Theory and methodology of scoring functions: tail properties, interval forecasts, and point processes

Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften der Universität Mannheim

vorgelegt von

Jonas Reiner Brehmer
aus Hünfeld

Mannheim 2020

Dekan: Dr. Bernd Lübcke, Universität Mannheim
Referent: Professor Dr. Martin Schlather, Universität Mannheim
Korreferenten: Dr. Kirstin Strokorb, Cardiff University, Professor Dr. Tilmann Gneiting,
Karlsruher Institut für Technologie

Tag der mündlichen Prüfung: 7. Dezember 2020

Abstract

Scoring functions are decision-theoretically principled tools to quantify predictive performance. A scoring function is strictly consistent for a statistical functional, e.g. the expectation, if its expected score is uniquely minimized by this functional. This thesis examines existence and properties of strictly consistent scoring functions for three kinds of functionals. Firstly, we introduce the class of max-functionals, which contains key characteristics from extreme value theory, for instance the extreme value index. We show that its members do not allow for forecast evaluation via strictly consistent scoring functions in a very strong sense. The second part develops results for interval forecasts. Strictly consistent scoring functions exist and can be characterized for two types, the equal-tailed and modal interval. However, for the shortest prediction interval, they are not available relative to practically relevant classes of distributions. Lastly, the third part introduces consistent scoring functions for point process characteristics, such as the intensity, which enables a novel approach to comparative forecast evaluation in this framework.

Zusammenfassung

Bewertungsfunktionen sind von entscheidungstheoretischen Prinzipien geleitete Hilfsmittel um den Erfolg von Vorhersagen zu bewerten. Eine Bewertungsfunktion ist strikt konsistent für ein statistisches Funktional, beispielsweise die Erwartung, wenn ihre erwartete Bewertung durch das Funktional eindeutig minimiert wird. Diese Arbeit untersucht Existenz und Eigenschaften strikt konsistenter Bewertungsfunktionen für drei Arten von Funktionalen. Zuerst führen wir die Klasse der Max-Funktionale ein, die zentrale Charakteristiken der Extremwerttheorie enthält, zum Beispiel den Extremwertindex. Wir zeigen, dass die Elemente dieser Klasse in einem starken Sinne keine Vorhersageauswertung mittels strikt konsistenter Bewertungsfunktionen zulassen. Der zweite Teil entwickelt Resultate für Intervallvorhersagen. Für zwei Typen von Intervallen, das "equal-tailed interval" und das Modalintervall, gibt es strikt konsistente Bewertungsfunkionen, die sich auch charakterisieren lassen. Für das kürzeste Prediktionsintervall sind jedoch, relativ zu praktisch relevanten Verteilungsklassen, keine solchen Funktionen verfügbar. Schließlich stellt der dritte Teil strikt konsistente Bewertungsfunktionen für Charakteristika von Punktprozessen vor, zum Beispiel für die Intensität. Dies ermöglicht eine neue Herangehensweise an die vergleichende Auswertung von Vorhersagen im Rahmen von Punktprozessen.

Acknowledgements

First of all, I want to thank my supervisors for their support and the enjoyable and fruitful scientific collaboration over the past years. I am indebted to Martin Schlather for motivating me to become a doctoral student at his chair. His confidence and optimism often provided much needed encouragement. I am very grateful to Kirstin Strokorb for undertaking this long-distance supervision as well as structuring and guiding my studies. Whenever needed, her advice was just a video call away. Lastly, I want to thank Tilmann Gneiting for his unfailing assistance, especially for sharing his invaluable experience and expertise in countless situations. His regular pointers to interesting open problems helped me develop and benefited this thesis a lot.
Special thanks go to Anja Gilliar for taking care of all administrative work and any day-to-day problems.

I am thankful for my great current and past colleagues at the University of Mannheim and the Heidelberg Institute for Theoretical Studies, who ensured a pleasant working environment and many enjoyable lunch times. In particular, I want to thank Christopher Dörr and Nicholas Schreck for proofreading and many helpful comments on earlier versions of this thesis.
Several other researchers contributed to this thesis via fruitful discussions and friendly advice. For this I want to thank Johannes Bracher, Timo Dimitriadis, Sebastian Engelke, Christopher Ferro, Tobias Fissler, Rafael Frongillo, Claudio Heinrich, Fabian Krüger, Sebastian Lerch, Marco Oesting, Patrick Schmidt, and Chen Zhou.

