Localizing Strictly Proper Scoring Rules*

Ramon F. A. de Punder
Department of Quantitative Economics
University of Amsterdam and Tinbergen Institute

Cees G. H. Diks†
Department of Quantitative Economics
University of Amsterdam and Tinbergen Institute

Roger J. A. Laeven
Department of Quantitative Economics
University of Amsterdam, CentER and EURANDOM

Dick J. C. van Dijk
Department of Econometrics
Erasumus University Rotterdam and Tinbergen Institute

December 24, 2023

*We are very grateful to Timo Dimitriadis, Tilmann Gneiting, Alexander Jordan, Frank Kleibergen, Siem Jan Koopman, Sebastian Lerch, Xiaochun Meng, Marc-Oliver Pohle, Johanna Ziegel and participants at various seminars and conferences, including at the Heidelberg Institute for Theoretical Studies, Tinbergen Institute, University of Copenhagen, the 42nd International Symposium on Forecasting in Oxford (July 2022), the 10th International Workshop on Applied Probability in Thessaloniki (June 2023), the 5th Quantitative Finance and Financial Econometrics International Conference in Marseille (June 2023), the 12th ECB Conference on Forecasting Techniques in Frankfurt (June 2023), the 16th Meeting of the Netherlands Econometric Study Group in Rotterdam (June 2023) and the International Association for Applied Econometrics Annual Conference in Oslo (June 2023), for their comments and suggestions. This research was supported in part by the Netherlands Organization for Scientific Research under grant NWO Vici 2020–2025 (Laeven).

†Corresponding author. Mailing Address: PO Box 15867, 1001 NJ Amsterdam, The Netherlands. Phone: +31 (0) 20 525 4252. Email: C.G.H.Diks@uva.nl.

Abstract

When comparing predictive distributions, forecasters are typically not equally interested in all regions of the outcome space. To address the demand for focused forecast evaluation, we propose a procedure to transform strictly proper scoring rules into their localized counterparts while preserving strict propriety. This is accomplished by applying the original scoring rule to a censored distribution, acknowledging that censoring emerges as a natural localization device due to its ability to retain precisely all relevant information of the original distribution. Our procedure nests the censored likelihood score as a special case. Among a multitude of others, it also implies a class of censored kernel scores that offers a multivariate alternative to the threshold weighted Continuously Ranked Probability Score (twCRPS), extending its local propriety to more general weight functions than single tail indicators. Within this localized framework, we obtain a generalization of the Neyman Pearson lemma, establishing the censored likelihood ratio test as uniformly most powerful. For other tests of localized equal predictive performance, results of Monte Carlo simulations and empirical applications to risk management, inflation and climate data consistently emphasize the superior power properties of censoring.

Keywords: Density forecast evaluation; Tests for equal predictive ability; Censoring; Likelihood ratio; CRPS.

1 INTRODUCTION

Over the past decades, probabilistic forecasts have garnered increasing attention across a variety of disciplines, primarily because they provide a more comprehensive understanding of the stochastic nature of a random variable under scrutiny than point forecasts (Dawid 1984). A cornerstone for the effective evaluation of such probabilistic forecasts is the use of strictly proper scoring rules (Gneiting and Raftery 2007; Brehmer and Gneiting 2020; Patton 2020), which have been widely advocated for their ability to ensure fair comparative assessments of different forecast methods. While the usefulness of regular probabilistic forecasting is well-recognized and well-understood, various applications, such as the assessment of large financial portfolio losses, inflation targets or temperature ranges, require a focused, localized evaluation of predictive distributions.

In this paper, we introduce a natural localization mechanism for strictly proper scoring rules that preserves strict propriety. By censoring (Bernoulli 1760; Tobin 1958) the ob-

servation and distribution before applying the original scoring rule, we find a sweet spot between retaining and discarding information when focusing the original distribution to a region of interest. Specifically, unlike existing approaches that employ conditional distributions, our method preserves the overall probability of receiving an observation in (or outside) the target region, obviously relevant when comparing various candidate distributions focused on the same area. Moreover, within the region of interest, our mechanism replicates the original distribution's shape, which is particularly beneficial when evaluating functionals specific to this region, like quantiles or conditional expectations. Our procedure can be used to generate a multitude of strictly locally proper scoring rules. These include as special cases the censored likelihood (CSL) score, proposed by Diks et al. (2011), and the threshold weighted Continuously Ranked Probability Score (twCRPS), proposed by Gneiting and Ranjan (2011), for weight functions for which Holzmann and Klar (2017a) have shown that the twCRPS is strictly locally proper. On the other hand, for weight functions for which the twCRPS is not strictly locally proper, our analysis delineates the adverse consequences arising from this failure in localization, and provides a strictly locally proper alternative.

The additional information retained by our censoring approach also translates into advantageous power properties of tests aimed to compare density forecasts on regions of interest. We prove a generalization of the Neyman Pearson (1933) lemma, revealing that the censored likelihood ratio leads to a Uniformly Most Powerful (UMP) test. By contrast, we provide explicit evidence that the conditional likelihood (CL) score does not admit a UMP test. Monte Carlo simulations and empirical applications analyze the power properties of the Diebold and Mariano (2002) (DM) type test statistic, within the framework of Giacomini and White (2006), based on conditional vis-à-vis censored scoring rules. Censored

scoring rules enhance power in all three Monte Carlo experiments we have conducted. Substantial spurious power is observed solely for conditional scoring rules, which also falter in terms of power when tails become proportional. In multiple empirical experiments, which span financial, macroeconomic and climate data, we integrate the DM tests into the Model Confidence Set (MCS) as proposed by Hansen et al. (2011). The MCSs resulting from censored scoring rules are typically much smaller than their conditional counterparts, aligning with the power enhancements due to censoring displayed by the Monte Carlo results.

Our research contributes to the literature on focused scoring rules, initiated by the weighted likelihood score of Amisano and Giacomini (2007). Diks et al. (2011) and Gneiting and Ranjan (2011) sought to correct the (regular) impropriety of this scoring rule by introducing the CL, CSL and twCRPS, respectively. Holzmann and Klar (2017a) substantially advanced focused scoring rules, by generalizing the case of the CL score to construct proportionally locally proper scoring rules, based on conditioning, from regular scoring rules other than the logarithmic scoring rule. They also show that strict local propriety of the ensuing scoring rules can be restored by adding an auxiliary weighted scoring rule, based on an arbitrary strictly proper scoring rule for the probability of an observation landing in the region of interest. Our work differs importantly from theirs by opting for censoring rather than conditioning as localization mechanism. Through censoring, we enable the direct application of the original scoring rule to the localized measure, thereby avoiding the introduction of an auxiliary scoring rule and preserving the original Bregman divergence. As detailed by Brehmer and Gneiting (2020, Theorem 1), the conditional scoring rules of Holzmann and Klar (2017a) can also be viewed as an extension of the weighted likelihood score refined through a 'properization' process. Consequently, properization is not a viable mechanism for retaining strict propriety of the original scoring rule.

Our research also rests upon a substantial body of research concerning regular strictly proper scoring rules and their associated divergence measures. While the formalization of strict propriety was rigorously achieved by Gneiting and Raftery (2007), scoring rules satisfying this property date back to at least the Quadratic Scoring rule of Brier (1950). Literature in this domain has evolved from an initial focus on discrete settings to a more general treatment. In this vein, we rely on the expanded frameworks of the Power (PowS_α) and PseudoSpherical (PsSphS_α) families as advocated by Gneiting and Raftery (2007) and Ovcharov (2018) rather than their discrete foundations and refer to Gneiting and Raftery (2007) for foundational references. Additionally, scoring rules are inherently connected with divergence measures; under the restriction of strict propriety, these measures are subsumed under Bregman divergences (Dawid 2007; Ovcharov 2018; Painsky and Wornell 2020). This effectively excludes f-divergences other than Kullback-Leibler divergence (Kullback and Leibler 1951), distinguished for its favorable properties (Liese and Vajda 2006).

Interest in targeting specific regions of predictive distributions has surged across diverse fields, underscored by analyses of extreme events in disciplines such as meteorology, climatology, hydrology, finance, and economics (Lerch et al. 2017). In financial risk management, attention is particularly concentrated on the left tail of return distributions, conforming to mandated risk metrics like Value-at-Risk and Expected Shortfall (Cont et al. 2010; Fissler et al. 2015). Analogously, in macroeconomics, concepts such as 'Inflation-at-Risk' and 'Growth-at-Risk' are emerging, signifying values that deviate significantly from benchmarks established by institutions like Central Banks (Adrian et al. 2019; Lopez-Salido and Loria 2020; Iacopini et al. 2023). In other scenarios, the emphasis might rest on the central region or on another specific region of the distribution, often dictated by external constraints or objectives. Examples range from optimizing growing conditions for specific

crops like tubers, to calibrating wind speeds for peak wind turbine performance, and regulating blood sugar levels for effective diabetes management. They necessitate region-specific performance evaluations aligned with the interest in particular outcomes. Accordingly, as illustrated by Lerch et al. (2017), it is crucial to distinguish between strict propriety and strict local propriety; failing to do so can result in misleading forecast results.

This paper is organized as follows. Section 2 provides the foundational concepts essential for the subsequent analysis. Section 3 introduces the Censored Scoring Rule and establishes its strict local propriety. This section also introduces the Z-Q-Randomization procedure, proven to be equivalent to the Censored Scoring Rule, and showcases a variety of examples. It concludes with a generalization of the Neyman Pearson lemma and the main results of the simulation study. Section 4 discusses the empirical performance of our approach. Section 5 concludes. In accompanying Supplementary Material, we provide the proofs of our results, derivations of the theoretical properties tabulated in Section 3, extensive details of the Monte Carlo study, and full tables underlying the performance reported in Section 4.

