-1} + \big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1}$ $-(\alpha-1)\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1} + (1-w(y))\bar{F}_w^{\alpha-1}}{\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$S_{w,h}^\flat(f,y)$	$w(y)\log f(y) + \big(1-w(y)\big)\log\bar{F}_w$	$w(y)\alpha f_w(y)^{\alpha-1} + \big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1}\|h\|_\alpha^\alpha$ $-(\alpha-1)\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\|h\|_\alpha^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1} + \big(1-w(y)\big)\bar{F}_w^{\alpha-1}\|h\|_\alpha^\alpha}{\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\|h\|_\alpha^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$\mathbb{H}_{S_w^\flat}(f)$	$\int\log(f)f_w\mathrm{d}\mu + \log\left(\bar{F}_w\right)\bar{F}_w + \int\log(w)f_w\mathrm{d}\mu$	$\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha$	$\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\right)^{\frac{1}{\alpha}}$
$\mathbb{D}_{S_w^\flat}(f\|g)$	$\int\log\left(\frac{f}{g}\right)f_w\mathrm{d}\mu + \log\left(\frac{\bar{F}_w}{\bar{G}_w}\right)\bar{F}_w$	$\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha - \int f_w g_w^{\alpha-1}\mathrm{d}\mu - \bar{F}_w\bar{G}_w^{\alpha-1}$ $-(\alpha-1)\big(\|g_w\|_\alpha^\alpha + \bar{G}_w^\alpha\big)$	$\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\right)^{\frac{1}{\alpha}} - \frac{\int f_w g_w^{\alpha-1}\mathrm{d}\mu + \bar{F}_w\bar{G}_w^{\alpha-1}}{\left(\|g_w\|_\alpha^\alpha + \bar{G}_w^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$

Note: $C(f,g) = \frac{\int fg\mathrm{d}\mu}{\sqrt{\int f^2\mathrm{d}\mu\int g^2\mathrm{d}\mu}}$ and $\mathbb{D}_S(f||g)$ denote the cosine similarity and the score divergence between f and g, respectively. $\mathbb{H}_S(f)$ denotes the negative entropy and the PowS_α and PsSphS_α families are restricted to $\alpha > 1$. Furthermore, $\not\ell_\alpha$ denotes the space for which the L^α-norm is finite, where μ is measure relative to which the densities p and f are defined, i.e. $\frac{\mathrm{dF}}{\mathrm{d}\mu}$, with $\mathrm{F} \ll \mu$. The common limiting case of the PowS_α and PsSphS_α remains to hold for conditioning and censoring, i.e. $\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PsSphS}_{\alpha,w}^x(f,y) = \lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PowS}_{\alpha,w}^x(f,y) = \mathrm{LogS}^x(f,y)$, $x \in \{\sharp,\flat\}$. The generalised censored distribution $S_{w,h}^\flat$ departs from a density h which support is a subset of $\{w=0\} \subseteq \mathcal{Y}$. The weight function is restricted accordingly.

4 Monte Carlo simulation

In this section, we compare the size and power properties of the conditional and censored scoring rules of a selection of regular scoring rules. The selected simulation design is similar to the ones conducted by Diks et al. (2011), Lerch et al. (2017) and Holzmann and Klar (2016). In particular, we use the score difference series of two candidates F and G, that is, realisations of $D = S(\mathrm{F}, Y) - S(\mathrm{G}, Y)$ to employ the Giacomini and White (2006) test, for the null hypothesis

$$\mathbb{H}_0 : \mathbb{E}_{\mathrm{P}} S(\mathrm{F}, Y) = \mathbb{E}_{\mathrm{P}} S(\mathrm{G}, Y),$$

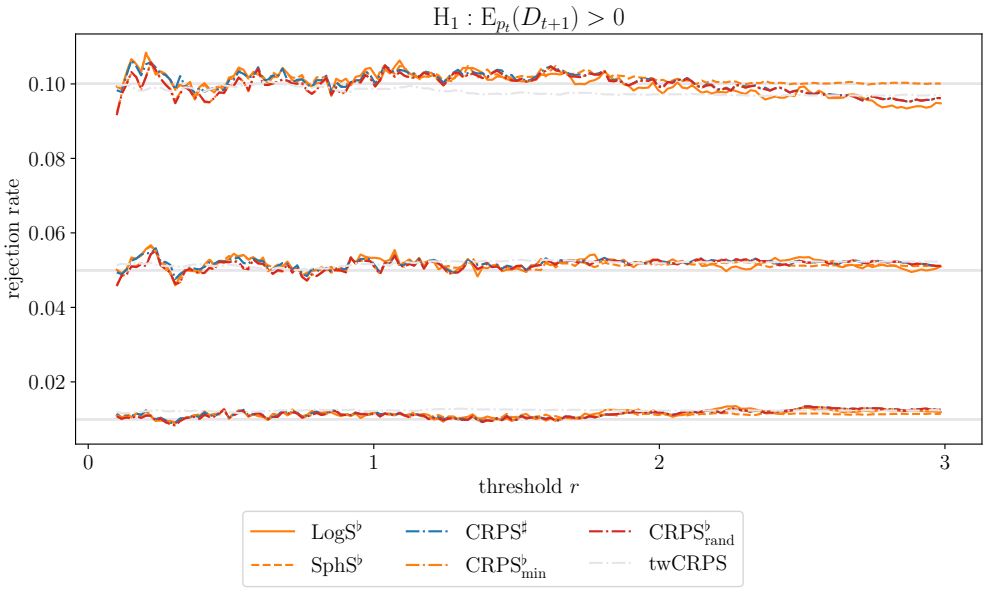
by means of the Diebold Mariano-type statistic $t_T = \frac{\frac{1}{T}\sum_{t=1}^{T} d_t}{\sqrt{\hat{\sigma}_t^2/T}}$, where $\hat{\sigma}_t$ should be a heteroskedasticity and autocorrelation-consistent (HAC) variance estimator in non-i.i.d. settings. This null hypothesis, which is equivalent to $\mathbb{H}_0 : \mathbb{D}_S(\mathrm{P}\|\mathrm{F}) = \mathbb{D}_S(\mathrm{F}\|\mathrm{G})$, is rejected if it is unlikely enough that quoting F instead of P leads to the same information loss as quoting G instead of P.

A natural conjecture is that strictly locally proper scoring rules generally lead to higher power since they are sensible with respect to all measurable aspects of the distribution. When comparing conditional and censored scoring rules, we thus expect the censored scoring rules to have more power. Though, it is worth mentioning that the simulation study of Diks et al. (2011) already presents some examples in which the conditional scoring rule indicates higher power. This can be explained by understanding that some candidates do not have aspects for which the conditional scoring rule has a blind spot. In such cases, the conditioning transformation can either enhance or alleviate the discriminative ability of the scoring rule when compared to censoring.

4.1 Size

As carefully explained by Diks et al. (2011), the null hypothesis of the GW test forces a particularly symmetric design. We adopt the design of Diks et al. (2011), using a centre-

Figure 2: Size properties GW test



indicator weight function $w(y) = \mathbb{1}_{[-r,r]}(y)$ combined with and i.i.d. standard normal DGP and normal candidates with unit variance and means $\mu_f = -0.2$ and $\mu_g = 0.2$. Due to the symmetry, the norms and $\bar{F}_w$-probabilities of the candidates are equivalent, leading to coinciding DM statistics based on QS and SphS scoring rules. Additionally, the equal norms and discrete probabilities also imply the censoring and conditioning rules to be equivalent within a semi-local scoring rule family since observations outside the region of interest obtain the same scores under both candidates in this case.

Figure 2 displays the rejection rates for rejection the null of equal predictive ability against the one-sided alternative that candidate f is statistically closer to p than g. The rejection rates are given at nominal significance levels of 0.01, 0.05 and 0.10, for focused versions of the LogS, SphS and CRPS scoring rules, based on 10,000 simulations. Given the discussion above, this gives a complete picture of the selection $\{\text{LogS}, \text{SphS}, \text{QS}, \text{CRPS}\} \times \{\sharp, \flat\}$. The twCRPS is added since it will also be included as benchmark in the power studies based on weight functions for which the censored CRPS variants do not reduce to the twCRPS (see Example 2). None of the displayed rejection rates give reason to doubt the size correctness of the tests.

4.2 Power

4.2.1 Laplace tails

Our first experiment studies the consequences of the lack of the conditional rule to disentangle two proportional tails when using the left tail indicator function $w(y) = \mathbb{1}_{(-\infty,r)}(y)$. In particular, we follow up on the Laplace example given by Figure 1a in Subsection 2.2, analysing two Laplace candidates with different location $\mu_f = -1$ and $\mu_g = 1$ but equivalent scale $\theta_f = \theta_g = 1$. Interestingly, even if $\mu_p \to \mu_f$, the conditional scoring rule does not have any power against the null of the candidates being statically equally far away from p, that is, for thresholds $r < \mu_f$, for which the conditional distributions on $(-\infty, r)$ coincide. Since movements of p in terms of μ_p are invisible through the lens of a conditional score divergence, this is essentially not a lack of power against $\mathbb{H}_0$, which is based on the conditional scoring rule. Yet, it is a lack of power against the distributions being statistically equally far away from the actual density on $\{w > 0\}$ through the lens of the regular score divergence and, therefore, still a lack of local discriminative ability. More fundamentally, the GW test degenerates in this case, as the score differences are exactly zero.

Leaving this extreme case, we analyse what happens if the scale parameters are not exactly the same, but close. Specifically, we let $\theta_f = 1$ and $\theta_g = 1.1$. Figure 3 shows the rejection rates of the GW test if the DGP is f (left-hand side) or g (right-hand side) in favour of f (top) or g (bottom). Since both candidates are now also different through the lens of the conditional rule, the subfigures on the diagonal display actual power while the off-diagonal ones show spurious power. Concerning the selection of scoring rules, it is good to remember that both censored CRPS rules (based on d_{min} and d_{rand}) coincide with each other and with the twCRPS, see Example 1 for details.

Three observations strike us. First of all, the increase in power from the conditional operator to the censoring operator is immense for all four scoring rules and thresholds $r < \mu_f$. The difference decreases over the interval $r \in (\mu_f, \mu_g)$, after which both conditioning and censoring have close to unit power. This observation is in line with the

Figure 3: Laplace experiment ($c = 20$)

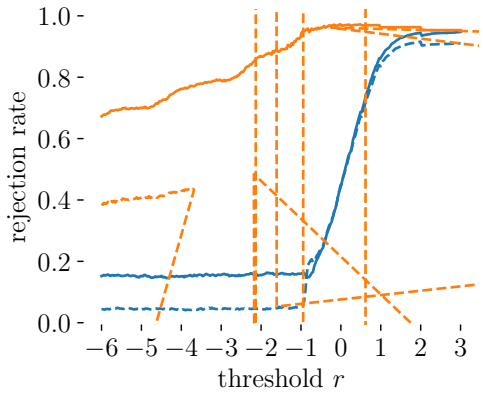
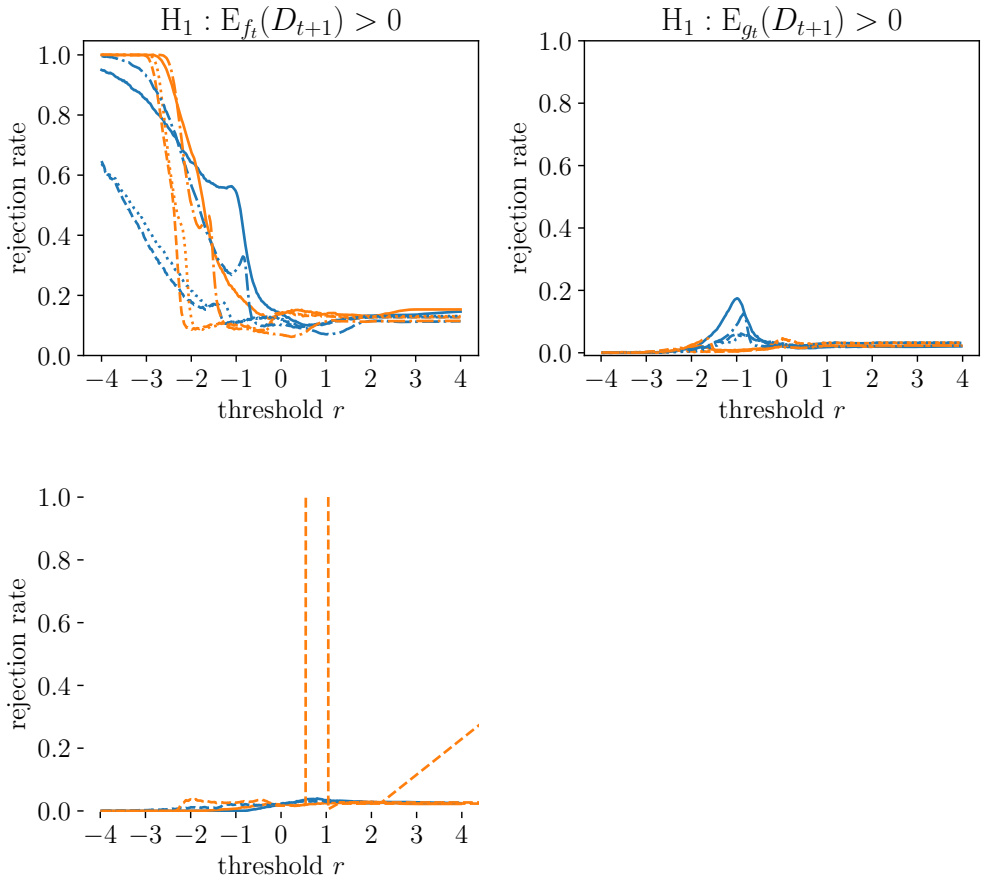


Figure 4: $\mathcal{N}(0,1)$ versus Student-$t(5)$: Left-tail ($c = 20$)