Finally, I am deeply grateful to my family and Ria for their continuous support and love throughout my studies and especially during the writing of this thesis.

I gratefully acknowledge financial support by the German Research Foundation (DFG) through the Research Training Group RTG 1953 'Statistical Modeling of Complex Systems and Processes'. I am also grateful for support by the Klaus Tschira Foundation.

Contents

Introduction

In all kinds of situations, people base their decisions on forecasts, which inform them on the likely implications of different actions. This practice ranges from everyday activities, where we can check weather forecasts to decide whether or not to carry an umbrella, to public infrastructure projects based on predicted traffic volume. Naturally, users of such forecasts want to know which of the competing sources of information will be most valuable for the decision at hand. If the phenomena of interest, e.g. amount of rain or traffic volume, can be quantified and exhibit some regular behavior, then statistical forecast evaluation methods can guide this search for high-quality forecasts. Specifically, one approach is to assume that the quantity of interest is a random variable Y which follows some unknown distribution F. A decision maker needs some kind of information on F in order to choose an appropriate action and thus asks several forecasters for their reports. After a period of time, these reports are then compared to a set of realizations of Y.

Scoring functions are widely used and well-studied tools to check whether one forecast outperforms its competitors (Gneiting, 2011a). The term 'scoring' usually refers to the act of assigning a real number, the 'score', to each pair of forecast and realized observation of Y. Then a low (or high, depending on convention) score signifies a good report. If the forecast is a statistical property or functional of the distribution of Y, e.g. the expectation or a quantile, such a mapping is usually called *scoring function*, whereas the term *scoring rule* is common when a full distribution is reported. In both cases, the key requirement is that forecasting the truth gives the minimal score in expectation: A scoring function is *consistent* for a statistical functional if the value of this functional for a distribution F is a minimizer of the expected score with respect to F. If the minimum is unique we call the scoring function strictly consistent and the corresponding functional *elicitable* (Lambert et al., 2008). Likewise, a scoring rule is (strictly) *proper* if the expected score with respect to F is (uniquely) minimized by F.

Consistency and propriety are desirable properties for forecast evaluation, since they can be used to ensure that rational and risk-neutral forecasters report truthfully: Faced with a payoff equal to the negative of their score, reporting their true belief maximizes their expected payoff. Hence, the interests of the forecasters are in line with the interests of the decision maker (Brier, 1950; McCarthy, 1956).

In probabilistic forecasting, i.e. when full predictive distributions are reported, a variety of strictly proper scoring rules are available (Gneiting and Raftery, 2007). Similarly, many statistical functionals, such as quantiles and expectiles, are elicitable with convenient characterizations of the corresponding classes of consistent scoring functions, see Gneiting (2011a) and the references therein. On the other hand, several widely considered characteristics fail to be elicitable, for instance the variance, the mode (Heinrich, 2014) and the prominent financial risk measure Expected Shortfall (ES) (Weber, 2006;

Gneiting, 2011a). As a consequence, the question which functionals are elicitable is a central problem in comparative forecast evaluation, with recent theoretical advances due to Lambert et al. (2008), Gneiting (2011a), and Steinwart et al. (2014) in the real-valued case and due to Frongillo and Kash (2015, 2020) and Fissler and Ziegel (2016) in the vector-valued case. This thesis addresses the question of elicitability for three special classes of statistical functionals, namely tail properties, interval forecasts, and characteristics of point processes.

Tail properties

Forecasts are particularly relevant for exceptional events such as natural disasters, or financial crashes, as these usually have severe consequences. Such events are often rare and observational data for statistical models is sparse. Hence, these circumstances call for forecast evaluation techniques that emphasize distribution tails rather than average behavior. Recent progress includes Friederichs and Thorarinsdottir (2012) who investigate the use of scoring rules for distribution classes central to extreme value theory, and Diks et al. (2011), Lerch et al. (2017), as well as Holzmann and Klar (2017) who consider weighted scoring rules for forecasts of distribution tails. An event-based approach to evaluate whether exceedances of high thresholds are predicted correctly is pursued by Stephenson et al. (2008) and Ferro and Stephenson (2011). Closely connected is the verification tool of Taillardat et al. (2019) which is based on the asymptotic behavior of the popular continuous ranked probability score (CRPS), conditional on high realizations.