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 a measurable space $(\mathcal{Y}, \mathcal{G})$. Denote by $\mathcal{P}$ a convex class of probability distributions on $(\mathcal{Y}, \mathcal{G})$. A *scoring rule* S assigns numerical values (scores) to observations $y \in \mathcal{Y}$ and distributions $\mathrm{F} \in \mathcal{P}$, through a mapping $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\} =: \bar{\mathbb{R}}$. Following Holzmann and Klar (2017a), we assume that a scoring rule S is measurable with respect to $\mathcal{G}$ and quasi-integrable with respect to all $\mathrm{P} \in \mathcal{P}$, for all $\mathrm{F} \in \mathcal{P}$, and such that $\mathbb{E}_{\mathrm{P}} S(\mathrm{F}, Y) < \infty$

and $\mathbb{E}_{\mathrm{P}} S(\mathrm{P}, Y) \in \mathbb{R}, \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$. The latter condition guarantees that the *score divergence*, $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) := \mathbb{E}_{\mathrm{P}} S(\mathrm{P}, Y) - \mathbb{E}_{\mathrm{P}} S(\mathrm{F}, Y)$, exists, and maps onto $(-\infty, \infty]$. Adhering to Gneiting and Raftery (2007), a minimal requirement for S is that it is *strictly proper*.

Definition 1 (Strictly proper scoring rule)**.** *A scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is proper relative to $\mathcal{P}$ if $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) \geq 0,\ \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$, and strictly proper if, additionally, $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) = 0$ if and only if $\mathrm{P} = \mathrm{F},\ \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$.*

Equivalently, a score divergence is a divergence measure (see e.g., Eguchi, 1985) if and only if S is 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 $\mathcal{Y}$, this divergence is known to be a Bregman (1967) divergence under the conditions listed by Ovcharov (2018). Two remarks are in place. First, distributions $\mathrm{F} \in \mathcal{P}$ are compared in terms of their P-expected score differences, whence it follows that uniqueness of members in $\mathcal{P}$ should formally be interpreted in terms of P-a.s. equivalence classes of P. For ease of exposition, we omit technicalities about P-a.s. equivalence throughout. Second, if there exists a σ-finite measure μ such that $\mathrm{F} \ll \mu,\ \forall \mathrm{F} \in \mathcal{P}$, with $\ll$ denoting absolute continuity, then scoring rules and associated definitions and results can easily be formulated relative to the class of induced μ-densities $f = \frac{\mathrm{dF}}{\mathrm{d}\mu}$, also denoted by $\mathcal{P}$, like classes of distributions functions F.

Gneiting and Raftery (2007) provide an extensive list of strictly proper scoring rules, which can be divided into *local* scoring rules and *distance-sensitive* scoring rules (Ehm and Gneiting 2012). We use the same distinction when discussing examples, yet allowing local scoring rules to also depend on the density via a global norm of the density, and refer to them henceforth as *semi-local*. In this subcategory, our focus lies on the Logarithmic (LogS), Quadratic (QS) and Spherical (SphS) scoring rules, along with their extensions to the Power (PowS_α) and PseudoSpherical (PsSphS_α) families. Our choice of distance-

sensitive scoring rules is confined to the Energy Scores (ES) subfamily, 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) as a special case.

2.2 Weighted scoring rules

Example 1 (The need to focus)**.** *Let Y be a random variable that follows a piecewise uniform distribution across the intervals $A = [0, 1)$, $B = [1, 2)$ and $C = [2, 3]$, with probabilities π_A, π_B and π_C, respectively. Figure 1 displays the densities and distribution functions of the true distribution* P *and two candidates* F *and* G*. Consider the CRPS, which is strictly proper and has score divergence $\mathbb{D}_{CRPS}(\mathrm{F}||\mathrm{G}) = \int_0^3 \big(F(s) - G(s)\big)^2 \mathrm{d}s$. From the right panel of Figure 1 it is apparent that $\mathbb{D}_{CRPS}(\mathrm{P}||\mathrm{F}) > \mathbb{D}_{CRPS}(\mathrm{P}||\mathrm{G})$. However, if only observations in B are pertinent, the ranking induced by $\mathbb{D}_{CRPS}$ fails because* F *coincides with* P *on B, that is, $\mathrm{P}(E \cap B) = \mathrm{F}(E \cap B), \forall E \in \mathcal{G}$, in contrast to* G*.*

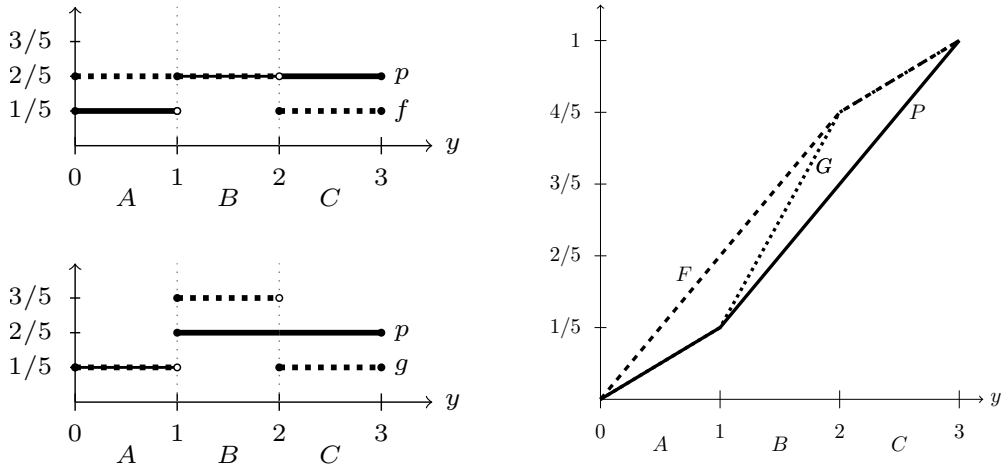

Figure 1: Densities (left) and distribution functions (right) of distributions F, G and true distribution P, all piecewise uniformly distributed on $[0, 3]$ but with different probabilities $\boldsymbol{\pi} := (\pi_A, \pi_B, \pi_C)'$. Specifically, $\boldsymbol{\pi}_p = (1/5, 2/5, 2/5)'$, $\boldsymbol{\pi}_f = (2/5, 2/5, 1/5)'$ and $\boldsymbol{\pi}_g = (1/5, 3/5, 1/5)'$.

As demonstrated by Example 1, it is imperative to adapt the scoring rule when particular outcomes are of importance. Otherwise, an excellent fit in non-critical regions of the

outcome space may obscure a poor fit in regions of actual relevance. Modeling the relative importance of outcomes $y \in \mathcal{Y}$ by a *weight function* $w \in \mathcal{W}$, with $\mathcal{W}$ consisting of all $\mathcal{G}$-measurable mappings $w : \mathcal{Y} \to [0, 1]$, the question arises how to transform the original scoring rule S given this weight function. We concur with the arguments put forward by Holzmann and Klar (2017a) that the weighted scoring rule, S_w, must be *localizing*. Specifically, for all outcomes, the variation in S_w should be solely dependent on changes in the distribution within the region of interest $\{w > 0\} := \{y \in \mathcal{Y} : w(y) > 0\}$; see Definition 2.

Definition 2 (Localizing weighted scoring rule)**.** *A weighted scoring rule S_w, that is, a map $S_\cdot : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ such that $S_w(\cdot, \cdot)$ is a scoring rule for each $w \in \mathcal{W}$, is localizing if for any P, $\mathrm{F} \in \mathcal{P}$, $w \in \mathcal{W}$, it holds that*

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

If a weighted scoring rule is non-localizing, this may cause what we refer to as a *localization bias*, as illustrated by Example 2.

Example 2 (Localization bias)**.** *Revisit Example 1. Suppose that the region of interest is B, with corresponding weight function $w(y) = \mathbb{1}_B(y)$. A prevalent weighted version of the CRPS is given by twCRPS$(F, y) = \int_B \big(F(s) - \mathbb{1}_{[y,\infty)}(s)\big)^2 \mathrm{d}s$, with score divergence $\mathbb{D}_{twCRPS}(\mathrm{P}\|\mathrm{F}) = \int_B \big(F(s) - G(s)\big)^2 \mathrm{d}s$; see Gneiting and Ranjan (2011). This weighted variant of the CRPS is clearly non-localizing, for instance, because its value is influenced by $\mathrm{F}(A)$, while $\mathrm{F}(A)$ is not implied by $\mathrm{F}(B)$, only the sum $\mathrm{F}(A) + \mathrm{F}(C)$ is. Consequently, the scoring rule depends on the distribution F outside B in a way that is not implied by F restricted to B. Its failure to be localizing introduces a bias in evaluating distributions over the region B. Indeed, by accounting for behavior of F and G on A (i.e., outside B) where G is closer to P than F (see Figure 1), the twCRPS inappropriately favors G on B.*

Example 3 (Improper localizing weighted scoring rule)**.** *We examine the weighted likelihood score* $wl(f, y) = \log f(y)\mathbb{1}_B(y)$ *proposed by Amisano and Giacomini (2007), in the context of Example 1. Although the unweighted logarithmic scoring rule is strictly proper and the weighted likelihood score is localizing, it is not locally proper, and still inappropriately favors* G*. Specifically, we have* $\log g(y) > \log p(y), \forall y \in B$*, which implies* $\mathbb{D}_{wl}(\mathrm{P}\|\mathrm{P}) > \mathbb{D}_{wl}(\mathrm{P}\|\mathrm{G})$*.*

Example 3 illustrates that localizing versions of strictly proper scoring rules are not automatically proper for all weight functions. In light of this, we focus on the subclass of localizing scoring rules that maintain this property. By construction, a localizing weighted scoring rule cannot be *strictly* proper unless $w(y) > 0, \forall y \in \mathcal{Y}$. This is because any distribution $\tilde{\mathrm{P}}$ equivalent to P on $\{w > 0\}$ but different on $\{w = 0\}$ will receive an identical score. Nonetheless, as illustrated by Example 4 below, some notion of local *strictness* remains advantageous. As recalled in the example, this is not achieved by the family of weighted scoring rules

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

analyzed in detail by Holzmann and Klar (2017a), where S is a regular scoring rule, $\mathrm{dF}_w := w\mathrm{dF}$ is the weighted kernel of distribution F and $\bar{F}_w = \int_{\mathcal{Y}}(1-w)\mathrm{dF}$. For indicator weight functions, $\mathrm{F}_w^\sharp$ simplifies to a conditional distribution on the region of interest. Henceforth, we refer to $S_w^\sharp$ as a *conditional scoring rule* for general weight functions.