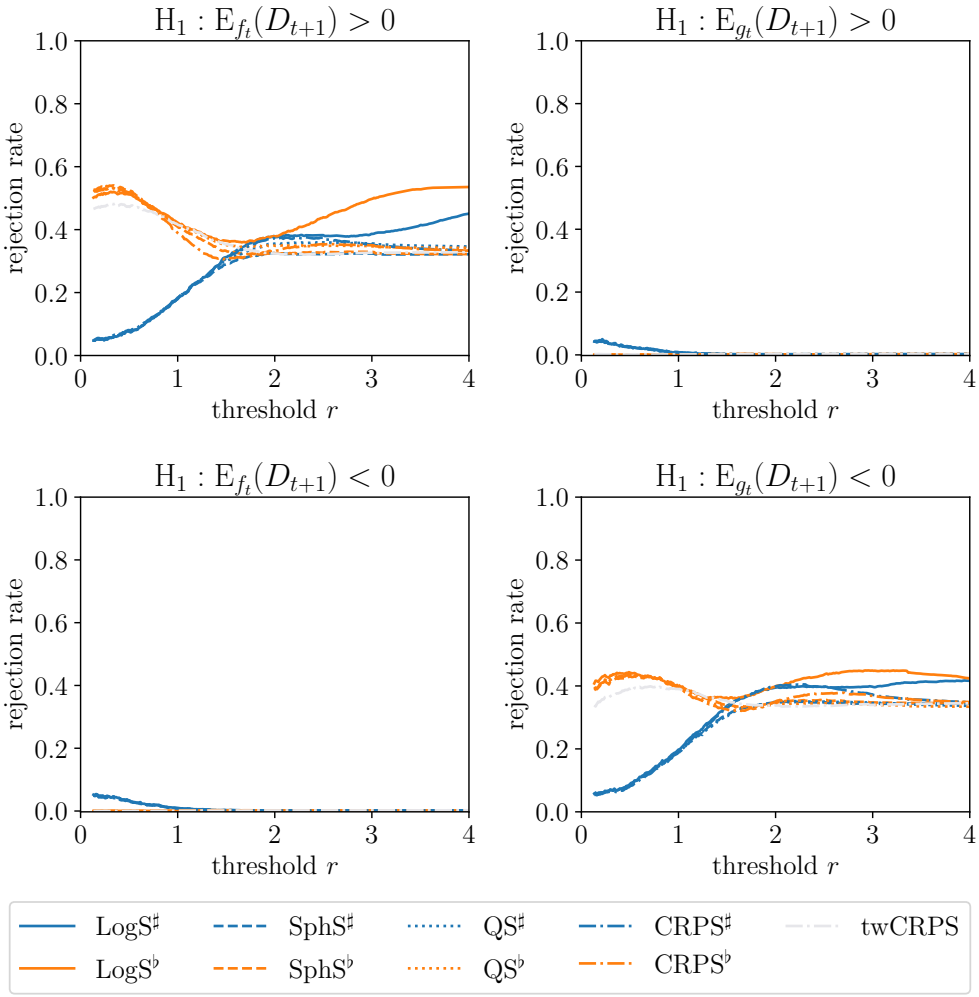
less monotonic, intersecting the graph of the competing focusing operator rejection rates. The latter occurs by construction since the densities of the candidates intersect as well, see Diks et al. (2011) for a discussion. Starting with the clearest differences, we note the spurious power humps of the conditional rules if the Student-$t(5)$ distribution is the DGP. In contrast, the censored scoring rules have almost no spurious power. Staying in the right column of Figure 4, the rejection rates in the bottom row reveal a clear preference for the censoring operator. Indeed, the exceptions of higher conditional power are rather weak, while the difference between the rejection rates (far) into the left-tail is particularly large for the Logarithmic and Spherical scoring rule. On the other hand, if the standard normal distribution is the DGP, then there is hardly (a difference in) spurious power. The differences between the rejection rates representing power are more extreme when the data is generated from the standard Normal distribution, yet so is their drop between $r = -2$ and $r = -1$, clouding a clear preference for one of the focusing operators for these intermediate tail values of r.

4.2.3 $\mathcal{N}(0,1)$ versus Student-$t(5)$: Centre

In our third Monte Carlo experiment, we focus on the centre of the candidate distributions by implementing the weight function $w(y) = \mathbb{1}_{[-r,r]}(y)$. Figure 5 displays the rejection rates for the same selection of regular scoring rules as in the previous experiments. Based on Figure 5, the added value of censoring relative to conditioning is overwhelming: Censoring leads to higher power and lower spurious power, in particular for values smaller than $r = 1$, which are of particular interest in applications.

As discussed in Example 2, the variants of the censored CRPS no longer coincide with each other, nor with the twCRPS. The $\text{CRPS}_w^\flat$ displayed in the Figure 5 is the generalised censored scoring rule based on the generalised censored measure in Equation (5). Due to the symmetry of the set up, there is visually no difference between using the suggested value $\gamma = \frac{1}{2}$ or the estimated proportion $\hat{\gamma}$. We have also calculated the $\text{CRPS}^\dagger(\text{F}, y)$ introduced in Subsection 3.4.2, which visually coincides with the twCRPS in this case. As can be seen from Figure 6, the alternative-distance based

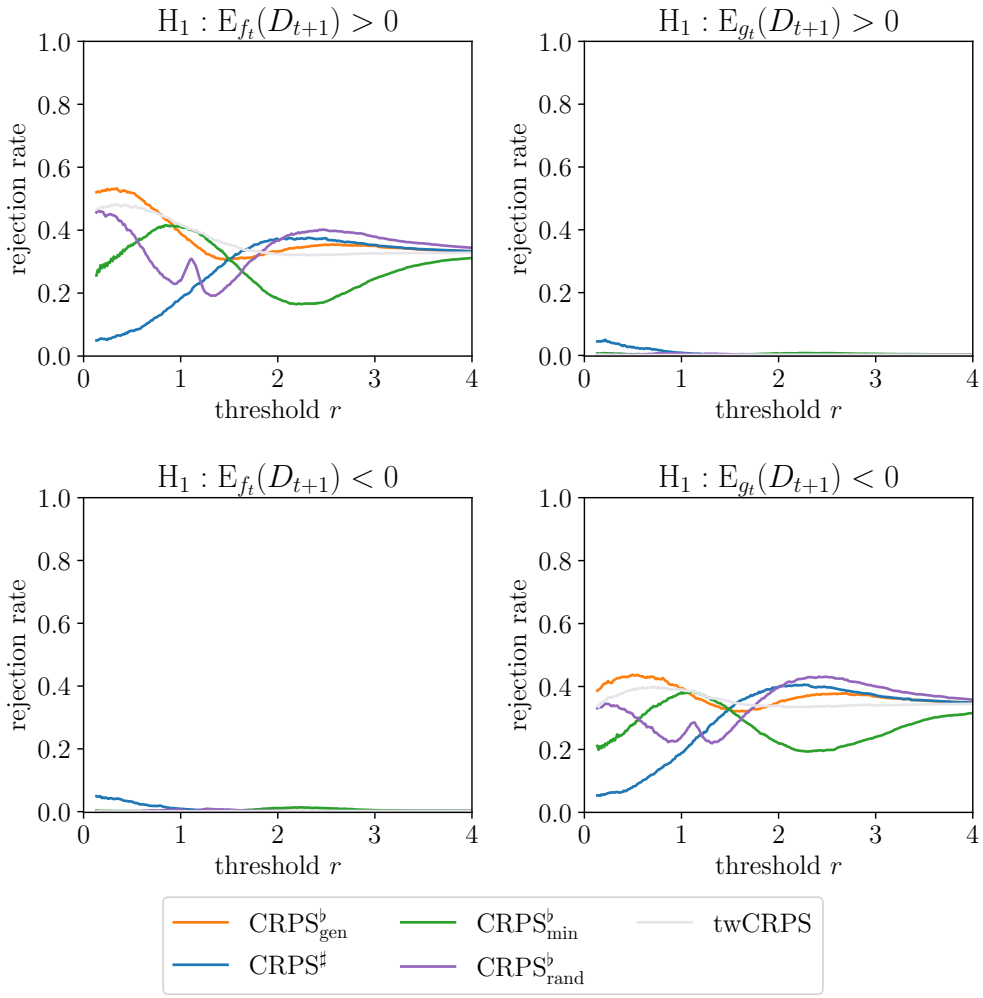
Figure 5: $\mathcal{N}(0,1)$ versus Student-$t(5)$: Centre ($c = 200$)



One-sided rejection rates of the GW-test of equal predictive ability of the candidates f_t (standard normal) and g_t (Student-$t(5)$) 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). Moreover, 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 $w(y) = 1_{(-r,r)}(y)$ and the number of expected observations in the region of interest is kept constant at $c = 200$.

CRPS variants do not coincide with the generalised censored CRPS, but lead to lower power than the generalised censored CRPS. Reconsidering the discussion of Subsection 3.4.2, this is unsurprising, since the generalised censored scoring rule incorporates more of the available information.

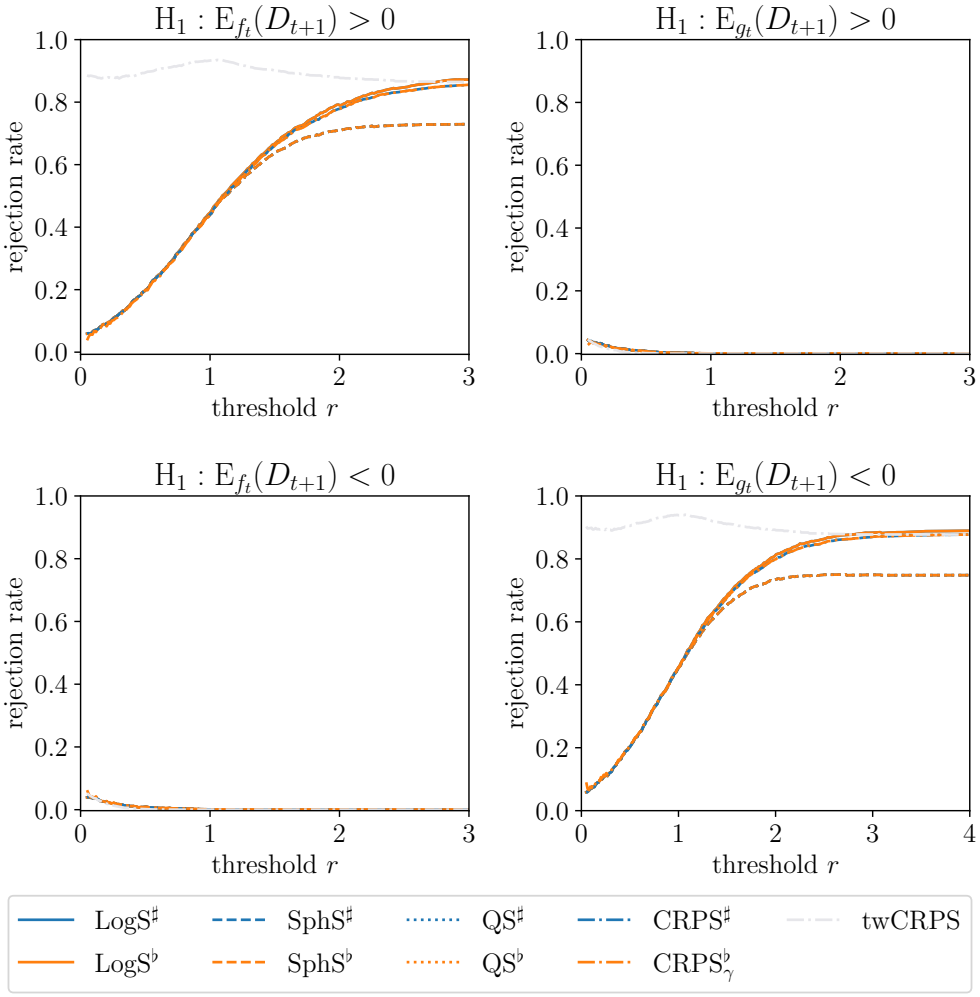
Figure 6: $\mathcal{N}(0,1)$ versus Student-$t(5)$: Centre ($c = 200$)



One-sided rejection rates of the GW test of equal predictive ability of the candidates f_t (standard normal) and g_t (Student-$t(5)$) 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). Moreover, 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 $w(y) = 1_{[-r,r]}(y)$ and the number of expected observations in the region of interest is kept constant at $c = 200$.

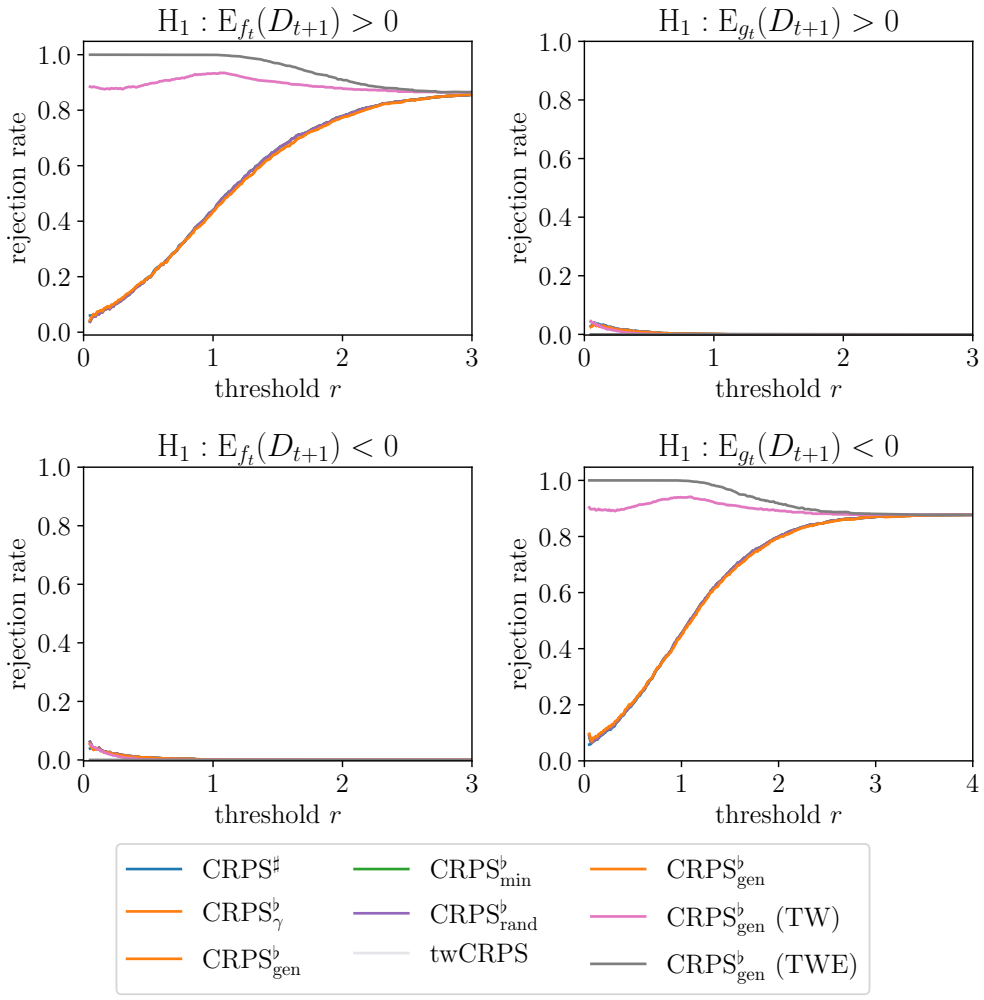
4.2.4 $\mathcal{N}(-0.2, 1)$ versus $\mathcal{N}(-0.2, 1)$: Centre ($c = 200$) [$\gamma_{\mathrm{F}} \neq \gamma_{\mathrm{P}}$]

Figure 7: $\mathcal{N}(-0.2, 1)$ versus $\mathcal{N}(-0.2, 1)$: Centre ($c = 200$)



One-sided rejection rates of the GW test of equal predictive ability of the candidates f_t ($\mathcal{N}(-0.2, 1)$) and g_t ($\mathcal{N}(0.2, 1)$) 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). Moreover, 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 $w(y) = 1_{[-r,r]}(y)$ and the number of expected observations in the region of interest is kept constant at $c = 200$.