In Chapter 2 we contribute to the problem of forecast evaluation for rare events by raising the fundamental question to what extent, and in which sense, statistical features of distribution tails are elicitable. To answer it, we introduce the class of max-functionals, which contains key characteristics from extreme value theory, e.g. the extreme value index. We show that under mild regularity assumptions its members fail to be elicitable in a very strong sense, which highlights limitations of scoring function-based forecast evaluation for tail properties.

Interval forecasts

A drawback of point forecasts is the loss of information when reducing a probability distribution to a real-number, e.g. mean or median. The simplest step towards a more informative forecast consists of reporting intervals, which provide an attractive and widely used way to convey information on the inherent uncertainty of the quantity of interest. In particular, one or multiple predictive intervals, which are designed to contain the observation with specified nominal probability, are requested implicitly or explicitly in a number of forecasting settings, including the Global Energy Forecasting Competition (Hong et al., 2016), the M4 Competition (Makridakis et al., 2020), the ongoing M5 Competition (M Open Forecasting Center, 2020), and the emerging COVID-19 Forecast Hub (Bracher et al., 2020). This highlights the need for sound evaluation methods which enable researchers and practitioners to compare interval forecasts and choose between different models for the generation of such intervals.

Early work on the evaluation problem for interval forecasts can be found in Aitchison and Dunsmore (1968), Winkler (1972), Casella et al. (1993), and Christoffersen (1998). Recently, Askanazi et al. (2018) have shown that many proposed scoring functions for intervals are (strictly) consistent for the equal-tailed interval, which lies between the $\frac{\alpha}{2}$- and $(1-\frac{\alpha}{2})$-quantiles. Hence, they are inappropriate for the shortest prediction interval. In Chapter 3 we address this issue and show that alternatives for the shortest interval are not readily available, as the shortest interval fails to be elicitable for many classes of distributions of practical interest. In contrast, consistent scoring functions for the equal-tailed interval are based on consistent scoring functions for quantiles and are thus well-understood. As a third, but conceptually different predictive interval, the modal interval admits a unique strictly consistent scoring function, up to equivalence. Our findings provide guidance in interval forecast evaluation and support recent choices of performance measures in forecast competitions.

Point process characteristics

In numerous forecasting settings the quantity of interest is not a one-dimensional number, but a complex point pattern in space and time, such that it can be modeled as a point process. Examples are abundant and range from quantitative criminology, which considers forecasts of increased criminal offenses in urban areas (Mohler et al., 2011; Flaxman et al., 2019) to epidemiology, which models when and where people catch diseases (Meyer and Held, 2014; Schoenberg et al., 2019). Moreover, fire departments monitor the sources and areas of wildfires (Peng et al., 2005; Xu and Schoenberg, 2011; Taylor et al., 2013) and statistical seismology quantifies properties of earthquakes such as time, epicenter, and magnitude (Bray and Schoenberg, 2013; Ogata, 2013). These applications demand reliable statistical methods for forecast evaluation and model selection in the point process framework.

Many existing approaches to model evaluation for point processes are due to applications in seismology, see e.g. Bray and Schoenberg (2013) for a review. In particular, the regional earthquake likelihood models (RELM) initiative (Field, 2007; Schorlemmer et al., 2007) set up testing centers to do prospective evaluation of seismological models. Bray and Schoenberg (2013) point out the connection between some of the used testing procedures and the scoring literature, by stating that “numerical tests such as the L-test, can be viewed as examples of scoring rules.” In Chapter 4 we make this connection explicit and derive scoring functions to compare point process models or forecasts. More precisely, we show that many widely used point process characteristics, such as the intensity or the product densities can be understood as elicitable statistical functionals. We illustrate the finite sample properties of the corresponding strictly consistent scoring functions via simulation experiments. These results offer a new and principled approach for the comparative assessment of forecasts and models, which encompasses several existing methods.

1 | Scoring functions and elicitability

This chapter fixes notation and definitions and presents some theoretical background concerning scoring functions and elicitability. It includes well-known results which we collect for the reader's convenience, but also new findings which play a key role in the subsequent chapters or are of independent interest. The chapter ends with a brief introduction to comparative forecast evaluation, the central application of consistent scoring functions in practice.