Example 4 (Proportionally locally proper)**.** *Consider the weighted scoring rule* $S_w^\sharp(\mathrm{F}, y)$ *in Equation* (1)*. This scoring rule is localizing and proper for weight functions for which it remains a scoring rule (see Section 2.1). Yet, when revisiting Example 1 with* $w(y) = \mathbb{1}_B(y)$*, we have that* $S_B^\sharp(\mathrm{F}, y) = S_B^\sharp(\mathrm{G}, y) = S_B^\sharp(\mathrm{P}, y), \forall y \in B$*, since* $S_w^\sharp$ *cannot discriminate between distributions that are proportional to each other on* $\{w > 0\}$*. Accordingly,*

$\mathbb{D}_{S_B^\sharp}(\mathrm{P}\|\mathrm{F}) = \mathbb{D}_{S_B^\sharp}(\mathrm{P}\|\mathrm{G}) = 0$, *while only* F *coincides with* P *on* B. *In other words, the score divergence* $\mathbb{D}_{S_B^\sharp}$ *of a candidate distribution and* P *is properly zero if, but not only if, the candidate coincides with* P *on* B, *as is the case for* F.

Motivated by Examples 2, 3 and 4, this paper posits the necessity for weighted scoring rules to be *strictly locally proper*, as articulated in Definition 3. Compared to the definition of strict propriety (Definition 1), strictness is only required locally. More precisely, equivalent distributions on $\{w > 0\}$ must have weighted score divergence zero and, vice versa, distributions at zero weighted score divergence of each other must be equivalent on $\{w > 0\}$, the latter ruling out the ambiguities highlighted in Example 4.

Definition 3 (Strictly locally proper scoring rule)**.** *A weighted scoring rule* $S_\cdot : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ *is locally proper relative to* $(\mathcal{P}, \mathcal{W})$ *if it is localizing and* $S_w(\cdot, \cdot)$ *is proper for each* $w \in \mathcal{W}$. *Furthermore, it is strictly locally proper relative to* $(\mathcal{P}, \mathcal{W})$ *if, additionally,*

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

3 THE CENSORED SCORING RULE

To overcome issues such as the non-locality and non-strictness of the weighted scoring rules discussed above, we propose to use censoring as focusing mechanism. Censoring (Bernoulli 1760) refers to the statistical concept used to model a variable under scrutiny whose value, upon measurement or observation, is only partially known (Tobin 1958). More formally, under censoring, for realizations of a random variable Y that occur in A^c, the complement of some $A \subseteq \mathcal{Y}$, it is only known that they are not in A. Realizations in A^c are hence indistinguishable under censoring and 'A^c' may therefore be viewed as a single realization of the censored random variable. To avoid confusion, we label realizations in A^c by '$*$'

rather than 'A^c' itself, which is nothing but an abstract event, interchangeable with 'NaN'.

To facilitate censoring mathematically, we let $\mathcal{Y}$ and $\mathcal{G}$ both contain '$*$' and set $\mathrm{F}(*) = 0$, $\forall \mathrm{F} \in \mathcal{P}$, the latter rendering a choice for $w(*) \in [0,1]$ irrelevant. So, if one has some random variable on a measurable space $(\mathcal{X}, \mathcal{A})$ in mind, this measurable space is extended to $(\mathcal{Y}, \mathcal{G}) = \big(\mathcal{X} \cup \{*\}, \sigma(\{\mathcal{A}, *\})\big)$, where $\sigma(\{\mathcal{A}, *\})$ denotes the smallest σ-algebra containing the collection $\{\mathcal{A}, *\}$. The *censored random variable*

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

then defines a map from a probability space $(\mathcal{Y}, \mathcal{G}, \mathrm{F})$ to $(\mathcal{Y}, \mathcal{G})$, $\forall \mathrm{F} \in \mathcal{P}$. Similar to the conditional distribution in Example 4, we extend the definition of the distribution of $Y_A^\flat$ from indicator functions $w(y) = \mathbb{1}_A(y)$ to general weight functions $w \in \mathcal{W}$. Specifically, we define the *censored distribution* as

$$\mathrm{dF}_w^\flat := \mathrm{dF}_w + \bar{F}_w \mathrm{d}\delta_*, \qquad \bar{F}_w = \int_{\mathcal{Y}} (1-w)\mathrm{dF}, \qquad w \in \mathcal{W},\ \mathrm{F} \in \mathcal{P}, \tag{2}$$

where δ_* denotes the Dirac measure at $*$, i.e., $\delta_*(E) = \mathbb{1}_E(*)$.

In case $\mathrm{F} \ll \mu, \forall \mathrm{F} \in \mathcal{P}$, we may work with the μ-densities $f \in \mathcal{P}$ instead, and their associated $(\mu + \delta_*)$-densities

$$f_w^\flat = w f \mathbb{1}_{y \neq *} + \bar{F}_w \mathbb{1}_{y=*}, \qquad w \in \mathcal{W},\ f \in \mathcal{P}. \tag{3}$$

A detailed proof of this result is deferred to Appendix B.1. Borowska et al. (2020) also work with an explicit formulation of the censored density, albeit restricted to $w(y) = \mathbb{1}_A(y)$, coinciding with $f_A^\flat$, in the context of maximum likelihood. To ease notation, we adopt the subscript A instead of $\mathbb{1}_A$ when referencing indicator functions. The symbols 'sharp' ($\sharp$) and 'flat' ($\flat$) reflect their respective operations: conditioning sharpens the density on A by a factor $1/(1-\bar{F}_w)$, whereas censoring flattens the shape outside A into a point mass.

3.1 Censored scoring

Ideally, the censored scoring rule would be given by

$$S_A^\flat(\mathrm{F}, y) = S(\mathrm{F}_A^\flat, y_A^\flat), \tag{4}$$

as this would fully respect the forecaster's specific choice of the regular scoring rule S. The censored scoring rule given by Definition 4 below 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. This will be particularly clear from the randomization perspective provided in Section 3.2, which yields a similar identity for general weight functions; see Equation (6).

Definition 4 (Censored scoring rule)**.** *Let $S : \mathcal{P}^\flat \times \mathcal{Y} \to \bar{\mathbb{R}}$, $\mathcal{P}^\flat = \{\mathrm{F}_w^\flat, \mathrm{F} \in \mathcal{P}, w \in \mathcal{W}\}$, denote a regular scoring rule. Then, the corresponding censored scoring rule is given by the map $S_\cdot^\flat : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$,*

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

where the censored distribution $\mathrm{F}_w^\flat$ is defined in Equation (2).

Theorem 1 establishes that the censored scoring rule is strictly locally proper.

Theorem 1. *Suppose that the regular scoring rule S is strictly proper relative to $\mathcal{P}^\flat$. Then, the censored scoring rule $S^\flat$ in Definition 4 is strictly locally proper relative to $(\mathcal{P}, \mathcal{W})$.*

Theorem 1 is a special case of the more general Theorem 2 below, hence its proof is subsumed in the proof of Theorem 2. The assumption imposed in Theorem 1 ensures that the regular scoring rule is well-defined with respect to mixed continuous-discrete distributions on measurable spaces extended by '$*$'. In Subsection 3.3, we will verify that this assumption holds in the examples discussed.

Let us provide some intuition for the result of Theorem 1. Given some weight function $w \in \mathcal{W}$, it is clear that censoring maintains a one-to-one correspondence with the original distribution on $\{w > 0\}$. This correspondence is invalidated by conditioning due to the additional normalization of the weighted kernel. This difference is very explicit for indicator weight functions since $\mathrm{F}^{\flat}_{A} = \mathrm{F}$, while $\mathrm{F}^{\sharp}_{A} \neq \mathrm{F}$, 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. Consequently, the assumed strict propriety of the original rule localizes to $\{w > 0\}$ for the censored scoring rule.

Leveraging this intuition, one might conjecture that more general transformations to the distribution that suitably replace the Dirac measure in Definition 4 by an arbitrary nuisance distribution may also be performed, as long as the transformation remains independent of the original distribution and 'identifiable' when comparing two candidate distributions. The latter requirement, formalized by Assumption 1 below, ensures that the *generalized censored scoring rule* in Definition 5 is still strictly locally proper.

Definition 5 (Generalized censored scoring rule)**.** *Let $S : \mathcal{P}^{\flat} \times \mathcal{Y} \to \bar{\mathbb{R}}$ denote a regular scoring rule and $\mathcal{H} \subseteq \mathcal{P}$ a class of nuisance distributions. The associated generalized censored scoring rule is given by the map $S^{\flat}_{\cdot,\cdot} : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \times \mathcal{H} \to \bar{\mathbb{R}}$,*

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

where $\mathrm{F}^{\flat}_{w,\mathrm{H}}$ is referred to as the generalized censored distribution of F and $\mathrm{H} \in \mathcal{H}$ denotes the distribution of the random variable Q.

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

The following theorem, the proof of which is contained in Appendix A.1, establishes the strict local propriety of the generalized scoring rule.