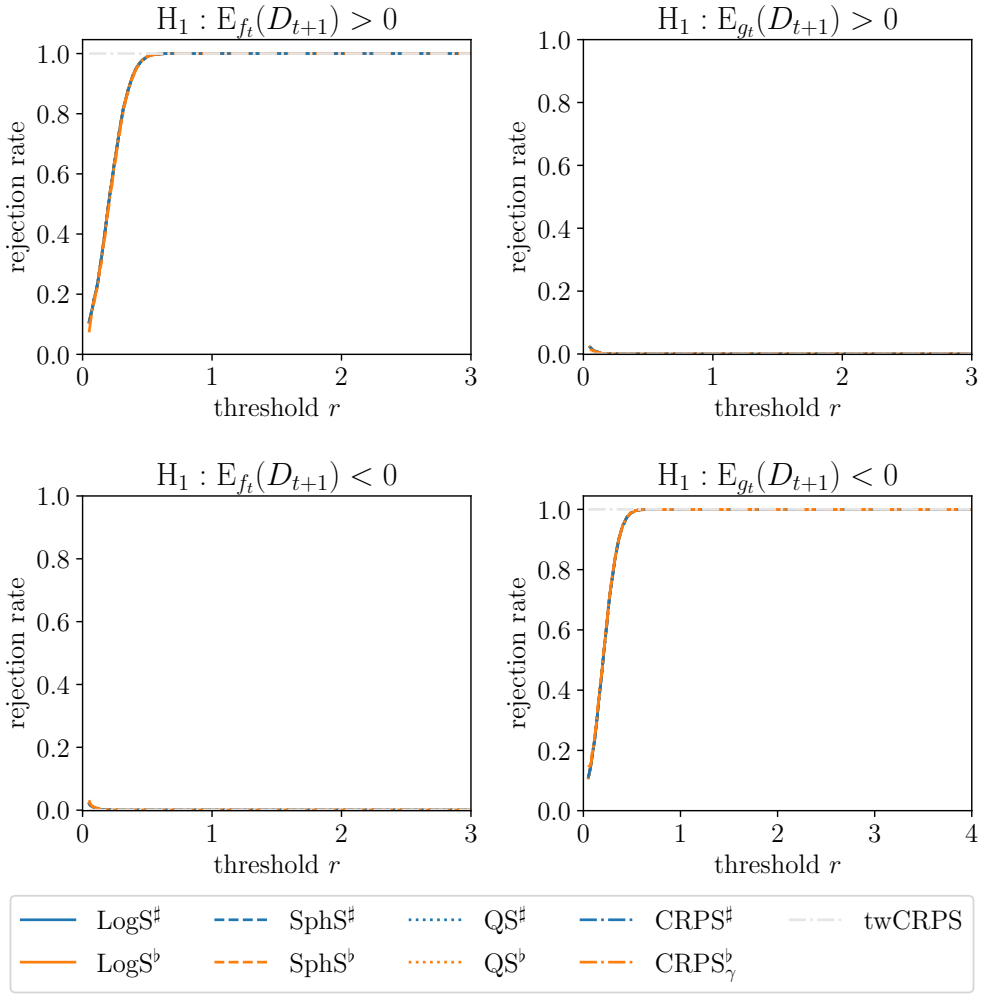
Figure 8: $\mathcal{N}(-0.2, 1)$ versus $\mathcal{N}(-0.2, 1)$: Centre ($c = 200$)



One-sided rejection rates of the GW test of equal predictive ability of the candidates f_t ($\mathcal{N}(-0.2, 1)$) and g_t ($\mathcal{N}(0.2, 1)$) 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). Moreover, 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 $w(y) = 1_{[-r,r]}(y)$ and the number of expected observations in the region of interest is kept constant at $c = 200$.

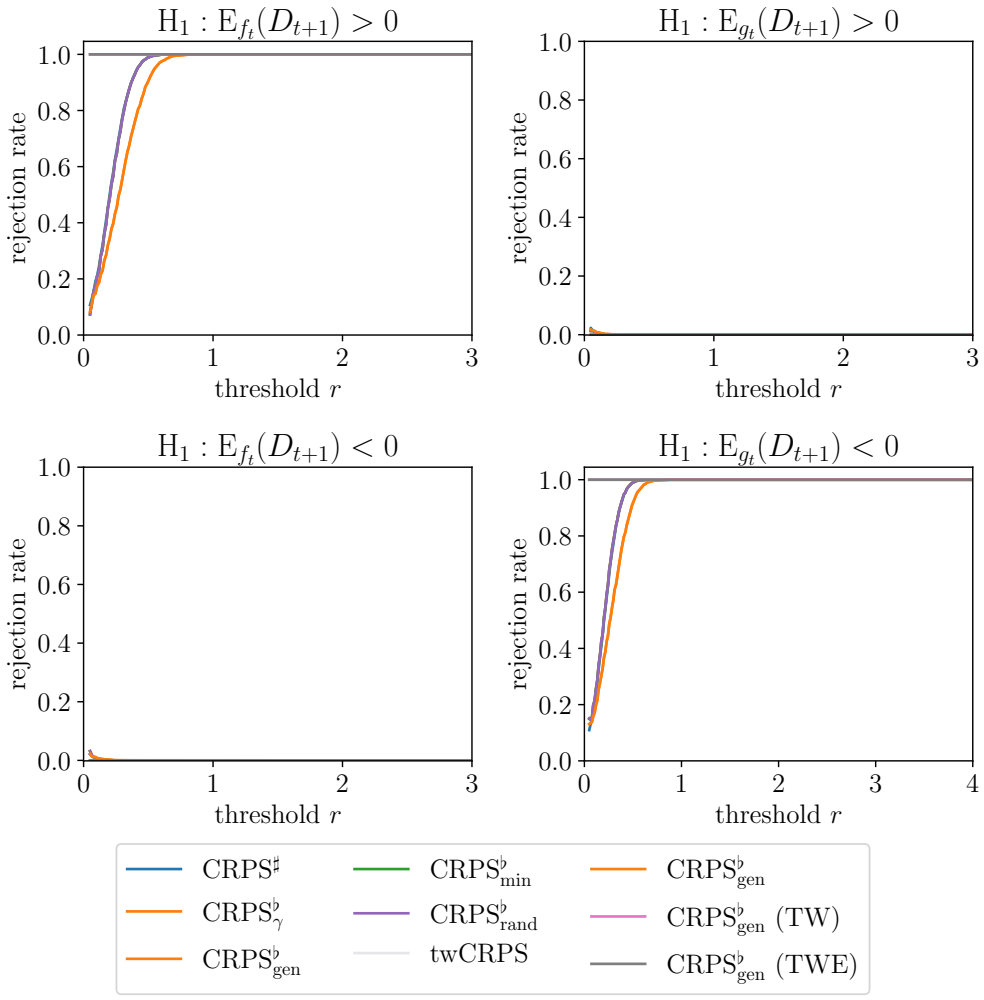
4.2.5 $\mathcal{N}(-1,1)$ versus $\mathcal{N}(-1,1)$: Centre ($c = 200$) [$\gamma_F \neq \gamma_P$]

Figure 9: $\mathcal{N}(-0.2,1)$ versus $\mathcal{N}(-0.2,1)$: Centre ($c = 200$)



One-sided rejection rates of the GW test of equal predictive ability of the candidates f_t ($\mathcal{N}(-1,1)$) and g_t ($\mathcal{N}(1,1)$) 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). Moreover, 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 $w(y) = 1_{[-r,r]}(y)$ and the number of expected observations in the region of interest is kept constant at $c = 200$.

Figure 10: $\mathcal{N}(-1,1)$ versus $\mathcal{N}(-1,1)$: Centre ($c = 200$)



One-sided rejection rates of the GW test of equal predictive ability of the candidates f_t ($\mathcal{N}(-1,1)$) and g_t ($\mathcal{N}(1,1)$) 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). Moreover, 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 $w(y) = 1_{[-r,r]}(y)$ and the number of expected observations in the region of interest is kept constant at $c = 200$.

5 Empirical Application

5.1 Risk management

Evaluating the downside risk of asset returns is a crucial task in risk management, particularly for meeting regulatory requirements related to risk measures like the Value-at-Risk ($\text{VaR}^q_{\hat{f}_t}$), which represents the q-th quantile of the model-based estimated density forecast $\hat{f}_t$ and the more recently required Expected Shortfall $\text{ES}^q_{\hat{f}_t}$, being the expected loss conditional on the the loss being below its $\text{VaR}^q_{\hat{f}_t}$. To fulfil this objective, we opt for a weight function of $w_t(y_t) = \mathbb{1}_{\left(-\infty, r_t^q\right]}(y_t)$, and choose for the variable of interest y_t 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. The dataset used for this study consists of 6,777 observations in total, spanning from January 2, 1996, to December 30, 2022, and obtained from Yahoo Finance.

Each of our selected forecast methods can be represented as

$$Y_t|\mathcal{F}_{t-1} \sim \mathcal{D}(\mu, \sigma_t^2, \boldsymbol{\vartheta}),$$

where $\mathcal{D}(\mu, \sigma_t^2, \boldsymbol{\vartheta})$ denotes a parametric family of distributions with mean μ, variance σ_t^2 and other parameters $\boldsymbol{\vartheta}$. We have also considered AR(1) and AR(5) models for the specification of the conditional mean, but did not find substantial improvements relative to the constant mean specification. In our analysis, we consider two conditional variance models: the GARCH(1,1) model proposed by Bollerslev (1987) and defined by the equation

$$\sigma_t^2 = \omega + \alpha(y_t - \mu)^2 + \beta\sigma_{t-1}^2,$$

and the RGARCH(1,1) model proposed by Hansen et al. (2012), which is given by

$$\sigma_t^2 = \omega + \alpha x_{t-1} + \beta \sigma_{t-1}^2,$$
$$x_t = \xi + \phi \sigma_t^2 + \tau z_t + \kappa(z_t^2 - 1) + u_t,$$

where x_t represents the realised measure, $z_t = (y_t - \mu)/\sigma_t$ and u_t denotes a white noise process with variance σ_u^2. The realised measure is downloaded from the Risklab page of Dacheng Xiu's website: `https://dachxiu.chicagobooth.edu/#risklab`. Notably, the equation for the realised measure is necessary for predictions beyond one step ahead, and the term $\tau z_t + \kappa(z_t^2 - 1)$ captures the leverage effect. Furthermore, we use either a normal distribution or a Student-t distribution with ν degrees of freedom. Although the heavy tails of the Student-t distribution have proven useful for daily stock return data, the normal distribution remains a common choice. We estimate the parameters using full-fledged maximum likelihood estimation, based on a rolling estimation window of length T_{est}.

5.1.1 Statistical comparison

As an empirical extension of our power analysis from Chapter 4, we compare the relative performance of the forecasting methods via the Model Confidence Set (MCS), as described by Hansen et al. (2011). The MCS procedure expands the GW hypothesis to larger sets of H_0-equivalent methods, employing an iterative elimination procedure based on the equivalence tests Tmax_k or TR_k, with k indicating the block length of the block bootstrap required for estimating the non-standard asymptotic distribution of these test statistics. Optimal power properties of censoring in the GW environment intuitively accelerate elimination in the MCS procedure, resulting in smaller MCS p-values and consequently, lower MCS cardinality.

Table 2 delineates the MCS p-values and the deduced $\text{MCS}_{0.90}$ and $\text{MCS}_{0.75}$ for the six previously mentioned forecast methods, based on their 1-step and 5-step ahead density forecasts. We present the outcomes for $q \in \{0.01, 0.1, 0.2\}$, using the TR_{20}

statistic with a block bootstrap size of $B = 10,000$ and an estimation window of $T_{\text{est}} = 1,000$, and later validate the stability of these findings. When examining the censored ($\flat$) and conditional ($\sharp$) columns, the reported results overwhelmingly support enhanced power via censoring. The cardinality of the censored MCS never surpasses its conditional counterpart, the censored MCS often being a (strict) subset of the conditional MCS. Notable differences, especially for $\text{MCS}_{0.75}$ of QS or CRPS at $q = 0.10$, underline the overall relative increase in cardinality of $\text{MCS}_{0.90}$ ($\text{MCS}_{0.75}$) by 78% (75%) when opting for conditioning over censoring. The significance of these reductions is further emphasised by the fact that the resulting MCS encompass more complex model specifications, which would be the optimal choices in the absence of parameter and forecasting uncertainty.

For $h = 5$, reductions through censoring also occur, albeit less frequently. Excluding the CRPS rule, the table suggests that censoring leads to a smaller $\text{MCS}_{0.90}$ ($\text{MCS}_{0.75}$) 3.0 (1.7) times more often. With the CRPS inclusion, this frequency equilibrates due to the stronger eliminative capacity of the conditional CRPS. Again, the differences can become quite sizeable, here on average leading to a relative increase of cardinality of 31% (41%), or even 50% (62%) excluding the CRPS, when using conditioning instead.

To ascertain the robustness of our findings, we expand our study across different estimation window lengths T_{est}, block lengths k, and the Tmax statistics for a more comprehensive set of quantiles, also including $q = 0.05, 0.15$ and 0.25. Table 3 highlights the consistency of these results, revealing only one occurrence of increased $\text{MCS}^{\flat}$-cardinality at both confidence levels and even no exception for the average relative increase of the size of from $\text{MCS}^{\flat}$ to $\text{MCS}^{\sharp}$. These findings align with reported differences between forecasting horizons in Table 2, more generally influenced by the relatively strong performance of the $\text{CRPS}^{\sharp}$ rule on $h = 5$ outcomes. Furthermore, the first row of Table 3 extends the results from Table 2 to all quantiles, showing that censoring typically reduces MCS cardinality 71% of the time and only infrequently, 4% of the time, increases it. The TR statistic typically intensifies elimination, bearing smaller MCS p-values, translating into movements outside the $\text{MCS}_{0.90}$ or $\text{MCS}_{0.75}$ for

the censored scoring rules, where MCS p-values were typically closer to the boundaries. Differences in outcomes due to varying block lengths ($k = 20$ and $k = 100$) are minor, and the influence of the estimation window lengths on the MCS p-values is not substantial.