1.1 Notation and definitions

For notation and basic definitions we follow Gneiting (2011a) and Fissler and Ziegel (2016). Let Y be a random variable taking values in some *observation domain* O which is a subset of some real vector space. Typical examples include $\mathbb{R}^d$, the set $\mathbb{N}_0$, corresponding to count data, or the space of point patterns (see Chapter 4). The Borel σ-algebra of O is denoted via $\mathcal{O}$ and $\mathcal{F}$ is a collection of probability distributions on $(\mathsf{O}, \mathcal{O})$, which represents the possible distributions for Y. Whenever convenient, we identify probability distributions with their cumulative distribution functions (CDFs). The *action domain* A holds all possible reports of the forecasters on which the decision maker can act.

A *functional* will be a mapping which represents a statistical property of the distributions in $\mathcal{F}$. In the general framework, a functional is *set-valued*, i.e. given by $T : \mathcal{F} \to 2^{\mathsf{A}}$, where 2^{A} denotes the power set of A. The set $T(F) \subseteq \mathsf{A}$ then consists of all correct forecasts if $F \in \mathcal{F}$ is true. In the special case where $T(F)$ reduces to a single value $t \in \mathsf{A}$ for all $F \in \mathcal{F}$ we use the more convenient definition $T : \mathcal{F} \to \mathsf{A}$ and call T *single-valued.* All results on set-valued functionals naturally transfer to this simpler setting.

A measurable function $h : \mathsf{O} \to \mathbb{R}$ is called $\mathcal{F}$*-integrable* if it is integrable with respect to all $F \in \mathcal{F}$. Analogously, a function $g : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is called $\mathcal{F}$-integrable if for all $x \in \mathsf{A}$ the function $y \mapsto g(x, y)$ is integrable with respect to all $F \in \mathcal{F}$. We use the short notation

$$\bar{h}(F) := \int_{\mathsf{O}} h(y) \, \mathrm{d}F(y) \quad \text{and} \quad \bar{g}(x, F) := \int_{\mathsf{O}} g(x, y) \, \mathrm{d}F(y)$$

for $\mathcal{F}$-integrable functions h, g and $x \in \mathsf{A}$, $F \in \mathcal{F}$ and let $\mathbb{E}_F$ denote the expectation operator when Y has distribution $F \in \mathcal{F}$, such that $\mathbb{E}_F h(Y) = \bar{h}(F)$.

A *scoring function* is an $\mathcal{F}$-integrable mapping $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. The central concepts connecting scoring functions and statistical functionals are consistency and elicitability.

Definition 1.1 (Consistency)**.** A scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is *consistent* for a functional $T : \mathcal{F} \to 2^{\mathsf{A}}$ relative to the class $\mathcal{F}$ if

$$\bar{S}(t, F) \le \bar{S}(x, F) \tag{1.1}$$

for all $F \in \mathcal{F}$, $t \in T(F)$, and $x \in \mathsf{A}$. It is *strictly consistent* for T if it is consistent for T and the equality $\bar{S}(t, F) = \bar{S}(x, F)$ implies $x \in T(F)$ for all $F \in \mathcal{F}$ and $x \in \mathsf{A}$.

Definition 1.2 (Elicitability)**.** A functional $T : \mathcal{F} \to 2^{\mathsf{A}}$ is *elicitable* if there exists a scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ that is strictly consistent for T relative to $\mathcal{F}$.

If a forecaster faces a penalty $S(x, y)$ for a report x and an outcome y, consistency of the scoring function S for the functional T ensures that any member of the forecaster's set of true beliefs $T(F)$ minimizes the expected penalty. Strict consistency ensures that *only* the values in $T(F)$ are minimizers, i.e. deviating from the truth leads to a higher expected penalty.

Although this work introduces consistency and elicitability from the perspective of forecast evaluation, both concepts are useful in other areas of statistics, too. They enable regression, e.g. quantile and expectile regression (Koenker, 2005; Newey and Powell, 1987), M-estimation (Huber and Ronchetti, 2009), and are central to various machine learning algorithms (Steinwart et al., 2014; Frongillo and Kash, 2020). In order to simplify the presentation we stick with the decision theoretic perspective for the rest of the thesis.

Since the ordering in (1.1) is not affected by scaling S with a positive constant or adding a report-independent function, we use the following definition (Gneiting, 2011a).

Definition 1.3 (Equivalence)**.** Let $S', S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be two scoring functions. We call S' *equivalent* to S if

$$S'(x, y) = cS(x, y) + h(y)$$

for some $c > 0$ and an $\mathcal{F}$-integrable function $h : \mathsf{O} \to \mathbb{R}$.