Theorem 2. *Suppose that: (i) the regular scoring rule S in Definition 5 is strictly proper relative to $\mathcal{P}^\flat$, and (ii) $\mathcal{W}$ and $\mathcal{H}$ are such that Assumption 1 is satisfied. Then, the generalized censored scoring rule $S^\flat$ in Definition 5 is strictly locally proper relative to $(\mathcal{P}, \mathcal{W}, \mathcal{H})$.*

We refer to H as a *nuisance* distribution since its sole role is to suitably allocate the probability mass $\bar{F}_w$. Correspondingly, the rationale behind the choice of H is to add as little information to the censored distribution as the regular scoring rule permits. For example, the choice $\mathrm{dH} = \mathrm{d}\delta_*$ provides no information about the location of $\bar{F}_w$, particularly appropriate for semi-local scoring rules. Yet, when dealing with scoring rules based on distribution functions, which are restricted to real numbers, the scoring rule demands information about the location of $\bar{F}_w$, e.g., incorporated by replacing δ_* by $\delta_{\mathbf{r}}$, where $\mathbf{r} \in \mathbb{R}^d$; see Section 3.3. Selecting $\delta_{\mathbf{r}}$ as nuisance distribution in such cases easily upholds Assumption 1 as a regularity condition, as it suffices to restrict to distributions without a point mass at $\mathbf{r}$ and/or weight functions satisfying $w(\mathbf{r}) = 0$. Additionally, with $\mathrm{F}(*) = 0$ by definition, Assumption 1 is trivially met for $\mathrm{dH} = \mathrm{d}\delta_*$, the choice of H in Theorem 1.

Finally, a corollary of Lemma A2 in the proof of Theorem 2 is that

$$\mathbb{D}_{S^\flat_{w,\mathrm{H}}}(\mathrm{F}\|\mathrm{G}) = \mathbb{D}_S(\mathrm{F}^\flat_{w,\mathrm{H}}\|\mathrm{G}^\flat_{w,\mathrm{H}}), \tag{5}$$

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 *localized divergence measures*, satisfying the properties of a divergence measure (see Section 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 $\mathrm{F}(E \cap \{w > 0\}) = \mathrm{G}(E \cap \{w > 0\})$, $\forall E \in \mathcal{G}$.

3.2 Z, Q-Randomization

The (generalized) censored scoring rule in Definition 4 (5) can alternatively be formulated in terms of a randomization procedure. This is particularly appealing for general weight functions for which it yields an identity similar to Equation (4) for indicator weight functions. This procedure relies on an auxiliary random variable Z_w, indicating, conditional on the realization 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 $Z_w|(Y = y) \sim \text{BIN}\big(1, w(y)\big)$. By working out the conditional expectation, it is obvious that $Y^\flat_w = \mathbb{E}_{Z_w|Y}\varphi(Y, Z_w)$ coincides with the specification of the censored random variable in Equation (2). 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-randomization* definition of the censored scoring rule reads

$$S^\flat_w(\mathrm{F}, y) = \mathbb{E}_{Z_w|(Y=y)} S(\mathrm{F}^\flat_w, y^\flat_{Z_w}), \tag{6}$$

which is equivalent to the censored scoring rule defined by Definition 4.

A similar line of reasoning holds for the generalized censored scoring rule. In addition to the auxiliary random variable Z_w, we introduce an independent random variable Q with distribution H. Rather than labeling the observation as censored, we now take a random draw from Q if $Z_w = 0$, i.e., we define

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

As anticipated, the distribution of $Y^\flat_{w,\mathrm{H}} = \mathbb{E}_{Z_w|Y,\mathrm{H}}\varphi_{w,\mathrm{H}}(y, Z_w, Q)$ coincides with the specification of $\mathrm{F}^\flat_{w,\mathrm{H}}$ in Definition 5. Additionally, the generalized censored scoring rule of

Definition 5 admits the Z, Q-*randomization* representation

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

which reduces to Equation (6) for $\mathrm{H} = \delta_*$.

3.3 Examples

We now apply our censoring procedure to the regular scoring rules defined in Subsection 2.1. Following the classification into semi-local and distance-sensitive scoring rules, we start by localizing the former class.

Semi-local scoring rules. Together with the main characteristics of the LogS, PowS_α and PsSphS_α families, Table 1 presents the localized versions of these families based on conditioning, censoring and generalized censoring. As displayed in Table 1, each of the regular families is strictly proper relative to $\mathcal{P}_\alpha$, the class of μ-densities with a finite L^α-norm, where $\alpha = 1$ for LogS. Hence, one can easily verify their strict propriety with respect to $\mathcal{P}^{\flat}_\alpha$ as required for Theorems 1 and 2, since $\|f^{\flat}_w\|^\alpha_\alpha \leq 1 + \|f\|^\alpha_\alpha < \infty$, $\forall f \in \mathcal{P}_\alpha$, $\forall w \in \mathcal{W}$.

Upon comparing the censored and conditioned versions of the rules in Table 1, 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 demonstrates that the semi-local censored scoring rules are independent of $*$, the label of a censored observation. The generalized censored scoring rules in Table 1 extend these findings. Specifically, these rules maintain invariance to the choice of the nuisance density on $\{w = 0\}$ upon normalization by the α-norm of h, i.e., to the class of densities $\tilde{h} = h/\|h\|_\alpha$, where $\alpha = 1$ for LogS. Since $\|h\|_1 = 1$, this means that LogS is invariant to the unnormalized choice of h, as can be seen from Table 1. Lastly, Table 1 includes the localized divergence measures $\mathbb{D}_{S^{\flat}_w}$, which are all localized Bregman divergences since all

Table 1: Examples of semi-local scoring rules.

Name	Logarithmic	Power family	PseudoSpherical family
		Regular	
$S(f,y)$	$\mathrm{LogS}(f,y) = \log f(y)$	$\mathrm{PowS}_\alpha(f,y) = \alpha f(y)^{\alpha-1} - (\alpha-1)\|f\|_\alpha^\alpha, \quad \alpha > 1$	$\mathrm{PsSphS}_\alpha(f,y) = \frac{f(y)^{\alpha-1}}{\|f\|_\alpha^{\alpha-1}}, \quad \alpha > 1$
Special cases	–	$\mathrm{QS}(f,y) = \mathrm{PowS}_2(f,y)$	$\mathrm{SphS}(f,y) = \mathrm{PsSphS}_2(f,y)$
		$\mathrm{LogS}(f,y) = \lim_{\alpha\downarrow 1} \mathrm{PowS}_\alpha(f,y)$	$\mathrm{LogS}(f,y) = \lim_{\alpha\downarrow 1} \mathrm{PsSphS}_\alpha(f,y)$
$\mathbb{D}_S(f\|g)$	$\mathrm{KL}(f\|g) = \mathbb{E}_f \log\left(\frac{f}{g}\right)$	$\|f\|_\alpha^\alpha - \alpha\int g^{\alpha-1}(f-g)\mathrm{d}\mu - \|g\|_\alpha^\alpha$	$\|f\|_\alpha - \frac{\int f g^{\alpha-1}\mathrm{d}\mu}{\|g\|_\alpha^{\alpha-1}}$
$\alpha = 2$	–	$\|f-g\|_2^2$	$\|f\|_2\,(1 - C(f,g))$
SP class	$\mathcal{P}_{\alpha=1}$	$\mathcal{P}_\alpha$	$\mathcal{P}_\alpha$
$\zeta(t)$	$t\log t$	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	
$S_w^\sharp(f,y)$	$w(y)\log\left(\frac{f(y)}{1-F_w}\right)$	$w(y)\left(\alpha\left(\frac{f_w(y)}{1-F_w}\right)^{\alpha-1} - (\alpha-1)\left\|\frac{f_w(y)}{1-F_w}\right\|_\alpha^\alpha\right)$	$w(y)\frac{f_w(y)^{\alpha-1}}{\|f_w\|_\alpha^{\alpha-1}}$
$S_w^\flat(f,y)$	$w(y)\log f(y) + \big(1-w(y)\big)\log\bar{F}_w$	$w(y)\alpha f_w(y)^{\alpha-1} + \big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1} - (\alpha-1)\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1} + (1-w(y))\bar{F}_w^{\alpha-1}}{\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$S_{w,h}^\flat(f,y)$	$w(y)\log f(y) + \big(1-w(y)\big)\log\bar{F}_w$	$w(y)\alpha f_w(y)^{\alpha-1} + \big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1}\|h\|_\alpha^\alpha - (\alpha-1)\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\|h\|_\alpha^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1} + \big(1-w(y)\big)\bar{F}_w^{\alpha-1}\|h\|_\alpha^\alpha}{\left(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\|h\|_\alpha^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$\mathbb{D}_{S_w^\flat}(f\|g)$	$\int\log\left(\frac{f_w}{g_w}\right)f_w\mathrm{d}\mu + \log\left(\frac{\bar{F}_w}{\bar{G}_w}\right)\bar{F}_w$	$\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha - \int f_w g_w^{\alpha-1}\mathrm{d}\mu - \bar{F}_w\bar{G}_w^{\alpha-1} - (\alpha-1)\big(\|g_w\|_\alpha^\alpha + \bar{G}_w^\alpha\big)$	$\big(\|f_w\|_\alpha^\alpha + \bar{F}_w^\alpha\big)^{\frac{1}{\alpha}} - \frac{\int f_w g_w^{\alpha-1}\mathrm{d}\mu + \bar{F}_w\bar{G}_w^{\alpha-1}}{\big(\|g_w\|_\alpha^\alpha + \bar{G}_w^\alpha\big)^{\frac{\alpha-1}{\alpha}}}$