5.1.2 Backtesting

Beyond the statistical assessment of forecast methods, we compute their 1- and 5-step ahead Value at Risk ($\mathrm{VaR}^q_{\hat{f}_t}$) and Expected Shortfall ($\mathrm{ES}^q_{\hat{f}_t}$). These measures provide only partial insight into the forecasts, since the tail component of the density forecast carries more comprehensive information than a single quantile ($\mathrm{VaR}^q_{\hat{f}_t}$) or conditional moment $\mathrm{ES}^q_{\hat{f}_t} = \mathbb{E}_{\hat{f}_t}\left(Y_{t+h} | Y_{t+h} \leq \mathrm{VaR}^q_{\hat{f}_t}\right)$. Notably, the conditioning in $\mathrm{ES}^q_{\hat{f}_t}$ is a quantile of the density forecast itself rather than $\hat{r}^q_t$, a.s. implying a discrepancy between the operational region of $\mathrm{ES}^q_{\hat{f}_t}$ and the focused scoring rules introduced above. Additionally, if the $\mathrm{VaR}^q_{\hat{f}_t}$ is quite off, then the 'risk' indicated by $\mathrm{ES}^q_{\hat{f}_t}$ can become quite detached from the true risk ES^q_p, where p denotes the density of the DGP. Hence, the $\mathrm{ES}^q_{\hat{f}_t}$ is (particularly) useful when the $\mathrm{VaR}^q_{\hat{f}_t}$ is accurate, i.e. we preferably have a good fit for the pair $(\mathrm{VaR}^q_{\hat{f}_t}, \mathrm{ES}^q_{\hat{f}_t})$, rather than just $\mathrm{ES}^q_{\hat{f}_t}$ itself.

We highlight a corollary before discussing results. Given a fixed level q, let r be such that $\mathrm{VaR}^q_{\hat{f}_t} \vee \mathrm{VaR}p^q \leq r$. A property of the censored scoring rule is its ability to render the true $(\mathrm{VaR}^q_p, \mathrm{ES}^q_p)$ pair, since

$$\mathbb{D}_{S^{\flat}_w}(p\|f) = 0 \implies (\mathrm{VaR}^q_p, \mathrm{ES}^q_p) = (\mathrm{VaR}^q_{\hat{f}_t}, \mathrm{ES}^q_{\hat{f}_t}), \tag{8}$$

where $w(y) = \mathbb{1}_{(-\infty,r)}(y)$. This is a direct consequence of (3), i.e. another corollary of Lemma 2, and holds also more generally for any functional on distributions on $\{w > 0\}$. In (sharp) contrast, $\mathbb{D}_{S^{\sharp}_w}(p\|f) = 0$ implies that $p \propto f$ on $(-\infty, r)$ and hence $(\mathrm{VaR}^q_p, \mathrm{ES}^q_p) \neq (\mathrm{VaR}^q_{\hat{f}_t}, \mathrm{ES}^q_{\hat{f}_t})$, unless $\bar{F}_w = \bar{P}_w$. Therefore, model selection based on censored scoring rules aligns more effectively with backtesting of functionals of the distribution compared to model selection based on conditional scoring rules.

Table 2 reports the backtesting results, demonstrating a consistent preference for Student-t_ν models by Expected Shortfall (ES) measures, with VaR's top models fluctuating between Student-t_ν and Normal. Interestingly, at $q = 0.10$ for $h = 1$, there's a significant discrepancy in methods preferred by the risk measures. The MCS accommodates this by retaining at least one model from both camps, the censored MCS favouring accurate $\text{VaR}^q_{\hat{f}_t}$ coverage slightly more than conditional ones. Table 3 supports this, revealing that censoring typically leads to greater alignment with the scoring rule in the VaR column, while conditioning excels in the ES column. Notably, these outcomes are often reached with smaller MCS under censoring. Hence, the small percentages of mismatches particularly underscore the usefulness of censoring, as it frequently identifies a compact set of models yielding reliable risk measures. For $h = 5$, the percentages are generally smaller. However, this is not because forecasting risk measures further into the future is easier. On the contrary, the MCS are less reduced for $h = 5$, likely due to the greater difficulty in distinguishing between models as a result of increased forecasting noise.

As discussed earlier, it is sensible to examine the pair $(\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t})$ rather than $\text{ES}^q_{\hat{f}_t}$ in isolation. Moreover, Equation (8) suggests that censoring may generate Model Confidence Sets (MCS) containing forecast models that produce $(\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t})$ pairs closer to the true pair. Support for this conjecture is found in Table 3. Despite often being smaller, the censored MCS contains well-fitted $(\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t})$ pairs, defined as 0% mismatches for both VaR and ES, more than twice as often (9 versus 4). If we accept up to 4% mismatches, the comparison remains favourable: 14 versus 6, endorsing censored MCS as a superior selection mechanism prior to VaR and ES calculations.

Table 2: Evaluating forecast methods in the left-tail

		LogS		QS		SphS		CRPS		Backtesting	
q	Method	♭	♯	♭	♯	♭	♯	♭	♯	**hitrate**	**ES**
						$h = 1$					
0.01	RGARCH-t_ν	**1.00**	**0.60**	**0.45**	**0.95**	**0.65**	**0.88**	**0.73**	**0.97**	**0.016**	**0.002***
	TGARCH-t_ν	**0.99**	**1.00**	**0.63**	**1.00**	**0.88**	**1.00**	**0.81**	**1.00**	**0.014**	**0.008***
	GARCH-t_ν	**0.53**	**0.69**	**0.34**	**0.84**	**0.65**	**0.88**	**0.81**	**0.91**	**0.015**	**0.044***
	RGARCH-$\mathcal{N}$	0.09	0.19	**1.00**	**0.95**	**1.00**	**0.66**	**0.81**	**0.97**	0.020	0.221
	TGARCH-$\mathcal{N}$	0.03	0.09	**0.63**	**0.95**	**0.88**	**0.38**	**1.00**	**0.97**	0.021	0.257
	GARCH-$\mathcal{N}$	0.01	0.09	**0.45**	**0.84**	**0.65**	**0.45**	**0.81**	**0.64**	0.023	0.294
0.10	RGARCH-t_ν	**1.00**	**0.73**	**1.00**	**0.70**	**0.35**	**0.88**	0.12	**0.74**	0.115	**0.074***
	TGARCH-t_ν	0.05	**1.00**	0.16	**1.00**	0.02	**1.00**	0.05	**1.00**	0.113	**0.045***
	GARCH-t_ν	0.00	**0.40**	0.01	**0.43**	0.00	**0.37**	0.01	0.20	0.118	**0.048***
	RGARCH-$\mathcal{N}$	0.05	0.03	**0.49**	**0.70**	**1.00**	0.21	**1.00**	**0.74**	**0.103***	0.160
	TGARCH-$\mathcal{N}$	0.00	0.01	0.06	**0.39**	0.00	0.03	0.12	0.04	**0.100***	0.168
	GARCH-$\mathcal{N}$	0.00	0.00	0.00	0.06	0.00	0.01	0.03	0.01	**0.103***	0.206
0.20	RGARCH-t_ν	**1.00**	**1.00**	**1.00**	**0.40**	**1.00**	**0.26**	0.10	**0.42**	**0.203***	**0.111**
	TGARCH-t_ν	0.02	0.23	0.10	0.06	0.02	0.02	0.01	0.14	**0.204***	**0.076**
	GARCH-t_ν	0.00	0.06	0.06	0.00	0.00	0.00	0.00	0.01	**0.210***	**0.084**
	RGARCH-$\mathcal{N}$	0.01	0.04	0.00	**1.00**	0.02	**1.00**	**1.00**	**1.00**	0.183	0.144
	TGARCH-$\mathcal{N}$	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.09	0.180	0.134
	GARCH-$\mathcal{N}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.186	0.152
						$h = 5$					
0.01	RGARCH-t_ν	**0.37**	**0.87**	0.12	**1.00**	0.24	**1.00**	**0.57**	**1.00**	**0.015**	**0.227**
	TGARCH-t_ν	**0.83**	**1.00**	**0.86**	**0.45**	**1.00**	**0.40**	**0.65**	**0.63**	**0.017**	**0.128***
	GARCH-t_ν	**1.00**	**0.96**	0.17	**0.38**	**0.47**	**0.40**	**0.65**	**0.44**	**0.017**	**0.054***
	RGARCH-$\mathcal{N}$	0.01	0.05	0.12	**0.81**	0.18	**0.40**	**0.57**	0.18	0.019	0.521
	TGARCH-$\mathcal{N}$	0.01	0.05	**1.00**	**0.75**	**1.00**	0.23	**1.00**	0.16	0.022	0.461
	GARCH-$\mathcal{N}$	0.01	0.04	0.17	**0.75**	**0.41**	**0.27**	**0.65**	0.09	0.022	0.480
0.10	RGARCH-t_ν	**0.46**	0.15	**0.35**	0.16	**0.55**	0.11	**0.47**	**0.43**	**0.100***	**0.156**
	TGARCH-t_ν	**1.00**	**0.36**	**1.00**	**1.00**	**1.00**	**0.41**	**1.00**	**1.00**	0.108	**0.096**
	GARCH-t_ν	**0.56**	**1.00**	**0.35**	**0.70**	**0.55**	**1.00**	**0.60**	**0.43**	0.113	**0.075***
	RGARCH-$\mathcal{N}$	0.00	0.00	0.01	0.16	**0.26**	0.03	**0.53**	0.01	0.090	0.242
	TGARCH-$\mathcal{N}$	0.00	0.00	0.01	0.16	0.21	0.00	**0.60**	0.00	**0.095***	0.242
	GARCH-$\mathcal{N}$	0.00	0.00	0.01	0.16	0.20	0.00	**0.46**	0.00	**0.099***	0.247
0.20	RGARCH-t_ν	0.15	**0.36**	0.00	**0.89**	0.02	**0.97**	**0.37**	**0.67**	0.175	**0.155**
	TGARCH-t_ν	**1.00**	**0.36**	**0.77**	**1.00**	**1.00**	**0.97**	**1.00**	**1.00**	**0.192***	**0.114**
	GARCH-t_ν	**0.78**	**1.00**	**1.00**	**0.89**	**0.50**	**0.97**	**0.37**	**0.67**	**0.201***	**0.086***
	RGARCH-$\mathcal{N}$	0.00	0.00	0.00	**0.89**	0.00	**1.00**	**0.32**	**0.67**	0.160	0.193
	TGARCH-$\mathcal{N}$	0.00	0.00	0.00	**0.79**	0.00	**0.90**	**0.37**	0.02	0.172	0.167
	GARCH-$\mathcal{N}$	0.00	0.00	0.00	**0.79**	0.00	**0.76**	0.23	0.01	**0.176**	0.185

Note: Columns titled by scoring rules present MCS p-values for assessing predictive equality of the forecast methods listed, h-steps ahead, based on LogS, QS, SphS and CRPS scoring rule variants, both censored (♭) and conditional (♯). The emphasis is on the left-tail, incorporated by weight function $w_t(y_t) = \mathbb{1}_{(-\infty, r_t^q]}(y_t)$, for various q values. Bold (underlined) p-values signify the forecast method's inclusion in $\text{MCS}_{0.75}$ ($\text{MCS}_{0.90}$). All MCS p-values utilise the TR20 statistic, implementing $B = 10,000$ block bootstrap simulations with blocklength $k = 20$. The estimation window is $T_{\text{est}} = 1,000$. The backtesting columns present the hit rate, instances beneath a method's VaR at level q, and the absolute difference between actual and density forecast-implied expected shortfall at level q. The top three models are bolded; an asterisk denotes non-rejection of null for correct coverage (hit rate) or difference (ES) at a 0.05 significance level.

Table 3: Overview and robustness of left-tail application

		$\text{MCS}_{0.90}$							$\text{MCS}_{0.75}$						
		|MCS|			VaR		ES		**|MCS|**			VaR		ES	
T_{est}	**Stat.**	$<$	$>$	%	$\flat$	$\sharp$	$\flat$	$\sharp$	$<$	$>$	%	$\flat$	$\sharp$	$\flat$	$\sharp$
							$h=1$								
1000	TR_{20}	**71**	4	128	17	17	17	**8**	**63**	8	104	**17**	21	25	**13**
	TR_{100}	**71**	4	130	17	17	17	**8**	**63**	8	104	**17**	17	17	**0**
	Tmax_{20}	**46**	13	58	**0**	13	4	4	**71**	8	107	**4**	13	17	**4**
	Tmax_{100}	**46**	13	48	**0**	13	4	4	**67**	13	102	**4**	13	17	**4**
750	TR_{20}	**54**	13	81	**13**	25	13	13	**50**	13	80	**13**	29	29	**17**
	Tmax_{20}	**33**	29	19	**0**	17	4	4	**63**	17	100	**4**	17	21	**8**
1250	TR_{100}	**63**	8	122	17	17	17	**0**	**50**	4	99	17	17	21	**17**
	Tmax_{20}	**58**	17	74	8	8	13	**0**	**50**	17	83	**8**	13	21	**17**
							$h=5$								
1000	TR_{20}	**38**	25	69	17	**0**	4	**0**	**50**	42	72	**17**	21	4	**0**
	TR_{100}	**38**	25	69	17	**0**	4	**0**	**50**	42	66	**17**	21	4	**0**
	Tmax_{20}	**42**	25	43	0	0	0	0	**50**	42	59	**0**	17	0	0
	Tmax_{100}	**38**	17	44	0	0	0	0	**46**	38	59	**0**	17	0	0
750	TR_{100}	33	**46**	44	**0**	8	0	0	33	**50**	54	**0**	25	0	0
	Tmax_{20}	**29**	25	53	**0**	4	0	0	**42**	38	53	**0**	17	0	0
1250	TR_{20}	**38**	33	61	4	4	0	0	**50**	33	61	**8**	25	0	0
	Tmax_{20}	**42**	29	46	**0**	4	0	0	**46**	42	46	**4**	17	0	0

Note: The table summarises MCS and backtesting results using varying values for estimation window T_{est}, equivalence test statistics TR_k and Tmax_k, blocklength k across forecast horizons $h=1$ and $h=5$, based on $B=10,000$ bootstrap replications. All values are percentages. Columns labelled with |MCS| refer to MCS cardinality. Across 24 combinations of $q \in \{0.01, 0.05, 0.10, 0.15, 0.20, 0.25\}$ and $S \in \{\text{LogS}, \text{QS}, \text{SphS}, \text{CRPS}\}$, the $<$ ($>$)-column displays the frequency of $|\text{MCS}^{\flat}_{1-\alpha}| < (>)|\text{MCS}^{\sharp}_{1-\alpha}|$ and the %-column indicates average relative cardinality increase from $\text{MCS}^{\sharp}_{1-\alpha}$ to $\text{MCS}^{\flat}_{1-\alpha}$. The VaR (ES) column shows the frequency of the $\text{MCS}_{1-\alpha}$ containing one of the top three models based on VaR (ES) backtesting results. Bolded numbers indicate strictly more smaller ($<$) or larger ($>$) $\text{MCS}^{\flat}_{1-\alpha}$ as well as strictly less times the $\text{MCS}_{1-\alpha}$ contains a top 3 VaR or ES method.