1.2 Basic results and examples

This section collects some examples of elicitable functionals and then presents some standard results on consistency and elicitability for later reference. For a deeper theoretical introduction see e.g. Gneiting (2011a), Fissler and Ziegel (2016), and Frongillo and Kash (2019, 2020).

Example 1.4 (Quantiles)**.** Let $\mathsf{O} = \mathsf{A} = \mathbb{R}$. For $\alpha \in (0, 1)$ an α-quantile of F is a point $x \in \mathbb{R}$ that satisfies $F(x-) \le \alpha \le F(x)$, where $F(x-) := \lim_{y \uparrow x} F(y)$ denotes the left-hand limit of F at x. The α-quantile functional $T_\alpha(F) := \{x \mid F(x-) \le \alpha \le F(x)\}$ is set-valued, and it is elicitable relative to any class $\mathcal{F}$. The strictly consistent scoring functions are equivalent to

$$S_\alpha(x, y) = (\mathbb{1}(y \le x) - \alpha)\,(g(x) - g(y))\,, \tag{1.2}$$

where g is $\mathcal{F}$-integrable and strictly increasing, see Gneiting (2011a,b) and references therein. ◇

Example 1.5 (Expectations)**.** Let $\mathsf{O} = \mathbb{R}$ and $\mathcal{F}_2$ be the class of distributions with finite second moments. It is well-known that if S is the quadratic loss $S(x, y) = (x - y)^2$, then the expected score function $x \mapsto \bar{S}(x, F)$ is minimized by the expectation $\mathbb{E}_F Y$ for all $F \in \mathcal{F}_2$. Stated differently, the single-valued functional $T(F) := \mathbb{E}_F Y$ is elicitable and S is strictly consistent for T relative to $\mathcal{F}_2$. $\diamond$

A more general result is that expectations of integrable functionals are always elicitable, i.e. finite second moments as in the previous example are not needed. To make this precise, let $\mathsf{A}, \mathsf{O} \subseteq \mathbb{R}^k$ and denote the subderivative (Rockafellar, 1970) of a convex function $f : \mathsf{A} \to \mathbb{R}^k$ at $x \in \mathbb{R}^k$ by $\nabla f(x)$. Then the function

$$b : \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \quad (x, y) \mapsto -f(x) - \nabla f(x)^\top (y - x), \tag{1.3}$$

is called a *Bregman function* for f. If f is strictly convex, we call b *strictly consistent.* Using these definitions we can formulate the following well-known result, see Savage (1971), Gneiting (2011a, Theorem 7), and Frongillo and Kash (2015, Theorem 13).

Theorem 1.6 (Elicitability of expectations)**.** *If $h : \mathsf{O} \to \mathbb{R}^k$ is $\mathcal{F}$-integrable, then the functional $T : \mathcal{F} \to \mathbb{R}^k$ defined via*

$$T(F) = \big(\bar{h}_1(F), \ldots, \bar{h}_k(F)\big)^\top$$

is elicitable and consistent scoring functions $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ are given by $S(x, y) = b(x, h(y))$, where b is a Bregman function. If b is strictly consistent, then S is strictly consistent for T.

Although there are many elicitable functionals beyond quantiles and expectations, there are also several examples which do not allow for consistent scoring functions. The simplest approach to check whether a functional is elicitable, is to study its values for convex combinations of distributions. The following proposition states the classical convex level sets (CxLS) result (Osband (1985, Proposition 2.5) and Gneiting (2011a, Theorem 6)) together with the refined CxLS* property of Fissler et al. (2020, Proposition 3.3).

Proposition 1.7 (convex level sets)**.** *Let $T : \mathcal{F} \to 2^{\mathsf{A}}$ be an elicitable functional. If $F_0, F_1 \in \mathcal{F}$ and $\lambda \in (0, 1)$ are such that $F_\lambda = \lambda F_1 + (1 - \lambda) F_0 \in \mathcal{F}$, then*

(i) $T(F_0) \cap T(F_1) \subseteq T(F_\lambda)$ (CxLS property)

(ii) $T(F_0) \cap T(F_1) \neq \emptyset \implies T(F_0) \cap T(F_1) = T(F_\lambda)$ (CxLS property)*

If T is a single-valued functional, the properties coincide and are simply referred to as CxLS. Under certain regularity conditions, convex level sets are also sufficient for elicitability, as demonstrated by Lambert et al. (2008) for distributions on finite sets and by Steinwart et al. (2014) for continuous densities on compact metric spaces. The most relevant examples of functionals that do not have convex level sets and thus fail to be elicitable, are the variance and the quantitative risk measure Expected Shortfall (ES), see Weber (2006) and Gneiting (2011a). However, even if a functional does not have CxLS this can often be overcome by pairing it with another functional such that the resulting vector becomes elicitable. This is demonstrated in Fissler and Ziegel (2016) for ES together with the risk measure Value at Risk (VaR). The next example collects the corresponding arguments for the variance.