NOTE: This table displays regular and focused scoring rules, divergences and associated properties based on two μ-densities f and g, living on the measurable space $(\mathcal{Y},\mathcal{G},\mu)$, equipped with the L^α-norm $\|f\|_\alpha = (\int_{\mathcal{Y}} f^\alpha \mathrm{d}\mu)^{1/\alpha}$. The common limiting case of PowS_α and PsSphS_α remains to hold for conditioning and censoring, i.e., $\lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PsSphS}^x_{\alpha,w}(f,y) = \lim_{\alpha\downarrow 1}\frac{1}{\alpha-1}\mathrm{PowS}^x_{\alpha,w}(f,y) = \mathrm{LogS}^x(f,y)$, $x \in \{\sharp,\flat\}$. $\mathbb{D}_S(f||g)$ denotes the score divergence of g from f and $C(f,g) = \int fg\mathrm{d}\mu/\sqrt{\int f^2\mathrm{d}\mu\int g^2\mathrm{d}\mu}$, the cosine similarity between f and g. The strict propriety class (SP class) is the class of probability measures relative to which the scoring rule is strictly proper. $\mathcal{P}_\alpha$ denotes the class of densities on $(\mathcal{Y},\mathcal{G},\mu)$ such that $\|f\|_\alpha < \infty, \ \ \forall f \in \mathcal{P}_\alpha$. The Bregman generator function $\zeta(t)$ parameterizes the subclass of separable Bregman divergences, consisting of the score divergences based on strictly proper scoring rules $S_\zeta : \mathcal{P}(\mathcal{Y},\mathcal{G}) \times \mathcal{Y} \to \mathbb{R}$ of the form $S_\zeta(p,y) = \zeta'(p(y)) - \int_{\mathcal{Y}}\zeta'(p(y))p(y) - \zeta(p(y))\mu(\mathrm{d}y)$. $S(\tilde{f},\tilde{y})$ denotes the score of the real-valued random variable $\tilde{Y} = bY + a$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}\backslash\{0\}$, with density $\tilde{f}(\tilde{y}) = \frac{1}{|b|}f\big(\frac{\tilde{y}-a}{b}\big)$. The presented results for the focused scoring rules are equivalent in the sense that they yield the same expected score. The generalized censored scoring rule $S_{w,h}^\flat$ departs from a density h of which the support is a subset of $\{w=0\} \subseteq \mathcal{Y}$. The weight function is restricted accordingly. Appendix C details the derivations of the results presented in this table.

regular divergences $\mathbb{D}_S$ in this table are necessarily of the Bregman type.

Distance-sensitive scoring rules. A rich class of distance-sensitive scoring rules is the Energy Score family given by

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

where $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Y}}$ denote independent copies of a d-dimensional random vector with distribution $\mathrm{F} \in \mathcal{P}_\beta$ and $\|\cdot\|_2$ denotes the Euclidean norm. Moreover, $\mathcal{P}_\beta$ denotes the class of Borel probability measures on $\mathbb{R}^d$ such that $\mathbb{E}_F\|\mathbf{Y}\|_2^\beta < \infty$, the class relative to which the ES_β family is known to be strictly proper $\mathcal{P}_\beta$ (Gneiting and Raftery 2007). In contrast to the semi-local scoring rules, the corresponding censored ES family is sensitive to $*$, particularly to the (yet undefined) distance $d(\mathbf{y}) = \|\mathbf{y} - *\|_2$. Specifically,

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

To define $d(\mathbf{y})$, one straightforward approach is to set $* \in \mathbb{R}^d$, thereby indicating the location of $\bar{F}_w$. It is important to recognize that the censored scoring rule's outcome is influenced by this choice, as $*$ now represents both the censored event and the value assigned post-censoring. Motivated by the empirical observation that weight functions often possess 'pivotal points', such as the edges of an indicator function or the center of a kernel (Gneiting and Ranjan 2011), we refrain from introducing some general concept for censored distances d. Instead, we allocate the residual mass $\bar{F}_w$ across a set of pivotal points $\mathbf{r}_1, \ldots, \mathbf{r}_k \in \mathbb{R}^d$, i.e.,

$$\mathrm{dF}^\flat_{w,\boldsymbol{\gamma}} := \mathrm{dF}_w + \bar{F}_w \sum_{i=1}^k \gamma_i \mathrm{d}\delta_{\mathbf{r}_i}, \qquad \boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_k)' \in \Delta(k), \tag{7}$$

where $\Delta(k)$ denotes the unit simplex. Adding information about the location of $\bar{F}_w$ when demanded by the scoring rule reflects the discussion of the selection of the nuisance distribution in Section 3.1. Furthermore, similar to the semi-local scoring rules, it is straightforward

to verify the strict propriety of the ES_β family relative to $\mathcal{P}_\beta^\flat$. Indeed, $\forall \mathrm{F} \in \mathcal{P}_\beta, w \in \mathcal{W}$, it follows that $\mathbb{E}_{\mathrm{F}^\flat_{w,\boldsymbol{\gamma}}}\|\mathbf{Y}\|_2^\beta = \int_{\mathbb{R}^d} \|\mathbf{y}\|_2^\beta \mathrm{F}_w(\mathrm{d}\mathbf{y}) + \bar{F}_w \sum_{i=1}^k \gamma_i \|\mathbf{r}_i\|_2^\beta < \mathbb{E}_\mathrm{F}\|\mathbf{Y}\|_2^\beta + \sum_{i=1}^k \|\mathbf{r}_i\|_2^\beta < \infty$. This approach being the conventional one for scoring rules sensitive to distance, we henceforth omit the dependence on $\boldsymbol{\gamma}$ in our notation.

Our approach to distance-sensitive scoring rules has some direct implications for the CRPS. First of all, for all left- and right-tail indicator functions, with pivotal point r, one can easily show that the $\mathrm{CRPS}_w^\flat$ coincides with twCRPS. In other words, $\mathrm{CRPS}_w^\flat =$ twCRPS, for all weight functions for which Holzmann and Klar (2017a, Theorem 5) proved that the twCRPS is strictly locally proper. Second, for other weight functions such as the center indicator functions for which the twCRPS loses its strict local propriety due to its non-localizing nature (see Example 2), $\mathrm{CRPS}^\flat$ serves as a strictly locally proper alternative. This alternative bears an additional parameter γ, the selection of which is contingent on the specific application. For the indicator function $w(y) = \mathbb{1}_{[r_1,r_2]}$ it is natural to choose $\gamma = \frac{1}{2}$ if $r_1 = -r$ and $r_2 = r$ and one aims to compare the predictive ability of two candidates that are both symmetric around zero. Moreover, in applications where empirical data are available, one can alternatively distribute $\bar{F}_w$ according to the empirical proportion of data falling into the left- and right tail. By using the same estimate $\hat{\gamma}$ for all considered candidate distributions, the censored scoring rule remains strictly locally proper. This would be different if one would use a candidate-derived probability for falling into either tail, e.g., by setting $\gamma_F = F(r_1)/\bar{F}_w$. Such a method results in a non-localizing scoring rule, as only the sum $F(r_1) + \big(1 - F(r_2)\big) = \bar{F}_w$ is implied by F on $[r_1, r_2]$. This non-localizing characteristic is evident in the twCRPS, which can be regarded as a *semi-censored scoring rule* $S_w^\dagger(\mathrm{F}, y) = S(\mathrm{F}_w^\dagger, y)$, where $\mathrm{dF}_w^\dagger = \mathrm{dF}_w + F(r_1)\mathrm{d}\delta_{r_1} + \big(1 - F(r_2)\big)\mathrm{d}\delta_{r_2}$, for $w(y) = \mathbb{1}_{[r_1,r_2]}(y)$. Another example of a semi-censored scoring rule is the centre 'censored'

log-likelihood introduced by Mitchell and Weale (2023) and Harvey and Liao (2023), which hence also fails to be strictly locally proper.

3.4 Localized Neyman Pearson

In anticipation of the applications contained in the next section, we now consider an explicit time-series context. Specifically, we 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 $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$, $\forall t$. The random variable of interest becomes Y_{t+1} conditional on $\mathcal{F}_t$, indicated by adding a subscript t to the notation of (predictive) distributions, distribution functions and μ-densities, assuming the existence of a dominating measure μ, $\forall t$. We adopt the same notation for objects related to Q_{t+1}. If desired, a generalization to a sequence of μ_t-densities is straightforward. Furthermore, the regions of interest $A_t \subseteq \mathcal{Y}$ are assumed to be $\mathcal{F}_t$-measurable.

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

$$\mathbb{H}_0 : p_t \mathbb{1}_{A_t} = f_{0t} \mathbb{1}_{A_t},\ \forall t \qquad \text{vs.} \qquad \mathbb{H}_1 : p_t \mathbb{1}_{A_t} = f_{1t} \mathbb{1}_{A_t}, \quad \forall t. \tag{8}$$

Although the predictive densities under the null and alternative hypotheses are assumed to be known, i.e., fixing $f_{jt},\ j \in \{0, 1\}$, 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. In other words, the densities $\left[f_{0t}\mathbb{1}_{A_t} + \big(\mathrm{F}_{0t}(A_t^c)/\mathrm{H}_t(A_t^c)\big) h_t \mathbb{1}_{A_t^c}\right]_{A_t}^{\flat}$ and $[f_{0t}]_{A_t}^{\flat}$ coincide, assuming $\mathrm{H}_t(A_t^c) > 0$. Here, $[\cdot]_w^{\flat}$ refers to censoring a distribution (function) and density according to Equations (2) and (3), respectively. Similarly, we use $[\cdot]_w^{\sharp}$ for

conditioning. This notation is particularly helpful in this subsection due to the additional subscripts related to time and hypotheses. Theorem 3 reveals that this setting admits a UMP test, reducing to the Neyman and Pearson (1933) lemma when $A_t = \mathcal{Y}$, $\forall t$. A detailed proof of this result is deferred to Appendix A.2.

Theorem 3 (Localized Neyman Pearson)**.** *For any given $\alpha \in (0,1)$, the UMP test of size α for testing problem* (8) *reads*

$$\phi_A^\flat(\mathbf{y}) = \begin{cases} 1, & \text{if } \lambda(\mathbf{y}) > c \\ \gamma, & \text{if } \lambda(\mathbf{y}) = c \\ 0, & \text{if } \lambda(\mathbf{y}) < c, \end{cases} \qquad \lambda(\mathbf{y}) := \frac{[f_1]_A^\flat(\mathbf{y})}{[f_0]_A^\flat(\mathbf{y})}, \quad [f_j]_A^\flat(\mathbf{y}) := \prod_{t=0}^{T-1} [f_{jt}]_{A_t}^\flat(y_{t+1}), \quad j \in \{0,1\},$$

where $\phi_A^\flat : \mathcal{Y}^T \to [0,1]$ denotes a test function specifying the rejection probability, c is the largest constant such that $[\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) \geq c\big) \geq \alpha$ and $[\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) \leq c\big) \geq 1-\alpha$, and $\gamma \in [0,1]$ is such that $\alpha = [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) > c\big) + \gamma [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) = c\big)$.