6 Conclusion

In many applications, forecasters are not equally interested in all possible outcomes of the random variable of interest. For such cases, we have motivated the use of censoring as focusing device. In particular, we have shown that focusing scoring rules by applying them to censored distributions leads to strictly locally proper scoring rules. To the best of our knowledge, we are first in deriving a transformation of the original scoring rule that preserves strict propriety.

Our approach offers considerable flexibility in terms of the original scoring rule, the weight function, and the outcome space. For specific choices, the (generalised) censored scoring rule delivers intuitively sound scoring rules that can easily be implemented by practitioners. When applied to the logarithmic scoring rule, our focusing technique produces the well-established censored likelihood score. Our framework also facilitates the derivation of weighted versions of the Energy Score family, which can be considered the multivariate equivalent of the twCRPS for left- or right-tail indicator functions. For other weight functions, the censored CRPS rule is strictly locally proper, while the twCRPS is not. By uncovering a procedure close to censoring, we established to clarify the localisation bias of the twCRPS and how one can derive a multivariate version of the twCRPS for the centre indicator function.

The censored likelihood score also appears in a second important result of this paper. In particular, we have shown that the UMP test for a localised version of the standard simple versus simple Neyman Pearson testing problem is based on the censored likelihood ratio. Furthermore, the results of our Monte Carlo study suggest that our theoretical findings spill over to the finite sample properties of other forecast evaluation tests. In our experiments, striking differences in power almost always favour censoring. In the empirical application, we identified an enhanced elimination of poorly performing models during the MCS process when using censoring instead of conditioning. Moreover, models that resulted in superior VaR and ES backtesting outcomes were more frequently incorporated in the MCS process when based on censoring.

Appendix

A Proofs

A.1 Proof Theorem 2

For clarity of exposition, we first prove the main ingredients of the proof via two isolated lemmas and a corollary.

Lemma 2. *Consider the censored scoring rule defined in Definition 6. $\forall w \in \mathcal{W}$ and $\mathrm{H} \in \mathscr{H}$, the following identity holds $\int_{\mathcal{Y}} S^{\flat}_{w,\mathrm{H}}(\mathrm{F}, y)\mathrm{P}(\mathrm{d}y) = \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w,\mathrm{H}}(\mathrm{d}y)$.*

Proof.

$$
\begin{aligned}
\int_{\mathcal{Y}} S^{\flat}_{w,\mathrm{H}}(\mathrm{F}, y)\mathrm{P}(\mathrm{d}y) &= \int_{\mathcal{Y}} \Big(w(y) S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y) + \big(1 - w(y)\big) \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, q)\mathrm{H}(\mathrm{d}q) \Big) \mathrm{P}(\mathrm{d}y), \\
&= \int_{\mathcal{Y}} w(y) S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}(\mathrm{d}y) + \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, q) \int_{\mathcal{Y}} \big(1 - w(y)\big)\mathrm{P}(\mathrm{d}y)\mathrm{H}(\mathrm{d}q), \\
&= \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}_w(\mathrm{d}y) + \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\bar{P}_w\mathrm{H}(\mathrm{d}y), \\
&= \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\big(\mathrm{P}_w(\mathrm{d}y) + \bar{P}_w\mathrm{H}(\mathrm{d}y)\big), \\
&= \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w,\mathrm{H}}(\mathrm{d}y).
\end{aligned}
$$

□

Lemma 3. *Consider two distributions* P *and* F *on the same measurable space $(\mathcal{Y}, \mathcal{G})$. On the same space, let their censored counterparts $\mathrm{P}^{\flat}_{w,\mathrm{H}}$ and $\mathrm{F}^{\flat}_{w,\mathrm{H}}$ be given by Definition 6. Then,*

$$
\mathrm{F}^{\flat}_{w,\mathrm{H}}(E) = \mathrm{G}^{\flat}_{w,\mathrm{H}}(E),\ \forall E \in \mathcal{G} \iff \mathrm{F}(E \cap \{w > 0\}) = \mathrm{G}(E \cap \{w > 0\}),\ \forall E \in \mathcal{G}.
$$

Proof. " $\Longrightarrow$ " We start with the most challenging direction, for which Assumption 1

is of critical importance. First, note that

$$
\begin{aligned}
&\mathrm{F}^{\flat}_{w,\mathrm{H}}(E) = \mathrm{G}^{\flat}_{w,\mathrm{H}}(E), \qquad \forall E \in \mathcal{G} \\
&\implies \mathrm{F}^{\flat}_{w,\mathrm{H}}(E \cap \{w = c\}) = \mathrm{G}^{\flat}_{w,\mathrm{H}}(E \cap \{w = c\}), \qquad \forall E \in \mathcal{G} \\
&\implies \int_{\mathcal{Y}} (1-w)\mathrm{dFH}(E \cap \{w = c\}) = \int_{\mathcal{Y}} (1-w)\mathrm{dGH}(E \cap \{w = c\}), \qquad \forall E \in \mathcal{G} \\
&\implies \int_{\mathcal{Y}} (1-w)\mathrm{dFH}(\{w = c\}) = \int_{\mathcal{Y}} (1-w)\mathrm{dGH}(\{w = c\}), \\
&\implies \int_{\mathcal{Y}} (1-w)\mathrm{dF} = \int_{\mathcal{Y}} (1-w)\mathrm{dG},
\end{aligned}
$$

where c denotes a constant such that Assumption 1 is satisfied. Then, exploit this equality to conclude

$$
\begin{aligned}
&\mathrm{F}^{\flat}_{w,\mathrm{H}}(E) = \mathrm{G}^{\flat}_{w,\mathrm{H}}(E), \qquad \forall E \in \mathcal{G} \\
&\implies \int_{\mathcal{Y}} w(y)\mathbb{1}_{y \in E}\mathrm{F}(\mathrm{d}y) = \int_{\mathcal{Y}} w(y)\mathbb{1}_{y \in E}\mathrm{G}(\mathrm{d}y), \qquad \forall E \in \mathcal{G} \\
&\implies \mathrm{F}(E \cap \{w > 0\}) = \mathrm{G}(E \cap \{w > 0\}), \qquad \forall E \in \mathcal{G}.
\end{aligned}
$$

" $\Longleftarrow$ " The other direction is somewhat trivial. Indeed,

$$
\begin{aligned}
&\mathrm{F}(E \cap \{w > 0\}) = \mathrm{G}(E \cap \{w > 0\}), \qquad \forall E \in \mathcal{G} \\
&\implies \int_{\mathcal{Y}} w(y)\mathbb{1}_{y \in E}\mathrm{F}(\mathrm{d}y) = \int_{\mathcal{Y}} w(y)\mathbb{1}_{y \in E}\mathrm{G}(\mathrm{d}y), \qquad \forall E \in \mathcal{G} \\
&\implies \int_{\mathcal{Y}} (1-w)\mathrm{dF} = \int_{\mathcal{Y}} (1-w)\mathrm{dG},
\end{aligned}
$$

and the two implied results jointly imply $\mathrm{F}^{\flat}_{w,\mathrm{H}}(E) = \mathrm{G}^{\flat}_{w,\mathrm{H}}(E),\ \forall E \in \mathcal{G}, \forall \mathrm{H} \in \mathscr{H}$. □

Corollary 4. *The censored scoring rule defined in Definition 6 is localising* $\forall \mathrm{H} \in \mathscr{H}$.

Proof. Suppose that $\mathrm{F}(E \cap \{w > 0\}) = \mathrm{G}(E \cap \{w > 0\}),\ \forall E \in \mathcal{G}$. Then, by Lemma 3, $\mathrm{F}^{\flat}_{w,\mathrm{H}}(E) = \mathrm{G}^{\flat}_{w,\mathrm{H}}(E), \forall E \in \mathcal{G}$, whence it follows that $S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y) = S^{\flat}_{w,\mathrm{H}}(\mathrm{F}, y), \forall y \in \mathcal{Y}$. □

We now turn to the main body of the proof. The definition of a strictly locally

proper scoring rule (Definition 4) and the definitions on which this definition is built, that is, the definition of a locally proper scoring rule (Definition 4) and a localising weighted scoring rule (Definition 3), reveal that we need to prove a list of three things $\forall \mathrm{H} \in \mathscr{H}$: (i) $S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y)$ must be localising relative to $\mathcal{W}$, (ii) $S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y)$ must be proper relative to $\mathscr{P}$, $\forall w \in \mathcal{W}$ and (iii) the if and only if statement in Definition 4. We prove them one by one.

(i) $S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y)$ is localising relative to $\mathcal{W}$, $\forall \mathrm{H} \in \mathscr{H}$, by Corollary 4.

(ii) Fix an arbitrary $w \in \mathcal{W}$ and $\mathrm{H} \in \mathscr{H}$. Since $\mathscr{P}^{\flat}_{w,\mathrm{H}} \subseteq \mathscr{P}^{\flat}$, S is strictly proper relative to $\mathscr{P}^{\flat}_{w,\mathrm{H}}$, i.e.

$$\int_{\mathcal{Y}} S(\mathrm{P}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w,\mathrm{H}}(\mathrm{d}y) \geq \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w,\mathrm{H}}(\mathrm{d}y), \quad \forall \mathrm{P}^{\flat}_{w,\mathrm{H}}, \mathrm{F}^{\flat}_{w,\mathrm{H}} \in \mathscr{P}^{\flat}_{w,\mathrm{H}}, \tag{A.1}$$

which is by definition of the class $\mathscr{P}^{\flat}_{w,\mathrm{H}} \equiv \{[\mathrm{P}]^{\flat}_{w,\mathrm{H}}, \mathrm{P} \in \mathscr{P}\}$ equivalent to

$$\int_{\mathcal{Y}} S([\mathrm{P}]^{\flat}_{w,\mathrm{H}}, y)[\mathrm{P}]^{\flat}_{w,\mathrm{H}}(\mathrm{d}y) \geq \int_{\mathcal{Y}} S([\mathrm{F}]^{\flat}_{w,\mathrm{H}}, y)[\mathrm{P}]^{\flat}_{w,\mathrm{H}}(\mathrm{d}y), \quad \forall \mathrm{P}, \mathrm{F} \in \mathscr{P}, \tag{A.2}$$

and hence, by Lemma 2, also

$$\int_{\mathcal{Y}} S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y)\mathrm{P}(\mathrm{d}y) \geq \int_{\mathcal{Y}} S^{\flat}_{w,\mathrm{H}}(\mathrm{F}, y)\mathrm{P}(\mathrm{d}y), \quad \forall \mathrm{P}, \mathrm{F} \in \mathscr{P}. \tag{A.3}$$

Therefore, $S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y)$ is proper relative to $\mathscr{P}$ by Definition 2.

(iii) Since S is strictly proper relative to $\mathscr{P}^{\flat}$ and hence $\mathscr{P}^{\flat}_{w,\mathrm{H}}$, it also follows that, $\forall w \in \mathcal{W}$ and $\mathrm{H} \in \mathscr{H}$,

$$\int_{\mathcal{Y}} S(\mathrm{P}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w,\mathrm{H}}(\mathrm{d}y) = \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w,\mathrm{H}}(\mathrm{d}y) \iff \mathrm{P}^{\flat}_{w,\mathrm{H}} = \mathrm{F}^{\flat}_{w,\mathrm{H}},$$

and thus, by Lemma 3,

$$\int_{\mathcal{Y}} S(\mathrm{P}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w}(\mathrm{d}y) = \int_{\mathcal{Y}} S(\mathrm{F}^{\flat}_{w,\mathrm{H}}, y)\mathrm{P}^{\flat}_{w,\mathrm{H}}(\mathrm{d}y) \iff \mathrm{P}(E \cap \{w > 0\}) = \mathrm{F}(E \cap \{w > 0\}),$$

$\forall E \in \mathcal{G}$, and hence, by Lemma 2, also

$$\int_{\mathcal{Y}} S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y)\mathrm{P}(\mathrm{d}y) = \int_{\mathcal{Y}} S^{\flat}_{w,\mathrm{H}}(\mathrm{F}, y)\mathrm{P}(\mathrm{d}y) \iff \mathrm{P}(E \cap \{w > 0\}) = \mathrm{F}(E \cap \{w > 0\}),$$

which is the desired if and only if statement of Definition 4.