Example 1.8 (Variance)**.** Let $\mathcal{F}_2$ be the class of distributions on $\mathbb{R}$ having finite second moments and set $T_k(F) := \mathbb{E}_F Y^k$ for all $k \in \mathbb{N}$. A simple construction shows that the variance functional $T_{\text{var}}(F) = T_2(F) - T_1(F)^2$ does not satisfy the CxLS property on any class $\mathcal{F} \subseteq \mathcal{F}_2$ on which T_1 is not constant. By Proposition 1.7 it thus fails to be elicitable relative to such classes. This issue can be addressed by considering the functional $T := (T_1, T_{\text{var}})^\top$ which connects to the functional $T' := (T_1, T_2)^\top$ via a bijection. Since T' consists of first and second moment only, it is elicitable by Theorem 1.6, and elicitability of T follows from the next Proposition. $\diamond$

The following result formalizes how bijective mappings ensure elicitability, see Osband (1985) and Gneiting (2011a, Theorem 4).

Proposition 1.9 (Revelation principle)**.** *Let A, A' be some sets and $g : \mathsf{A} \to \mathsf{A}'$ a bijection with inverse g^{-1}. Let $T : \mathcal{F} \to 2^{\mathsf{A}}$ and $T_g : \mathcal{F} \to 2^{\mathsf{A}'}$ defined via $T_g(F) := g(T(F))$ be functionals. Then T is elicitable if and only if T_g is elicitable. A function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is a strictly $\mathcal{F}$-consistent scoring function for T if and only if $S_g : \mathsf{A}' \times \mathsf{O} \to \mathbb{R}$, $(x, y) \mapsto S_g(x, y) := S(g^{-1}(x), y)$ is a strictly $\mathcal{F}$-consistent scoring function for T_g.*

Analogous to changes of the action domain, there is a straightforward connection between consistency and the observation domain. It resembles the findings on weighted functionals as discussed in Gneiting and Ranjan (2011) and Gneiting (2011a, Theorem 5).

Proposition 1.10 (Transformation principle)**.** *Let $g : \mathsf{O}' \to \mathsf{O}$ be a measurable mapping and $\mathcal{F}'$ a set of distributions on O' and define $g(\mathcal{F}') := \{F \circ g^{-1} \mid F \in \mathcal{F}'\}$. If there is an elicitable functional $T : g(\mathcal{F}') \to 2^{\mathsf{A}}$, then the functional T' defined via $T'(F) := T(F \circ g^{-1})$ is elicitable and (strictly) consistent scoring functions for T' are given by $S'(x, y) = S(x, g(y))$, where S is a (strictly) consistent scoring function for T.*

1.3 Structural results on elicitable functionals

This section presents two general results which illustrate characteristics of elicitable functionals by specifying which functionals *fail* to be elicitable. The first finding relies on convex combinations, similar to the CxLS properties (see Proposition 1.7), and its idea will be central for various results in the chapters below. The second result considers functionals which are invariant under certain transformations, e.g. translations, and is of independent interest. Lastly, we connect to continuity properties of elicitable functionals.

1.3.1 Convex combinations

As discussed in the previous section, the CxLS properties are central tools to show non-elicitability in many cases. An important example of a functional which fails to be elicitable, even though it has the CxLS* property, is the mode if it is defined for absolutely continuous distributions, see Heinrich (2014). We state a refined result on the behavior of elicitable functionals on convex combinations of distributions.