For $T \equiv 1$, the test in Theorem 3 reduces to the UMP test for a single observation proposed by Holzmann and Klar (2017b). Moreover, Corollary 1 reveals that the test in Theorem 3 can alternatively be formulated in terms of the CSL introduced by Diks et al. (2011). Corollary 2 endorses that the conditional operator does not bear a UMP test too, making the censored operator preferable to its conditional counterpart in the setting of this subsection. The proofs of Corollaries 1 and 2 are deferred to Appendices B.2 and B.3.

Corollary 1. *An alternative formulation of the UMP test for testing problem* (8) *is given by the test defined in Theorem 3 with $\lambda(\mathbf{y})$ replaced by $\tilde{\lambda}(\mathbf{y}) := \sum_{t=0}^{T-1} \big(\mathit{LogS}_{A_t}^\flat(f_{1t}, y_{t+1}) - \mathit{LogS}_{A_t}^\flat(f_{0t}, y_{t+1})\big)$, i.e., in terms of the CSL.*

Corollary 2. *For testing problem* (8)*, the test $\phi_A^\sharp$, which is defined as $\phi_A^\flat$ upon replacing $\flat$ by $\sharp$, is not UMP.*

3.5 Monte Carlo study

Employing a simulation design similar to Diks et al. (2011), Holzmann and Klar (2017b) and Lerch et al. (2017), we analyze in Monte Carlo simulations the size and power properties of the Giacomini and White (2006) test based on conditional and censored scoring rules. In this subsection, we summarize the main findings; the simulation results are described in full detail in Appendix D. The test we employ relies on the score difference series of two candidates $\hat{f}_t$ and $\hat{g}_t$, that is, realizations of $D^x_{t+1} := S^x_w(\hat{f}_t, Y_{t+1}) - S^x_w(\hat{g}_t, Y_{t+1})$, where $x \in \{\sharp, \flat\}$, in testing the null hypothesis $\mathbb{H}_0 : \mathbb{E}_{p_t} S^x_w(\hat{f}_t, Y_{t+1}) = \mathbb{E}_{p_t} S^x_w(\hat{g}_t, Y_{t+1})$, by means of

$$t_{T_{\text{est}},n} := \frac{1}{n} \sum_{t=T_{\text{est}}}^{T-1} \frac{d^x_{t+1}}{\sqrt{\hat{\sigma}^2_{T_{\text{est}},n}/n}}, \qquad n := T - T_{\text{est}},$$

where T_{est} denotes the length of the estimation window, and $\hat{\sigma}^2_{T_{\text{est}},n}$ is a heteroskedasticity and autocorrelation-consistent (HAC) variance estimator in non-i.i.d. settings. This null hypothesis, which is equivalent to $\mathbb{H}_0 : \mathbb{D}_{S^x_w}(p_t \| \hat{f}_t) = \mathbb{D}_{S^x_w}(p_t \| \hat{g}_t)$, is rejected if it is sufficiently unlikely that the localized score divergence from p_t to $\hat{f}_t$ and p_t to $\hat{g}_t$ coincide.

Appendix D.1 first confirms the good size properties of both conditional and censored scoring rules. A natural conjecture is that strictly locally proper scoring rules generally lead to higher power since they are sensitive with respect to all measurable aspects of the distribution. Yet, the dependence of the null hypothesis on the scoring rule makes the null and rejection sets dependent on the scoring rule too, obstructing theoretical results such as Theorem 3. Nevertheless, the power results displayed in Appendix D.2, are clearly in favor of censoring as localization mechanism: censoring yields both higher power and lower spurious power compared to conditioning in all three Monte Carlo experiments that we have conducted. In the left-tail application for standard Normal and Student-t candidates, the differences are less monotonic than in the two other experiments, due to the fact that the scores intersect by construction for the selection of candidates.

4 EMPIRICAL PERFORMANCE

In this section, we assess the empirical performance of censoring versus conditioning by comparing the MCS implied by conditional and censored scoring rules. As delineated by Hansen et al. (2011), the MCS procedure expands the Giacomini and White (2006) hypothesis to larger sets of $\mathbb{H}_0$-equivalent methods, employing an iterative elimination procedure to test the null of equal predictive performance of all methods in an initial set $\mathcal{M}_0$. This can be achieved by combining the relative scores into either TR:= $\max_{i,j\in\mathcal{M}_k} |t_{i,j}|$ or Tmax:= $\max_{i\in\mathcal{M}_k} t_i$, where $t_{i,j}$ refers to the t-statistic of the relative scores between methods i and j and t_i to the $t_{T,T_{\text{est}}}$-statistic of the relative scores between method i and the average score over all methods in $\mathcal{M}_k$, the set of methods that have survived until the k-th elimination round. Favorable power properties of censoring in the Giacomini and White (2006) environment intuitively accelerate elimination in the MCS procedure, resulting in smaller MCS p-values and, consequently, reduced cardinality. We present results at the 0.90 and 0.75 confidence levels, utilizing the TR statistic as benchmark with a block bootstrap with $B = 10,000$ replications and block length $k = 5$, unless stated otherwise. Our results are robust to variations in these parameters. When CRPS$^\flat$ and twCRPS differ, we include the twCRPS for reference. We quantify differences in cardinality in absolute terms, framed as the proportion of cases wherein the number of methods in MCS$^\flat$ is (strictly) smaller than MCS$^\sharp$, the MCS under censoring and conditioning. Additionally, we provide the factor by which the cardinality of the MCS expands when conditioning is adopted in lieu of censoring.

4.1 Risk management

Evaluating the downside risk of asset returns is a crucial task in risk management, particularly for compliance with regulatory requirements related to risk measures such as 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 mandated Expected Shortfall $\text{ES}^q_{\hat{f}_t}$, which quantifies expected losses conditional on exceeding $\text{VaR}^q_{\hat{f}_t}$. To achieve this, we opt for a weight function $w_t(y_t) = \mathbb{1}_{\left(-\infty,\hat{r}^q_t\right)}(y_t)$ and choose as 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 closing price on day t, adjusted for stock splits and dividends. The data consists of 6,777 daily observations, spanning from January 2, 1996, to December 30, 2022, sourced from Yahoo Finance.

All selected forecast methods conform to $Y_t|\mathcal{F}_{t-1} \sim \mathcal{D}(\mu, \sigma_t^2, \boldsymbol{\vartheta})$, denoting a parametric family of distributions with mean μ, variance σ_t^2 and other parameters collected in $\boldsymbol{\vartheta}$. While we evaluated AR(1) and AR(5) models for the conditional mean, they did not yield significant improvements over a constant mean specification. We consider three conditional variance models: the GARCH(1,1) model proposed by Bollerslev (1986), the more general TGARCH(1,1) model introduced by Glosten et al. (1993):

$$\sigma_t^2 = \omega + \alpha(y_t - \mu)^2 + \beta\sigma_{t-1}^2 + \gamma(y_t - \mu)^2\mathbb{1}_{y_t-\mu\leq 0}, \tag{9}$$

which reduces to GARCH(1,1) for $\gamma = 0$, and the RGARCH(1,1) model developed by Hansen et al. (2012), given by

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

where x_t represents the realized measure[1], $z_t = (y_t - \mu)/\sigma_t$, and u_t denotes a white noise process with variance σ_u^2. We combine each of the volatility models with a standard normal and Student-t_ν distribution, comprising six forecast methods in total. We estimate all parameters via maximum likelihood on a rolling window of length $T_{\text{est}} = 1,000$.

Table 2 reveals stark differences in the cardinality of $\text{MCS}^\flat$ and $\text{MCS}^\sharp$, particularly at the shortest forecast horizon $h = 1$. At a 0.90 confidence level and $h = 1$, $\text{MCS}^\sharp$ is smaller

[1]Downloaded from `https://dachxiu.chicagobooth.edu/#risklab`

only in one case across the examined quantiles and scoring rules, namely for $q = 0.25$ and $S = \text{QS}$, see Table E.1.a. Equality in MCS size occurs mainly for higher quantiles, where information scarcity with respect to the distributions on $(-\infty, \hat{r}_t^q)$ is less critical. For $h = 1$, $\text{MCS}^\sharp$ contains more than twice the number of methods compared to $\text{MCS}^\flat$ on average. For $h = 5$, the differential reduces but remains substantial, averaging around a factor 1.7.

Examining the composition of the MCSs reveals that the censored MCSs are often a subset of the conditional MCSs, when $|\text{MCS}^\flat| \leq |\text{MCS}^\sharp|$. The significance of reductions due to censoring is further emphasized by the fact that the resulting MCSs encompass more complex model specifications, which would be the optimal choices in the absence of parameter and forecasting uncertainty. Robustness checks, pertaining to k and T_{est}, confirm the stability with respect to these parameters (see Table E.1.b). Additionally, the use of the TR statistic tends to expedite model elimination, yielding smaller MCS p-values compared to Tmax; this acceleration, however, is consistent across both censoring and conditioning.

Beyond the statistical assessment of forecast methods, we compute their 1- and 5-step ahead Value at Risk ($\text{VaR}^q_{\hat{f}_t}$) and Expected Shortfall ($\text{ES}^q_{\hat{f}_t}$). 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 ($\text{VaR}^q_{\hat{f}_t}$) or conditional moment $\text{ES}^q_{\hat{f}_t} = \mathbb{E}_{\hat{f}_t}\left(Y_{t+h} | Y_{t+h} \leq \text{VaR}^q_{\hat{f}_t}\right)$. Notably, the conditioning in $\text{ES}^q_{\hat{f}_t}$ is a quantile of the density forecast itself rather than $\hat{r}_t^q$, a.s. implying a discrepancy between the operational region of $\text{ES}^q_{\hat{f}_t}$ and the focused scoring rules introduced above.