But then, as we have verified each of the listed conditions (i) to (iii), we have shown that $S^{\flat}_{w,\mathrm{H}}(\mathrm{P}, y)$ is strictly locally proper relative to $(\mathscr{P}, \mathcal{W})$, $\forall \mathrm{H} \in \mathscr{H}$. □

A.2 Proof Lemma 1

Due to the integral over $\mathcal{Y}(\mathcal{I}_{A^c})$, any test ψ_{h_1} is constant in arguments varying in $\mathcal{Y}(\mathcal{I}_{A^c})$. We can use this observation to simplify the size of a test ψ_{h_1}. In particular, $\forall h_1 \in \mathscr{h}$, we have that

$$\begin{aligned}
\sup_{p_0 \in \mathscr{p}_0} \mathbb{E}_{p_0}\psi_{h_1} &= \left(\prod_{t\in\mathcal{I}_{A^c}} \mathrm{F}_0(A_t^c)\right) \sup_{h_0\in\mathscr{h}} \int_{\mathcal{Y}^T} \psi_{h_1} \prod_{t\in\mathcal{I}_A} f_{0t}\mathbb{1}_{A_t} \prod_{t\in\mathcal{I}_{A^c}} [h_{0t}]^{\sharp}_{A_t^c}\mathbb{1}_{A_t^c}\mathrm{d}\mu_t \\
&= \left(\prod_{t\in\mathcal{I}_{A^c}} \mathrm{F}_0(A_t^c)\right) \int_{\mathcal{Y}(\mathcal{I}_A)} \psi_{h_1} \prod_{t\in\mathcal{I}_A} f_{0t}\mathbb{1}_{A_t}\mathrm{d}\mu_t \\
&= \left(\prod_{t\in\mathcal{I}_{A^c}} \mathrm{F}_0(A_t^c)\right) \int_{\mathcal{Y}^T} \phi^*_{h_1} \prod_{t\in\mathcal{I}_{A^c}} [h_{1t}]^{\sharp}_{A_t^c}\mathbb{1}_{A_t^c}\mathrm{d}\mu_t \prod_{t\in\mathcal{I}_A} f_{0t}\mathbb{1}_{A_t}\mathrm{d}\mu_t \\
&\leq \left(\prod_{t\in\mathcal{I}_{A^c}} \mathrm{F}_0(A_t^c)\right) \sup_{h_0\in\mathscr{h}} \int_{\mathcal{Y}^T} \phi^*_{h_1} \prod_{t\in\mathcal{I}_{A^c}} [h_{0t}]^{\sharp}_{A_t^c}\mathbb{1}_{A_t^c}\mathrm{d}\mu_t \prod_{t\in\mathcal{I}_A} f_{0t}\mathbb{1}_{A_t}\mathrm{d}\mu_t \\
&= \sup_{p_0\in\mathscr{p}_0} \mathbb{E}_{p_0}\phi^*_{h_1} \\
&\leq \alpha,
\end{aligned}$$

since $\phi^*_{h_1} \in \Phi(\alpha)$. Hence, $\psi_{h_1} \in \Phi(\alpha)$. □

A.3 Proof Corollary 1

Fix an arbitrary $h_1 \in \mathscr{h}$. Since $\Psi(\alpha) \subseteq \Phi(\alpha)$, we trivially have that $\max_{\phi\in\Phi(\alpha)} \mathbb{E}_{p_1}\phi \geq \max_{\psi\in\Psi(\alpha)} \mathbb{E}_{p_1}\psi$. Now suppose that $\max_{\phi\in\Phi(\alpha)} \mathbb{E}_{p_1}\phi < \max_{\psi\in\Psi(\alpha)} \mathbb{E}_{p_1}\psi$. Then, we can always define the test $\tilde{\psi} = \int_{\mathcal{Y}(\mathcal{I}_{A^c})} \phi^* \prod_{t\in\mathcal{I}_{A^c}} [h_{1t}]^{\sharp}_{A_t^c}\mathbb{1}_{A_t^c}\mathrm{d}\mu_t$, with $\phi^* \in \arg\max_{\phi\in\Phi(\alpha)} \mathbb{E}_{p_1}\phi$,

satisfying $\mathbb{E}_{p_1}\phi^* = \mathbb{E}_{p_1}\tilde{\psi}$. But, by Lemma 1, $\tilde{\psi} \in \Psi(\alpha)$, in which case $\max_{\phi\in\Phi(\alpha)} \mathbb{E}_{p_1}\phi = \max_{\psi\in\Psi(\alpha)} \mathbb{E}_{p_1}\tilde{\psi}$, contradicting the assumed strict inequality. □

A.4 Proof Theorem 3

For any fixed $h_1 \in \mathscr{h}$, the most powerful test of size α is a solution to the following restricted maximisation problem

$$\begin{aligned}
\max_{\phi\in\Phi(\alpha)} \mathbb{E}_{p_1}\phi &= \max_{\boldsymbol{\alpha}\in\Delta_{\bar{T}}(\alpha)} \sum_{k=0}^{T}\sum_{s=1}^{\binom{T}{k}} \max_{\phi_{k,s}\in\Phi(\alpha_{k,s})} \mathbb{E}_{p_1}\left(\phi_{k,s}|y_t \in A_t, \forall i \in \mathcal{I}_A(k,s) \wedge y_t \in A_t^c, \forall i \in \mathcal{I}_{A^c}(k,s)\right) \\
&= \max_{\boldsymbol{\alpha}\in\Delta_{\bar{T}}(\alpha)} \sum_{k=0}^{T}\sum_{s=1}^{\binom{T}{k}} \max_{\phi_{k,s}\in\Psi(\alpha_{k,s})} \mathbb{E}_{p_1}\left(\phi_{k,s}|y_t \in A_t, \forall i \in \mathcal{I}_A(k,s) \wedge y_t \in A_t^c, \forall i \in \mathcal{I}_{A^c}(k,s)\right) \\
&= \max_{\boldsymbol{\alpha}\in\Delta_{\bar{T}}(\alpha)} \sum_{k=0}^{T}\sum_{s=1}^{\binom{T}{k}} \max_{\phi_{k,s}\in\Psi(\alpha_{k,s})} \left(\prod_{t\in\mathcal{I}_{A^c}} \mathrm{F}_1(A_t^c)\right) \int_{\mathcal{Y}(\mathcal{I}_A)} \phi_{k,s} \prod_{t\in\mathcal{I}_A} f_{1t}\mathbb{1}_{A_t}\mathrm{d}\mu_t \\
&= \max_{\boldsymbol{\alpha}\in\Delta_{\bar{T}}(\alpha)} \sum_{k=0}^{T}\sum_{s=1}^{\binom{T}{k}} \max_{\phi_{k,s}\in\Psi(\alpha_{k,s})} \int_{\mathcal{Y}^T} \phi_{k,s} \prod_{t=0}^{T-1} \mathrm{d}[\mathrm{F}_t]_{A_t}^{\flat} \\
&= \max_{\boldsymbol{\alpha}\in\Delta_{\bar{T}}(\alpha)} \sum_{k=0}^{T}\sum_{s=1}^{\binom{T}{k}} \max_{\phi_{k,s}\in\Phi(\alpha_{k,s})} \int_{\mathcal{Y}^T} \phi_{k,s} \prod_{t=0}^{T-1} \mathrm{d}[\mathrm{F}_t]_{A_t}^{\flat},
\end{aligned}$$

where $\bar{T} = \sum_{k=0}^{T}\binom{T}{k}$ and $\Delta_{\bar{T}}(\alpha_0) = \{\boldsymbol{\alpha}_0 \in [0,\alpha_0]^{\bar{T}} : \boldsymbol{\iota}_{\bar{T}}'\boldsymbol{\alpha}_0 = \alpha_0\}$, with $\boldsymbol{\iota}_{\bar{T}}$ denoting column vector of ones of length $\bar{T}$. The first equality exploits that the test function can be decomposed into test functions operating on a single part of the partitioning of the outcome space $\mathcal{Y}^T$, in which case the maximisation problem can be split into finding an optimal test on each of the partitioned parts conditional on the amount of size spent on each part and the optimal distribution of size over the partition of the outcome space.

The second equality holds by Corollary 1, the third equality uses that the optimal test is constant in arguments varying in A^c, the fourth equality holds by definition of the censored measure and the fifth equality uses that all tests that are non-constant in arguments varying in A^c map under the censored measure onto tests that are constant in arguments varying in A^c.

Finally, the result follows by observing that the final maximisation problem is equivalent to finding the optimal test $\phi_A^\flat$ for the testing problem $\mathbb{H}_j : p_j = \prod_{t=0}^{T-1} [f_j]_{A_t}^\flat$, $j \in \{0, 1\}$, for which $\phi_A^\flat$ is the UMP test by the Fundamental Lemma of Neyman and Pearson (1933). By the equivalence, $\phi_A^\flat$ is, for any $h_1 \in \mathscr{H}$, also the most powerful test for testing problem (6). But, since the test $\phi^\flat$ is independent of h_1, it is the UMP test for testing problem (6). □

References

Amisano, G. and R. Giacomini (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics 25*(2), 177–190.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, 1–45.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, 542–547.

Borowska, A., L. Hoogerheide, S. Koopman, and H. van Dijk (2020). Partially censored posterior for robust and efficient risk evaluation. *Journal of Econometrics 217*(2), 335–355.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics 7*(3), 200–217.

Brehmer, J. and T. Gneiting (2020). Properization: constructing proper scoring rules via bayes acts. *Annals of the Institute of Statistical Mathematics 72*(3), 659–673.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review 78*(1), 1–3.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General) 147*(2), 278–290.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics 59*(1), 77–93.

Diks, C. and H. Fang (2020). Comparing density forecasts in a risk management context. *International Journal of Forecasting 36*(2), 531–551.

Diks, C., V. Panchenko, O. Sokolinskiy, and D. van Dijk (2014). Comparing the accuracy of multivariate density forecasts in selected regions of the copula support. *Journal of Economic Dynamics and Control 48*, 79–94.

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

Eguchi, S. et al. (1985). A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima mathematical journal 15*(2), 341–391.

Ehm, W. and T. Gneiting (2012, FEB). Local proper scoring rules of order two. *Annals of Statistics 40*(1), 609–637.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222*(594-604), 309–368.

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

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

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

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society 14*, 107–114.

Good, I. (1971). Comment on "measuring information and uncertainty.". *Foundation of Statistical Inference*, 265–273.

Hansen, P. R., Z. Huang, and H. H. Shek (2012, SEP-OCT). Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics 27*(6, SI), 877–906.

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

Hersbach, H. (2000, OCT). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*(5), 559–570.

Holzmann, H. and B. Klar (2016). Weighted scoring rules and hypothesis testing. Available at arXiv:1611.07345v2.

Holzmann, H. and B. Klar (2017). Focusing on regions of interest in forecast evaluation. *Annals of applied statistics 11*(4), 2404–2431.

Iacopini, M., F. Ravazzolo, and L. Rossini (2022). Proper scoring rules for evaluating density forecasts with asymmetric loss functions. *Journal of Business & Economic Statistics*, 1–15.

Jose, V. R. (2009). A characterization for the spherical scoring rule. *Theory and Decision 66*(3), 263–281.

Jose, V. R. R., R. F. Nau, and R. L. Winkler (2008). Scoring rules, generalized entropy, and utility maximization. *Operations research 56*(5), 1146–1157.

Kole, E., T. Markwat, A. Opschoor, and D. Van Dijk (2017). Forecasting value-at-risk under temporal and portfolio aggregation. *Journal of Financial Econometrics 15*(4), 649–677.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The annals of mathematical statistics 22*(1), 79–86.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 106–127.

Liese, F. and I. Vajda (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory 52*(10), 4394–4412.

Matheson, J. and R. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*(10), 1087–1096.

Merton, R. C. (1971). Optimum consumption and portfolio rules in a continuous-time model. pp. 621–661. Elsevier.

Neyman, J. and E. S. Pearson (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231*(694-706), 289–337.

Opschoor, A., D. van Dijk, and M. van der Wel (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics 32*(7), 1298–1313.

Ovcharov, E. (2018). Proper scoring rules and bregman divergence. *Bernoulli 24*(1), 53–79.

Painsky, A. and G. W. Wornell (2019). Bregman divergence bounds and universality properties of the logarithmic loss. *IEEE Transactions on Information Theory 66*(3), 1658–1673.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics 38*(4), 796–809.

Roby, T. B. (1964). Belief states: A preliminary empirical study. Technical report, Tufts University Medford MA.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*(336), 783–801.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics 1*(1), 43–61.

Shuford, E. H., A. Albert, and H. E. Massengill (1966). Admissible probability measurement procedures. *Psychometrika 31*(2), 125–145.

Székely, G. J. and M. L. Rizzo (2005). A new test for multivariate normality. *Journal of Multivariate Analysis 93*(1), 58–80.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 24–36.

Toda, M. (1963). Measurement of subjective probability distribution. rep. no. 3, 1963, state college, pennsylvania. *Institute for Research*.

Supplementary Material: A General Procedure for Localising Strictly Proper Scoring Rules

Ramon de Punder* Cees Diks* Roger Laeven* Dick van Dijk†

June 23, 2023

Abstract

This document contains additional proofs, derivations and examples of the paper 'Localising Strictly Proper Scoring Rules'.

*University of Amsterdam, Tinbergen Institute
†Erasmus University Rotterdam, Tinbergen Institute

Contents

A Additional Proofs

A.1 Proof censored density

We defined the measures μ and F to the extended measurable space $(\mathcal{Y}^*, \mathcal{G}^*)$ by putting $\mu^*(E) = \mu(E \backslash \{*\})$, if $* \in E$ and $\mu^*(E) = \mu(E)$, otherwise, $\forall E \in \mathcal{G}^*$. To simplify the notation, we drop the subscript $*$ in the notation of the extended measures, while still considering all measures with respect to the extended measurable space $(\mathcal{Y}^*, \mathcal{G}^*)$

Since $(\mu + \delta_*)(E) = 0$ implies that both $\mu(E) = 0$ and $\delta_*(E) = 0$, $\forall E \in \mathcal{G}^*$, we have that both $\mu \ll \mu + \delta_*$ and $\delta_* \ll \mu + \delta_*$. As a consequence,

$$f^{\flat}_{w,h} := \frac{\mathrm{dF}^{\flat}_{w}}{\mathrm{d}(\mu + \delta_*)} = w \frac{\mathrm{dF}}{\mathrm{d}(\mu + \delta_*)} + \bar{F}_w \frac{\mathrm{d}\delta_*}{\mathrm{d}(\mu + \delta_*)}$$

is the censored $(\mu + \delta_*)$-density of $\mathrm{F}^{\flat}_{w}$.

We can simplify this density as follows. Understand

Put together, we arrive at

$$f^{\flat}_{w,h}(y) = w(y)\frac{\mathrm{dF}}{\mathrm{d}\mu}(y)\mathbb{1}_{\mathcal{Y}\setminus\{*\}}(y) + \bar{F}_w\mathbb{1}_*(y) = w(y)f(y)\mathbb{1}_{y\neq *} + \bar{F}_w\mathbb{1}_{y=*}, \quad y \in \mathcal{Y},$$

where f denotes the μ-density of F. □

A.2 Proof Corollary 2

The test based on $\tilde{\lambda}(\mathbf{y})$ is equivalent to the UMP test in Theorem 3, since

$$\begin{aligned}
\tilde{\lambda}(\mathbf{y}) &= \sum_{t=0}^{T-1} \left(S^{\mathrm{csl}}_{A_t}(f_{1t}, y_{t+1}) - S^{\mathrm{csl}}_{A_t}(f_{0t}, y_{t+1})\right) \\
&= \sum_{t=0}^{T-1} \left(\log\left([f_{1t}]^{\flat}_{A_t}(y_{t+1})\right) - \log\left([f_{0t}]^{\flat}_{A_t}(y_{t+1})\right) \right) \\
&= \log\lambda(\mathbf{y})
\end{aligned}$$

and hence $\lambda(\mathbf{y}) \underset{<}{\overset{\geq}{}} c \iff \tilde{\lambda}(\mathbf{y}) \underset{<}{\overset{\geq}{}} \tilde{c}$, with $\tilde{c} = \log c$.