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

$$\mathbb{D}_{S^\flat_w}(p_t \| \hat{f}_t) = 0 \implies (\text{VaR}^q_{p_t}, \text{ES}^q_{p_t}) = (\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t}), \tag{10}$$

Table 2: Changes in MCS cardinality between censored and conditional scoring rules.

	Tail(s)						Interval					
	$\lvert\mathbf{MCS}_{0.90}\rvert$			$\lvert\mathbf{MCS}_{0.75}\rvert$			$\lvert\mathbf{MCS}_{0.90}\rvert$			$\lvert\mathbf{MCS}_{0.75}\rvert$		
h	$\leq$	$<$	$\sharp/\flat$	$\leq$	$<$	$\sharp/\flat$	$\leq$	$<$	$\sharp/\flat$	$\leq$	$<$	$\sharp/\flat$
Risk Management												
1	96%	71%	2.28	92%	63%	2.04						
5	75%	38%	1.69	58%	50%	1.72						
Inflation												
6	100%	92%	2.00	92%	83%	2.93	100%	83%	2.91	100%	100%	3.72
12	75%	50%	1.86	67%	58%	2.38	100%	67%	2.35	83%	75%	2.72
24	92%	75%	2.86	92%	58%	3.31	100%	67%	2.33	100%	92%	3.23
Climate												
1	87%	58%	2.01	75%	50%	1.74	92%	42%	1.54	83%	42%	1.46
2	87%	50%	1.63	87%	38%	1.40	100%	67%	1.67	100%	58%	1.58
3	83%	50%	1.80	83%	42%	1.35	100%	58%	1.58	100%	25%	1.25

NOTE: This table presents changes in cardinality of the MCS in absolute and relative terms, at confidence levels 0.75 and 0.90, across different forecast horizons h. Columns labeled $\leq$ ($<$) display the percentage of cases where $\text{MCS}^\flat$ contains (strictly) fewer forecast methods than $\text{MCS}^\sharp$ and the column labeled $\sharp/\flat$ reports the factor $|\text{MCS}^\sharp|/|\text{MCS}^\flat|$. Each of the results represents an average over a set of levels or quantiles q and scoring rules $S \in \{\text{LogS}, \text{QS}, \text{SphS}, \text{CRPS}\}$. The regions of interest for inflation are defined as $A_q = [2-q, 2+q]$ and its complement, where $q \in \{1, 1.5, 2\}$. For the climate data, $A_q = (r_q, \infty)$, where r_q is the empirical q-th quantile of the estimation window, with $q \in \{0.75, 0.80, 0.85, 0.90, 0.95, 0.99\}$ or $A_q = [18-q, 18+q]$ for $q \in \{1, 2, 4\}$. Complete MCS details and associated p-values are provided in Appendix E. The p-values are obtained via a block bootstrap of $B = 10,000$ replications, with block length $k = 5$, or $k = 200$ for the climate data.

where $w_t(y_t) = \mathbb{1}_{(-\infty, r_t)}(y_t)$. This is a direct consequence of (5), i.e., another corollary of Lemma A1, and holds also more generally for any functional of distributions on $\{w > 0\}$. In (sharp) contrast, $\mathbb{D}_{S_w^\sharp}(p_t \| \hat{f}_t) = 0$ implies that $p_t \propto \hat{f}_t$ on $(-\infty, r_t)$ and hence $(\text{VaR}_{p_t}^q, \text{ES}_{p_t}^q) \neq (\text{VaR}_{\hat{f}_t}^q, \text{ES}_{\hat{f}_t}^q)$, 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.

Thus, censoring is designed to generate MCSs containing forecast models that produce $(\text{VaR}_{\hat{f}_t}^q, \text{ES}_{\hat{f}_t}^q)$ pairs closer to the true pair. Support for this conjecture is found in Ta-

ble E.1.b. While 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 favorable: 14 versus 7, endorsing censored MCS as a superior selection mechanism for VaR and ES calculations.

4.2 Inflation

We next focus on forecasting inflation, a subject recently gaining prominence. Guided by the inflation target of 2% set by the Federal Reserve System (FED)[2] and European Central Bank (ECB)[3], we center our study on the range $A_q = [2-q, 2+q]$, where $q > 0$, employing the weight function $w(y_t) = \mathbb{1}_{A_q}(y_t)$. Simultaneously, we consider policymakers' concerns for deviations beyond A_q, termed 'Inflation at Risk' (Lopez-Salido and Loria 2020), utilizing the complement weight function $w(y_t) = \mathbb{1}_{A_q^c}(y_t)$.

While the evaluation ingredients remain almost exactly the same, the unique characteristics of the inflation time series necessitate an adapted set of forecast methods. We closely align with the methodology presented by Medeiros et al. (2021), using the same 122 variables from the FRED-MD database ($\mathbf{x}_t$), spanning January 1960 to December 2015. This timeframe encompasses a total of 672 monthly observations, with the final 180 being out-of-sample relative to the initial estimation window. While using the same baseline U.S. consumer price index $\text{CPI}_t =: P_t$ inflation as Medeiros et al. (2021), we follow Stock and Watson (2002) and Borup et al. (2022) by analyzing the h-step ahead forecasts of the accumulated series $y^h_{t+h} = (1200/h)\log\left(P_{t+h}/P_t\right)$, instead of the accumulation of the individual h-step ahead forecasts of the monthly rate. This direct approach is standard in the literature and especially advantageous for density forecasts, as accumulating densities

[2]Source: `https://federalreserve.gov/monetarypolicy/files/fomc_longerrungoals.pdf`

[3]Source: `https://ecb.europa.eu/mopo/implement/app/html/index.en.html`

is more complex than aggregating point forecasts.

Each of the forecast methods under consideration can be represented as

$$y_{t+h}^h = \mu_{j,t+h}^h(\mathbf{x}_t) + u_{t+h}^h, \qquad u_{t+h}^h|\mathcal{F}_t \sim \mathcal{N}_{\text{TP}}(0, \sigma_1, \sigma_2), \qquad \sigma_1, \sigma_2 > 0,$$

where $\mathcal{N}_{\text{TP}}(0, \sigma_1, \sigma_2)$ denotes the two-piece normal distribution. For the conditional mean $\mu_{j,t+h}^h$, we take the following subset of models listed by Medeiros et al. (2021): Random Walk, Auto-Regressive model (AR), Bagging, Complete Subset Regression (CSR), Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest models. The implementation specifics of these models are elaborated upon in Section 4 of Medeiros et al. (2021). The density of the two-piece normal distribution reads

$$f(y; \mu, \sigma_1, \sigma_2) = \frac{2}{\sigma_1 + \sigma_2} \left(\phi\left(\frac{y-\mu}{\sigma_1}\right) \mathbb{1}_{y<\mu} + \phi\left(\frac{y-\mu}{\sigma_2}\right) \mathbb{1}_{y\geq\mu} \right), \qquad \sigma_1, \sigma_2 > 0,$$

where $\phi(z)$ denotes the density of the standard normal distribution. This distributional choice is congruent with the underlying statistical model employed in the fan charts published by the Monetary Policy Committee of the Bank of England (Clements 2004; Mitchell and Hall 2005; Gneiting and Ranjan 2011).

The summary results presented in Table 2 show the difference between the cardinality of the $\text{MCS}^\flat$ and $\text{MCS}^\sharp$, averaged over $q \in \{1, 1.5, 2\}$. Table 2 reveals a distinct and pronounced preference for censoring. Notably, the cardinalities of $\text{MCS}^\flat$ are generally — with 'generally' here not seldom verging on unanimity — smaller than those of $\text{MCS}^\sharp$. This is especially salient in the Center case, where the $\text{MCS}^\flat$ are almost always weakly smaller than the corresponding $\text{MCS}^\sharp$. While it is unsurprising, given these results, that the relative increase in set cardinality when opting for conditioning over censoring is positive, the specific magnitudes of these increases even (substantially) exceed 100%. This is a striking finding; it effectively indicates that $\text{MCS}^\sharp$ consistently encompasses more than twice the number of methods compared to $\text{MCS}^\flat$, thereby making the use of $\text{MCS}^\sharp$ hard to defend.

The differences between the MCS variants are clearly highlighted by the p-values presented in Table E.2.b, which also offers more detailed insights. For $q = 1$ the cardinality of $\text{MCS}^{\sharp}_{0.90}$ consistently exceeds or equals that of $\text{MCS}^{\flat}_{0.90}$ with the sole exceptions occurring in tail cases predicated on the CRPS for $h = 12$ and $h = 24$, and QS for $h = 12$. These exceptions feature a marginal difference of one. At a confidence level of 0.75, a similar trend is observed, albeit without the QS exception for the tail case but with two additional exceptions for the center case at $h = 12$ in both the QS and CRPS rules.

Finally, a closer look at the differences between the twCRPS and $\text{CRPS}^{\flat}$ is in place. In the Center panel, we observe that the $\text{CRPS}^{\flat}$ is preferred to the twCRPS for $h = 6$ and $h = 24$, for both $q = 1$ and $q = 1.5$. For $h = 12$, the differences are less pronounced, slightly favoring the twCRPS for $q = 1$ and $q = 1.5$, but not for $q = 2$.