A.3 Proof Corollary 3

We show that $\phi^{\sharp}_A$ is not UMP by a specific counterexample in which the power of $\phi^{\sharp}_A$ is strictly smaller than the power of $\phi^{\flat}_A$. In particular, suppose that $T = 1$ and consider two densities f_0 and f_1 that are different on $A = [r, \infty)$, for some constant $r > 0$. Furthermore, assume that

$$\int_{\{y:\lambda(y)>r\}}^{\infty} \mathrm{F}_0(\mathrm{d}y) > \alpha, \qquad \lambda(y) = \frac{f_1(y)}{f_0(y)}. \tag{A.1}$$

For $T = 1$, the likelihood ratios of the conditional and censored test simplify to

$$\begin{aligned}
\lambda^{\sharp}_A(y) &= \frac{[f_1]^{\sharp}_A(y)}{[f_0]^{\sharp}_A(y)} = \frac{\frac{f_1(y)}{\mathrm{F}_1(A)}}{\frac{f_0(y)}{\mathrm{F}_0(A)}}\mathbb{1}_A(y) = \frac{\mathrm{F}_0(A^c)}{\mathrm{F}_1(A^c)}\frac{f_1(y)}{f_0(y)}\mathbb{1}_A(y) \\
\lambda^{\flat}_A(y) &= \frac{[f_1]^{\flat}_A(y)}{[f_0]^{\flat}_A(y)} = \frac{f_1(y)}{f_0(y)}\mathbb{1}_A(y) + \frac{\mathrm{F}_1(A^c)}{\mathrm{F}_0(A^c)}\mathbb{1}_{A^c}(y).
\end{aligned}$$

Due to restriction (A.1), the corresponding critical regions $C^\sharp = \left[c^\sharp, \infty\right)$ and $C^\flat = \left[c^\flat, \infty\right)$ are both contained in A. Hence, an example in which $\sharp$ has higher power than $\flat$, would not only be a counterexample to Theorem 3 but also to the fundamental lemma of Neyman and Pearson (1933).

There exist many examples for which the power of the censored test is strictly larger than the power of the conditional test. For instance, suppose that $y \sim \text{Exp}(\theta_j)$, $j \in \{0, 1\}$, with $\theta_0 > \theta_1$. Then, the critical regions follow from the equation

$$\alpha = \int_{\{y:\lambda(y)>c^*\}}^{\infty} \theta_0 \mathrm{e}^{-\theta_0 y} \mathrm{d}y = \int_{\{y:a^*\left(\frac{\theta_1}{\theta_0}\right)\mathrm{e}^{-(\theta_1-\theta_0)y}>c^*\}}^{\infty} \theta_0 \mathrm{e}^{-\theta_0 y} \mathrm{d}y = 1 - F_0\left(\frac{1}{\theta_0 - \theta_1} \log\left(\frac{\theta_0}{\theta_1}\right) \frac{c^*}{a^*}\right),$$

where $a^\sharp = \frac{1-F_0(r)}{1-F_1(r)} = \mathrm{e}^{-(\theta_0-\theta_1)r}$ and $a^\flat = 1$. Isolating c^*, gives

$$c^* = ba^*, \qquad b = \frac{\theta_0}{\theta_1} \mathrm{e}^{(\theta_0-\theta_1)F_0^{-1}(1-\alpha)} > 0.$$

Now, the power of the conditional test is only weakly larger than the power of the censored test, if

$$\int_{\{y:\lambda(y)>c^\sharp\}}^{\infty} \theta_1 \mathrm{e}^{-\theta_1 y} \mathrm{d}y \geq \int_{\{y:\lambda(y)>c^\flat\}}^{\infty} \theta_1 \mathrm{e}^{-\theta_1 y} \mathrm{d}y \iff c^\sharp \geq c^\flat \iff (\theta_0 - \theta_1) r \leq 0.$$

But then, as $\theta_0 > \theta_1$ and $r > 0$, it follows that the power of the conditional test is always strictly smaller than the power of the censored test. Consequently, the conditional test $\phi_A^\sharp$ is not UMP. □

B Derivations Table 1

The results in Table 1 hold under the assumption that all $\mathrm{F} \in \mathscr{P}$ are Borel measures on $\mathbb{R}^d$ satisfying $\mathrm{F}(r) = 0$, with $r \in \mathbb{R}^d$. Furthermore, the assumption on h and w in the generalised censored scoring rule examples can predominantly be simplified due to the observation that $w(y)h(y) = 0$, $\forall y \in \mathcal{Y}$ and $f_w(y) = 0, \forall y \in \{w = 0\}$.

B.1 LogS

$$
\begin{aligned}
\mathrm{LogS}(\tilde{f},\tilde{y}) &= \log \tilde{f}(\tilde{y}) = \log f(y) - \log|b| \overset{\text{eqv.}}{=} \log f(y).\\
\mathrm{LogS}^{\sharp}_{w}(f,y) &= w(y)\log\left(\frac{w(y)f(y)}{1-\bar{F}_w}\right)\\
&\overset{\text{eqv.}}{=} w(y)\log\left(\frac{f(y)}{1-\bar{F}_w}\right)\\
&= S^{\mathrm{cl}}_{w}(f,y),\\
\mathrm{LogS}^{\flat}_{w}(f,y) &= w(y)\log\big(w(y)f(y)\mathbb{1}_{y\neq r} + \bar{F}_w\mathbb{1}_{y=r}\big) + \big(1-w(y)\big)\log\bar{F}_w\\
&\overset{\mu\text{-a.e.}}{=} w(y)\log\big(w(y)f(y)\big) + \big(1-w(y)\big)\log\bar{F}_w\\
&\overset{\text{eqv.}}{=} w(y)\log\big(f(y)\big) + \big(1-w(y)\big)\log\bar{F}_w\\
&= S^{\mathrm{csl}}_{w}(f,y).\\
\mathrm{LogS}^{\flat}_{w,h}(f,y) &= w(y)\log f^{\flat}_{w,h}(y) + \big(1-w(y)\big)\int_{\mathcal{Y}}\log f^{\flat}_{w,h}(q)h(q)\mathrm{d}q,\\
&= w(y)\Big(\log\big(f_w(y)\big)\mathbb{1}_{w>0} + \log\big(\bar{F}_w h(y)\big)\mathbb{1}_{w=0}\Big)\\
&\qquad + \big(1-w(y)\big)\int_{\{w=0\}}\Big(\log\big(f_w(q)\big)\mathbb{1}_{w>0} + \log\big(\bar{F}_w h(q)\big)\mathbb{1}_{w=0}\Big)h(q)\mathrm{d}q,\\
&= w(y)\log f_w(y) + \big(1-w(y)\big)\int_{\{w=0\}}\log\big(\bar{F}_w h(q)\big)\,h(q)\mathrm{d}q,\\
&\overset{\text{eqv.}}{=} w(y)\log f(y) + \big(1-w(y)\big)\log\bar{F}_w,\\
&= S^{\mathrm{csl}}(f,y).
\end{aligned}
$$

B.2 PsSphS$_\alpha$

$$
\mathrm{PsSphS}_{\alpha}(\tilde{f},\tilde{y}) = \frac{\tilde{f}(\tilde{y})}{\|\tilde{f}\|_{\alpha}^{\alpha-1}} = \frac{\left(\frac{1}{|b|}\right)^{\alpha-1} f(y)^{\alpha-1}}{\left(\frac{1}{|b|}\right)^{\frac{(\alpha-1)^2}{\alpha}}\|f\|_{\alpha}^{\alpha-1}} = \left(\frac{1}{|b|}\right)^{\frac{\alpha-1}{\alpha}}\mathrm{PsSphS}_{\alpha}(f,y).
$$

Next, we show the limit. Rescaling the PsSphS_α family by a factor $\frac{1}{\alpha-1}$, we obtain

$$\begin{aligned}
\lim_{\alpha\downarrow 1} \frac{1}{\alpha-1}\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1} &= \lim_{\alpha\downarrow 1} \frac{(\alpha-1)\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1}}{(\alpha-1)^2} \\
&= \lim_{\alpha\downarrow 1} \frac{\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1} + (\alpha-1)\left(\log\left(\frac{f(y)}{\|f\|_\alpha}\right) + (\alpha-1)\left(\frac{f(y)}{\|f\|_\alpha}\right)^{-1}\frac{\partial}{\partial\alpha}\frac{f(y)}{\|f\|_\alpha}\right)\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1}}{2(\alpha-1)} \\
&= \frac{1}{2}\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1} + \frac{1}{2}\lim_{\alpha\downarrow 1}\log\left(\frac{f(y)}{\|f\|_\alpha}\right)\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1} \\
&\quad + \frac{1}{2}\lim_{\alpha\downarrow 1}(\alpha-1)\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-2}\frac{\partial}{\partial\alpha}\frac{f(y)}{\|f\|_\alpha},
\end{aligned}$$

and hence

$$\lim_{\alpha\downarrow 1} \frac{1}{\alpha-1}\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1} = \log f(y), \tag{A.2}$$

since $\|f\|_1 = 1$. It might be helpful to note that the second equality in the first display follows from L'Hôpital's rule combined with the following derivative

$$\frac{\partial}{\partial\alpha}\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1} = \log\left(\left(\frac{f(y)}{\|f\|_\alpha}\right) + (\alpha-1)\left(\frac{f(y)}{\|f\|_\alpha}\right)^{-1}\frac{\partial}{\partial\alpha}\frac{f(y)}{\|f\|_\alpha}\right)\left(\frac{f(y)}{\|f\|_\alpha}\right)^{\alpha-1}.$$

For the conditional PsSphS_α family, we find

$$\begin{aligned}
\mathrm{PsSphS}^\sharp_{\alpha,w}(f,y) &= w(y)\frac{\left(\frac{f_w(y)}{1-F_w}\right)^{\alpha-1}}{\left(\int_{\mathcal{Y}}\left(\frac{f_w}{1-F_w}\right)^\alpha \mathrm{d}\mu\right)^{\frac{\alpha-1}{\alpha}}} \\
&= w(y)\frac{f_w(y)^{\alpha-1}}{\|f_w\|_\alpha^{\alpha-1}} \\
&= w(y)\left(\frac{f_w(y)^\alpha}{\|f_w\|_\alpha^\alpha}\right)^{\frac{\alpha-1}{\alpha}}
\end{aligned}$$

By the close similarity with Equation (A.2), it is uncomplicated to obtain the following

limit

$$\begin{aligned}\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PsSphS}^{\sharp}_{\alpha,w}(f,y) &= w(y)\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\left(\frac{f_w(y)}{\|f_w\|_\alpha}\right)^{\alpha-1}\\ &= w(y)\log\left(\frac{f_w(y)}{\|f_w\|_1}\right)\\ &= w(y)\log f^{\sharp}_w(y)\\ &= \mathrm{LogS}^{\sharp}_w(f,y),\end{aligned}$$

since $\|f_w\|_1 = \int_{\mathcal{Y}} w f \mathrm{d}\mu = 1 - \bar{F}_w$. Clearly, this result also follows directly from the linearity of limits, as

$$\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PsSphS}^{\sharp}_{\alpha}(f,y) = w(y)\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PsSphS}_{\alpha}(f^{\sharp}_w,y) = w(y)\log f^{\sharp}_w(y) = \mathrm{LogS}^{\sharp}_w(f,y). \tag{A.3}$$

Moreover, for the censored PsSphS_α family, it follows that

$$\begin{aligned}\mathrm{PsSphS}^{\flat}_w(f,y) &= \frac{w(y)\left(f_w(y)\mathbb{1}_{y\neq r} + \bar{F}_w\mathbb{1}_{y=r}\right)^{\alpha-1} + \big(1-w(y)\big)\bar{F}_w^{\alpha-1}}{\left(\int_{\mathcal{Y}}\left(f_w(y)\mathbb{1}_{y\neq r} + \bar{F}_w\mathbb{1}_{y=r}\right)^{\alpha}(\mu+\delta_r)(\mathrm{d}y)\right)^{\frac{\alpha-1}{\alpha}}}\\ &= \frac{w(y)\left(f_w(y)^{\alpha-1}\mathbb{1}_{y\neq r} + \bar{F}_w^{\alpha-1}\mathbb{1}_{y=r}\right) + \big(1-w(y)\big)\bar{F}_w^{\alpha-1}}{\left(\int_{\mathcal{Y}}\left(f_w(y)\right)^{\alpha}\mathrm{d}y + \bar{F}_w^{\alpha}\right)^{\frac{\alpha-1}{\alpha}}}\\ &\overset{\mu\text{-a.e.}}{=} \frac{w(y)f_w(y)^{\alpha-1} + \big(1-w(y)\big)\bar{F}_w^{\alpha-1}}{\left(\|f_w(y)\|_\alpha^{\alpha} + \bar{F}_w^{\alpha}\right)^{\frac{\alpha-1}{\alpha}}}.\end{aligned}$$

For the limit of $\alpha \downarrow 1$, we cannot directly apply Equation (A.2) as we did for the

conditional case. Nevertheless, we obtain a similarly satisfying result, namely

$$\begin{aligned}
\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PsSphS}_w^{\flat}(f,y) &= w(y)\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\left(\frac{f_w(y)}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)^{\alpha-1}\\
&\quad+\big(1-w(y)\big)\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\left(\frac{\bar F_w}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)^{\alpha-1}\\
&= w(y)\left(\lim_{\alpha\downarrow 1}\log\left(\frac{f_w(y)}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)\left(\frac{f_w(y)}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)^{\alpha-1}\right.\\
&\quad\left.+\lim_{\alpha\downarrow 1}(\alpha-1)\left(\frac{f_w(y)}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)^{\alpha-2}\frac{\partial}{\partial\alpha}\frac{f_w(y)}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)\\
&\quad+\big(1-w(y)\big)\left(\lim_{\alpha\downarrow 1}\log\left(\frac{\bar F_w}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)\left(\frac{\bar F_w}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)^{\alpha-1}\right.\\
&\quad\left.+\lim_{\alpha\downarrow 1}(\alpha-1)\left(\frac{\bar F_w}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)^{\alpha-2}\frac{\partial}{\partial\alpha}\frac{\bar F_w}{\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)^{\frac{1}{\alpha}}}\right)\\
&= w(y)\log f_w(y)+\big(1-w(y)\big)\log\bar F_w\\
&= \mathrm{LogS}_w^{\flat}(f,y),
\end{aligned}$$

where we have used that $\|f_w\|_1+\bar F_w = 1-\bar F_w+\bar F_w = 1$.