4.3 Climate

In a third application, we generate density forecasts for Dutch daily average temperature data, extending the data and methodology of Franses et al. (2001) and Tol (1996). We maintain focus on volatility clustering and changing asymmetries in past temperature to volatility relations, along with accounting for seasonal variations in the mean and variance. Contrary to Franses et al. (2001), we use daily observations instead of the implied weekly averages. The dataset spans from February 1, 2003, to January 31, 2023, with the first $T_{\text{est}} = 2922$ days serving as the initial estimation window. Our models closely follow the GARCH, QGARCH-I, and QGARCH-II specifications as in Franses et al. (2001), but with alterations in seasonal trend estimation. Specifically, we use local day averages for the mean and a sine function for volatility, as opposed to a quadratic function. The models

can be formalized as: $Y_t|\mathcal{F}_{t-1} \sim \mathcal{D}(\mu_t, \sigma_t^2, \boldsymbol{\vartheta})$, where $\mu_t = m_{t|t-1} + \phi y_{t-1}$ and

$$\sigma_t^2 = \varphi(t; \omega_0, \omega_1) + \alpha\big(y_{t-1} - \mu_{t-1} - \varphi(t; \gamma_0, \gamma_1)\big)^2 + \beta\sigma_{t-1}^2.$$

Here, $m_{t|t-1}$ is the average temperature of days with the same day number in the estimation window, that is, all $s \in [t - T_{\text{est}}, t-1]$ such that $\tilde{T}_s = \tilde{T}_t$, where $\tilde{T}_t = \min(T_t, 365)$, in which T_t is the day number, with $T_t = 1$ on the first of February. The latter choice exploits the periodic pattern revealed by Figure 1 in Franses et al. (2001), which we model by $\varphi(t; \theta_0, \theta_1) = \theta_0 + \theta_1|\sin(\pi/365 \cdot \tilde{T}_t)|$. These models are combined with both Normal and Student-t_ν distributions to produce six forecast methods.

The summary findings are presented in the Climate panel of Table 2, focussing on the right tail $(\hat{r}_t^q, \infty)$ and the interval $[18 - q, 18 + q]$. The latter interval has its roots in the agricultural literature, corresponding to the optimal temperature for tuber growth, agreed to be approximately 18 degrees Celsius (Struik 2007, Section 18.5.5). Analyzing the results of this interval case, it is observed that there are no instances where conditioning leads to a smaller MCS for $h = 2$ and $h = 3$ and almost no such cases for $h = 1$, similar to the inflation interval case. Relative to inflation, there is a notable increase in cases in which the MCSs possess identical cardinality, which is also reflected by the smaller factors $|\text{MCS}^\sharp|/|\text{MCS}^\flat|$. The MCS p-values reported in Table E.3.b reveal that the MCSs are consistently small in the interval case, frequently including one or both of the QGARCH-II methods. This observation suggests that the preference for censoring, as depicted in Table 2, translates into the censored scoring rule's more effective recognition of the QGARCH-II methods' pronounced superiority. Table E.3.b further demonstrates that the performance of the $\text{CRPS}^\flat$ and twCRPS is closely matched.

The results for the right tail example, corresponding to (exceedingly) high daily temperatures, exhibit parallels with the left-tail risk management application. In particular, the

cardinalities of the censored MCSs are typically smaller than their conditional counterparts; these disparities diminish as forecasting horizons extend. The tails panel of Table E.3.b reveals that, although to a lesser degree and particularly at elevated levels of q, the MCS often comprise relatively compact sets, encompassing one or both of the QGARCH-II methods.

5 CONCLUSION

In many applications, forecasters are particularly interested in specific areas of the outcome space. Addressing this, we champion censoring as focusing device, demonstrating that applying scoring rules to censored distributions results in strictly locally proper scoring rules. To the best of our knowledge, we are the first to derive a transformation of the original scoring rule that preserves strict propriety. Our approach features high flexibility, applicable across varied scoring rules, weight functions, and outcome spaces. For specific choices, the censored scoring rule yields intuitively appealing rules apt for practical use. For instance, we recover the twCRPS for tail indicators, while solving its localization bias for other weight functions.

Our second theoretical contribution, a generalization of the Neyman Pearson lemma, revolves around the censored likelihood score. We have shown that the UMP test of the localized Neyman Pearson hypothesis is a censored likelihood ratio test, reducing to the original lemma if the weight function is one for all outcomes. By contrast, the conditional likelihood ratio test is not UMP. Monte Carlo simulations incorporate the Giacomini and White test to assess the power properties of conditional versus censored scoring rules based on the score differences between two candidates. The findings endorse the superior power properties of censoring, extending beyond the stylized scenario in which the candidates' tails are close to proportional.

To analyze real performance, we use the size of the Model Confidence Set (MCS) as an indicator of power. Notably, in our inflation example — where the number of observations is characteristically low, akin to many macro-applications — the frequency with which the censored MCS is strictly smaller than the conditional MCS strikes, as does the difference in cardinality. These observations hold across different horizons, whether centered on the 2% target or its complement. In focused forecast assessments of S&P500 and temperature data, a comparable pattern emerges, corroborating the enhanced power of censoring.

SUPPLEMENTARY MATERIAL

All proofs and additional theoretical results, the Monte Carlo analysis, and full tables on the empirical performance are provided in an online supplementary document. (.pdf)

References

Adrian, T., N. Boyarchenko, and D. Giannone (2019), "Vulnerable Growth", *American Economic Review*, *109*(4), 1263–1289.

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 Mortalite Causee par la Petite Verole, et des Avantages de l'Inoculation Pour la Prevenir", *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, 1–45.

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, *31*(3), 307–327.

Borowska, A., L. Hoogerheide, S. J. Koopman, and H. K. Van Dijk (2020), "Partially Censored Posterior for Robust and Efficient Risk Evaluation", *Journal of Econometrics*, *217*(2), 335–355.

Borup, D., P. Goulet Coulombe, D. Rapach, E. C. M. Schütte, and S. Schwenk-Nebbe (2022). "The Anatomy of Out-of-Sample Forecasting Accuracy". FRB Atlanta Working Paper No. 2022-16. DOI: 10.29338/wp2022-16. Available at `https://papers.ssrn.com/abstract=4278745`.

Bregman, L. (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. R. 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.

Clements, M. P. (2004), "Evaluating the Bank of England Density Forecasts of Inflation", *The Economic Journal*, *114*(498), 844–866.

Cont, R., R. Deguest, and G. Scandolo (2010), "Robustness and Sensitivity Analysis of Risk Measurement Procedures", *Quantitative Finance*, *10*(6), 593–606.

Dawid, A. P. (1984), "Statistical Theory: The Prequential Approach", *Journal of the Royal Statistical Society. Series A (General)*, *147*(2), 278–292.

Dawid, A. P. (2007), "The Geometry of Proper Scoring Rules", *Annals of the Institute of Statistical Mathematics*, *59*(1), 77–93.

Diebold, F. X. and R. S. Mariano (2002), "Comparing Predictive Accuracy", *Journal of Business & Economic Statistics*, *20*(1), 134–144.

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. (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), "Local Proper Scoring Rules of Order Two", *The Annals of Statistics*, *40*(1), 609–637.

Fissler, T., J. F. Ziegel, and T. Gneiting (2015). "Expected Shortfall is Jointly Elicitable with Value at Risk - Implications for Backtesting". DOI: 10.48550/ARXIV.1507.00244. Available at `https://arxiv.org/abs/1507.00244`.

Franses, P. H., J. Neele, and D. Van Dijk (2001), "Modeling Asymmetric Volatility in Weekly Dutch Temperature Data", *Environmental Modelling & Software*, *16*(2), 131–137.

Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability", *Econometrica*, *74*(6), 1545–1578.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993), "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", *The Journal of Finance*, *48*(5), 1779–1801.

Gneiting, T. and A. E. 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.

Hansen, P. R., Z. Huang, and H. H. Shek (2012), "Realized GARCH: A Joint Model for Returns and Realized Measures of Volatility", *Journal of Applied Econometrics*, *27*(6), 877–906.

Hansen, P. R., A. Lunde, and J. Nason (2011), "The Model Confidence Set", *Econometrica*, *79*(2), 453–497.

Harvey, A. and Y. Liao (2023), "Dynamic Tobit Models", *Econometrics and Statistics*, *26*, 72–83.

Holzmann, H. and B. Klar (2017a), "Focusing on Regions of Interest in Forecast Evaluation", *The Annals of Applied Statistics*, *11*(4), 2404–2431.

Holzmann, H. and B. Klar (2017b). "Weighted Scoring Rules and Hypothesis Testing". Available at https://arxiv.org/abs/1611.07345v2.

Iacopini, M., F. Ravazzolo, and L. Rossini (2023), "Proper Scoring Rules for Evaluating Density Forecasts with Asymmetric Loss Functions", *Journal of Business & Economic Statistics*, *41*(2), 482–496.

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*, *32*(1), 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.

Lopez-Salido, D. and F. Loria (2020). "Inflation at Risk". Finance and Economics Discussion Series 2020-013. Washington: Board of Governors of the Federal Reserve System. Avalaible at https://doi.org/10.17016/FEDS.2020.013.

Medeiros, M. C., G. F. R. Vasconcelos, A. Veiga, and E. Zilberman (2021), "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods", *Journal of Business & Economic Statistics*, *39*(1), 98–119.

Mitchell, J. and S. G. Hall (2005), "Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation", *Oxford Bulletin of Economics and Statistics*, *67*(s1), 995–1033.

Mitchell, J. and M. Weale (2023), "Censored Density Forecasts: Production and Evaluation", *Journal of Applied Econometrics*, *38*(5), 714–734.

Neyman, J. and E. 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.

Ovcharov, E. Y. (2018), "Proper Scoring Rules and Bregman Divergence", *Bernoulli*, *24*(1), 53–79.

Painsky, A. and G. W. Wornell (2020), "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.

Stock, J. H. and M. W. Watson (2002), "Macroeconomic Forecasting Using Diffusion Indexes", *Journal of Business & Economic Statistics*, *20*(2), 147–162.

Struik, P. C. (2007). "Chapter 18 - Responses of the Potato Plant to Temperature". In D. Vreugdenhil, J. Bradshaw, C. Gebhardt, F. Govers, D. K. L. Mackerron, M. A. Taylor, and H. A. Ross (Eds.), *Potato Biology and Biotechnology*, pp. 367–393. Amsterdam: Elsevier Science B.V.

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables", *Econometrica*, *26*(1), 24–36.

Tol, R. S. (1996), "Autoregressive Conditional Heteroscedasticity in Daily Temperature Measurements", *Environmetrics*, *7*(1), 67–75.