B.3 $\mathbf{PowS}_\alpha$

$$\begin{aligned}
\mathrm{PowS}_\alpha(\tilde f,\tilde y) &= \alpha\big(\tilde f(\tilde y)\big)^{\alpha-1}-(\alpha-1)\|\tilde f\|_\alpha^\alpha\\
&= \alpha\left(\frac{1}{|b|}\right)^{\alpha-1}f(y)-(\alpha-1)\left(\frac{1}{|b|}\right)^{\alpha-1}\|f\|_\alpha^\alpha\\
&= \left(\frac{1}{|b|}\right)^{\alpha-1}\mathrm{PowS}_\alpha(f,y),
\end{aligned}$$

as

$$\begin{aligned}\|\tilde{f}\|_\alpha^\alpha &= \int_{\tilde{\mathcal{Y}}} \tilde{f}(\tilde{y})^\alpha \mu(\mathrm{d}\tilde{y}) \\ &= \left(\frac{1}{|b|}\right)^{\alpha-1} \int_{\tilde{\mathcal{Y}}} \left(f\left(\frac{\tilde{y}-a}{b}\right)\right)^\alpha \frac{1}{|b|}\mu(\mathrm{d}\tilde{y}) \\ &= \left(\frac{1}{|b|}\right)^{\alpha-1} \int_{\mathcal{Y}} \left(f\left(y\right)\right)^\alpha \mu(\mathrm{d}y) \\ &= \left(\frac{1}{|b|}\right)^{\alpha-1} \|f\|_\alpha^\alpha.\end{aligned}$$

Next, we verify the limit for the non-focused family. Specifically,

$$\begin{aligned}\lim_{\alpha\downarrow 1} \frac{1}{\alpha-1}\mathrm{PowS}_\alpha &= \lim_{\alpha\downarrow 1} \frac{1}{\alpha-1}\left(\alpha f(y)^{\alpha-1} - (\alpha-1)\|f\|_\alpha^\alpha\right) \\ &= \lim_{\alpha\downarrow 1} \frac{(\alpha-1)\alpha f(y)^{\alpha-1}}{(\alpha-1)^2} - 1 \\ &= \lim_{\alpha\downarrow 1} \frac{\alpha f(y)^{\alpha-1} + (\alpha-1)f(y)^{\alpha-1}\big(1+\alpha\log f(y)\big)}{2(\alpha-1)} - 1 \\ &= \frac{1}{2}\left(\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\alpha f(y)^{\alpha-1} - 1\right) + \frac{1}{2}\left(\lim_{\alpha\downarrow 1} f(y)^{\alpha-1}\big(1+\alpha\log f(y)\big) - 1\right)\end{aligned}$$

and hence

$$\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PowS}_\alpha(f,y) = \log f(y).$$

Furthermore, the conditional version of the PowS_α family displayed in Table 1 is nothing but a direct application of the conditioning procedure. For the limit of the $\mathrm{PowS}^\sharp_{\alpha,w}$, we recall Equation (A.3) and immediately conclude that $\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PowS}^\sharp_{\alpha,w}(f,y) = \mathrm{LogS}^\sharp_w(f,y)$.

Turning to the censored focusing method, we recall from the analysis in Appendix B.2 that $\|f^\flat_w\|_\alpha^\alpha = \|f_w(y)\|_\alpha^\alpha + \bar{F}_w^\alpha$. Using this result, we obtain

$$\begin{aligned}\mathrm{PowS}^\flat_{\alpha,w}(f,y) &= w(y)\alpha\left(f_w(y)\mathbb{1}_{y\neq r} + \bar{F}_w\mathbb{1}_{y=c}\right)^{\alpha-1} + \big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1} - (\alpha-1)\|f^\flat_w\|_\alpha^\alpha \\ &\overset{\mu\text{-a.e.}}{=} w(y)\alpha f_w(y)^{\alpha-1} + \big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1} - (\alpha-1)\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\right),\end{aligned}$$

which bears the following limit

$$\begin{aligned}\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PowS}^{\flat}_{\alpha,w}(f,y) &= w(y)\lim_{\alpha\downarrow 1}\frac{(\alpha-1)\alpha f_w(y)^{\alpha-1}}{(\alpha-1)^2}+\big(1-w(y)\big)\lim_{\alpha\downarrow 1}\frac{(\alpha-1)\alpha \bar{F}_w^{\alpha-1}}{(\alpha-1)^2}-1\\ &= \frac{1}{2}\lim_{\alpha\downarrow 1}\left(w(y)\left(\frac{1}{\alpha-1}\alpha f_w(y)^{\alpha-1}-1\right)+\big(1-w(y)\big)\left(\frac{1}{\alpha-1}\alpha \bar{F}_w^{\alpha-1}-1\right)\right)\\ &\quad+\frac{1}{2}\lim_{\alpha\downarrow 1}\left(w(y)\left(f_w(y)^{\alpha-1}\big(1+\alpha\log f_w(y)\big)-1\right)\right.\\ &\quad\left.+\big(1-w(y)\big)\left(\bar{F}_w^{\alpha-1}\big(1+\alpha\log \bar{F}_w\big)-1\right)\right).\end{aligned}$$

Therefore,

$$\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PowS}^{\flat}_{\alpha,w}(f,y) = w(y)\log f_w(y)+\big(1-w(y)\big)\log \bar{F}_w = \mathrm{LogS}^{\flat}_w(f,y).$$

C Examples

C.1 Propriety CRPS* for Example 2

The extended metric $d^*:\mathbb{R}\cup\{*\}\times\mathbb{R}\cup\{*\}\to\mathbb{R}$ is given by

$$d^*(x,y)=\begin{cases}|x-y|, & \text{if } x\in\mathbb{R}, y\in\mathbb{R}\\ d(x), & \text{if } x\in\mathbb{R}, y=*\\ d(y), & \text{if } x=*, y\in\mathbb{R}\\ 0, & \text{if } x=y=*.\end{cases}$$

It is worth mentioning that Theorem 1 of Székely and Rizzo (2005) does not require $d(y)$ to be a metric. In particular, the triangle inequality does not need to hold. Therefore, the selected distance $d^*(y)=|y-r_1|\vee|y-r_2|$, as this implies a continuous $d^*(x^{\flat}_w,y^{\flat}_w)$ on $\{w>0\}\cup\{*\}\times\{w>0\}\cup\{*\}$. Indeed, for any $x<r$, it follows that $\lim_{y\downarrow r_1}d(x,y)=|x-r_1|=\lim_{y\to *}d(x,y)$, while for any $x>r$, it follows that $\lim_{y\uparrow r_2}d(x,y)=|x-r_2|=\lim_{y\to *}d(x,y)$. Furthermore, let $x_1,\ldots,x_n$ be n distinct but arbitrary outcomes in $\{w>0\}$. Consider the case, where one outcome is $*$, wlog

the last one. Then

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j d^*(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j d(\tilde{x}_i, \tilde{x}_j) \leq 0, \quad \tilde{x}_i = x_i \mathbb{1}_{\{w>0\}}(x_i) + r_k \mathbb{1}_{\{*\}},$$

with strict equality if and only if $a_i = 0,\ \forall i$ since d is a strictly negative definite kernel on $\mathbb{R} \times \mathbb{R}$ and $x_1, \ldots, x_{n-1}, r_k | x_i \in \{w > 0\}\ \forall i$ is a particular set of outcomes in $\mathbb{R},\ \forall k \in \{1, 2\}$. If none of the outcomes is $*$, we can borrow the strictly negative definiteness of d on $\mathbb{R} \times \mathbb{R}$ in a similar way, by understanding that any set of arbitrary outcomes in $\{w > 0\}$ is a particular set of arbitrary outcomes in $\mathbb{R}$. But then, Theorem 1 of Székely and Rizzo (2005) applies, whence the CRPS is also proper with respect to $(\mathbb{R}^*, d^*)$.

C.2 Localised NP for $T = 1$

Consider the special case $T = 1$. For one observation, it is straightforward to derive a most powerful test on A^c. For any $h_1 \in \mathscr{H}$, the relevant maximisation problem simplifies to

$$\begin{aligned}
\max_{\phi \in \Phi(\alpha)} \mathbb{E}_{p_1} \phi(y) &= \max_{\phi \in \Phi(\alpha)} \left\{ \mathbb{E}_{f_1} \phi_A(y) + \mathrm{F}_1(A^c) \mathbb{E}_{[h_1]^\sharp_{A_c}} \phi_{A^c}(y) \right\} \\
&= \max_{\alpha_A \leq \alpha} \left\{ \max_{\phi_A \in \Phi_A(\alpha_A)} \{\mathbb{E}_{f_1} \phi_A(y)\} + \mathrm{F}_1(A^c) \max_{\phi_{A^c} \in \Phi_{A^c}(\alpha - \alpha_A)} \left\{ \mathbb{E}_{[h_1]^\sharp_{A_c}} \phi_{A^c}(y) \right\} \right\} \\
&= \max_{\alpha_A \leq \alpha} \left\{ \max_{\phi_A \in \Phi_A(\alpha_A)} \{\mathbb{E}_{f_1} \phi_A(y)\} + \mathrm{F}_1(A^c) \frac{\alpha - \alpha_A}{\mathrm{F}_0(A^c)} \mathbb{1}_{A^c} \right\}.
\end{aligned}$$

After all, rejecting with probability $\frac{\alpha - \alpha_A}{\mathrm{F}_0(A^c)}$ if $y \in A^c$ is optimal since this is size correct and any more complicated test function ϕ_{A^c} has lower power. This can be verified as follows. For all level $\alpha - \alpha_A$ tests ϕ_{A^c}, i.e. $\phi_{A^c} \in \Phi_{A^c}(\alpha - \alpha_A)$, we have that

$$\begin{aligned}
\mathrm{F}_1(A^c) \mathbb{E}_{[h_1]^\sharp_{A_c}} \phi_{A^c}(y) &\leq \mathrm{F}_1(A^c) \sup_{h_1 \in \mathscr{H}} \{\mathbb{E}_{[h_1]^\sharp_{A_c}} \phi_{A^c}(y)\} \\
&= \mathrm{F}_1(A^c) \sup_{h_0 \in \mathscr{H}} \{\mathbb{E}_{[h_0]^\sharp_{A_c}} \phi_{A^c}(y)\} \\
&\leq \mathrm{F}_1(A^c) \frac{\alpha - \alpha_A}{\mathrm{F}_0(A^c)}.
\end{aligned}$$

Consequently, the test

$$\phi^*_{A^c}(y) = \frac{\alpha - \alpha_A}{\mathrm{F}_0(A^c)}, \qquad y \in A^c,$$

is most powerful against any other test $\phi^*_{A^c}(y)$ of size $\alpha - \alpha_A$.

This solution, also documented by Holzmann and Klar (2016), coincides with the UMP test given by Theorem 3. Indeed, suppose that the size α is such that $\frac{\mathrm{F}_1(A^c)}{\mathrm{F}_0(A^c)} = c$, i.e. not all of the size is spent on A, then the randomisation probability γ in Theorem 3 is such that

$$\alpha = \alpha_A + \gamma \mathrm{F}_0\left(\lambda(y) = \frac{\mathrm{F}_1(A^c)}{\mathrm{F}_0(A^c)}\right) = \alpha_A + \gamma \mathrm{F}_0\left(A^c\right) \implies \gamma = \frac{\alpha - \alpha_A}{\mathrm{F}_0(A^c)}.$$

C.3 CRPS

$$\begin{aligned}
\mathrm{CRPS}^\flat_{w_1}(\mathrm{F}, y) &= w_1(y)\mathrm{CRPS}(\mathrm{F}^\flat_{w_1}, y) + \big(1 - w_1(y)\big)\mathrm{CRPS}(\mathrm{F}^\flat_{w_1}, r) \\
&= \mathbb{1}_{(-\infty,r)}(y)\left(\int_{-\infty}^{r}\big(F(s) - \Delta_y(s)\big)^2\mathrm{d}s + \int_{r}^{\infty}(1-1)^2\mathrm{d}s\right) \\
&\quad + \big(1 - \mathbb{1}_{(-\infty,r)}(y)\big)\left(\int_{-\infty}^{r}\big(F(s) - \Delta_r(s)\big)^2\mathrm{d}s + \int_{r}^{\infty}(1-1)^2\mathrm{d}s\right) \\
&= \mathbb{1}_{(-\infty,r)}(y)\left(\int_{-\infty}^{r}\big(F(s) - \Delta_y(s)\big)^2\mathrm{d}s + \int_{r}^{\infty}(1-1)^2\mathrm{d}s\right) \\
&\quad + \big(1 - \mathbb{1}_{(-\infty,r)}(y)\big)\left(\int_{-\infty}^{r}\big(F(s) - \Delta_y(s)\big)^2\mathrm{d}s + \int_{r}^{\infty}(1-1)^2\mathrm{d}s\right) \\
&= \int_{-\infty}^{\infty} w_1(s)\big(F(s) - \Delta_y(s)\big)^2\mathrm{d}s
\end{aligned}$$

$$
\begin{aligned}
\mathrm{CRPS}^{\flat}_{w_2}(\mathrm{F}, y) &= w_2(y)\mathrm{CRPS}(\mathrm{F}^{\flat}_{w_2}, y) + \big(1 - w_2(y)\big)\mathrm{CRPS}(\mathrm{F}^{\flat}_{w_2}, r) \\
&= \mathbb{1}_{(r,\infty)}(y) \left(\int_{-\infty}^{r} (0-0)^2 \mathrm{d}s + \int_{r}^{\infty} \big(F(s) - \Delta_y(s)\big)^2 \mathrm{d}s \right) \\
&\qquad + \big(1 - \mathbb{1}_{(r,\infty)}(y)\big) \left(\int_{-\infty}^{r} (0-0)^2 \mathrm{d}s + \int_{r}^{\infty} \big(F(s) - \Delta_r(s)\big)^2 \mathrm{d}s \right) \\
&= \mathbb{1}_{(r,\infty)}(y) \left(\int_{-\infty}^{r} (0-0)^2 \mathrm{d}s + \int_{r}^{\infty} \big(F(s) - \Delta_y(s)\big)^2 \mathrm{d}s \right) \\
&\qquad + \big(1 - \mathbb{1}_{(r,\infty)}(y)\big) \left(\int_{-\infty}^{r} (0-0)^2 \mathrm{d}s + \int_{r}^{\infty} \big(F(s) - \Delta_y(s)\big)^2 \mathrm{d}s \right) \\
&= \int_{-\infty}^{\infty} w_2(s) \big(F(s) - \Delta_y(s)\big)^2 \mathrm{d}s
\end{aligned}
$$

References

Holzmann, H. and B. Klar (2016). Weighted scoring rules and hypothesis testing. Available at arXiv:1611.07345v2.

Neyman, J. and E. S. Pearson (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231*(694-706), 289–337.

Székely, G. J. and M. L. Rizzo (2005). A new test for multivariate normality. *Journal of Multivariate Analysis 93*(1), 58–80.