Theorem 1.11. *Let $T : \mathcal{F} \to 2^{\mathsf{A}}$ be a functional, and let $F_0, F_1 \in \mathcal{F}$ be such that $F_\lambda = \lambda F_1 + (1 - \lambda) F_0 \in \mathcal{F}$ for all $\lambda \in (0, 1)$. If there are $t_0 \in T(F_0) \backslash T(F_1)$ and*

$t_1 \in T(F_1)\backslash T(F_0)$ *such that for every* $\lambda \in (0,1)$ *it holds that either* $t_0 \in T(F_\lambda)$ *and* $t_1 \notin T(F_\lambda)$*, or* $t_1 \in T(F_\lambda)$ *and* $t_0 \notin T(F_\lambda)$*, then* T *is not elicitable.*

Proof. Let t_0, t_1 be as stated and set $F_\lambda := \lambda F_1 + (1-\lambda)F_0$. Suppose that S is a strictly consistent scoring function for T. Linearity of expectations yields

$$\bar{S}(t_0, F_\lambda) - \bar{S}(t_1, F_\lambda) = \lambda \left[\bar{S}(t_0, F_1) - \bar{S}(t_1, F_1)\right] + (1-\lambda) \left[\bar{S}(t_0, F_0) - \bar{S}(t_1, F_0)\right],$$

where the first difference is positive, while the second is negative. Hence, $\bar{S}(t_0, F_\lambda) = \bar{S}(t_1, F_\lambda)$ for some $\lambda \in (0,1)$. Since either $t_0 \in T(F_\lambda)$ and $t_1 \notin T(F_\lambda)$, or $t_1 \in T(F_\lambda)$ and $t_0 \notin T(F_\lambda)$, we arrive at a contradiction. □

The assertion of Theorem 1.11 overlaps with part (ii) of Proposition 1.7 in the sense that if $T(F_0) \cap T(F_1) \neq \emptyset$ and the conditions of Theorem 1.11 hold, then T cannot have the CxLS* property and thus fails to be elicitable. If $T(F_0) \cap T(F_1) = \emptyset$, Theorem 1.11 provides a novel result, since Proposition 1.7(ii) does not address this situation.

For single-valued functionals $T : \mathcal{F} \to \mathsf{A}$ Theorem 1.11 simplifies: If there exist $F_0, F_1 \in \mathcal{F}$ such that $T(F_0) \neq T(F_1)$ and

$$T(\lambda F_1 + (1-\lambda)F_0) \in \{T(F_0), T(F_1)\} \quad \text{for all } \lambda \in (0,1),$$

then T is not elicitable. This formulation illustrates that, loosely speaking, single-valued elicitable functionals cannot be piecewise constant on convex combinations of distributions. The following result for single-valued functionals is a simple consequence of this observation.

Corollary 1.12. *Let* $\mathcal{F}$ *be convex. If* $T : \mathcal{F} \to \mathsf{A}$ *is a non-constant finite-valued functional, then it is not elicitable.*

Proof. For $F_0, F_1 \in \mathcal{F}$ define $F_\lambda := \lambda F_1 + (1-\lambda)F_0$, set $t_0 := T(F_0)$ and $t_1 := T(F_1)$ and let $t_0 \neq t_1$. By assumption, there are some $t_2, \ldots, t_n \in \mathsf{A}$ such that $T(F_\lambda) \in \{t_0, t_1, \ldots, t_n\}$ for all $\lambda \in [0,1]$. If there is an $i \in \{0, \ldots, n\}$ such that the set $\{\lambda \in [0,1] \mid T(F_\lambda) = t_i\}$ is not an interval, then Proposition 1.7 implies that T is not elicitable. However, if they are intervals, we can find a new convex combination of distributions $F_0', F_1' \in \mathcal{F}$ such that Theorem 1.11 is applicable. □

Connection to identifiability Corollary 1.12 can be interpreted as an analogon to the statement in Frongillo and Kash (2020) that '*no nonconstant finite property is identifiable*'. Following Gneiting (2011a) and Fissler and Ziegel (2016) a function $V : \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ is a *strict identification function* for $T : \mathcal{F} \to 2^{\mathsf{A}}$ relative to the class $\mathcal{F}$ if it is $\mathcal{F}$-integrable and

$$\bar{V}(x, F) = 0 \quad \Longleftrightarrow \quad x \in T(F)$$

for all $x \in \mathsf{A}$ and $F \in \mathcal{F}$. The functional T is called *identifiable* if there exists a strict identification function for it. Identification functions can be interpreted as derivatives of scoring functions, a connection which is called *Osband's principle*, see Gneiting (2011a) and Fissler and Ziegel (2016) for precise results. The following is an analogon to Theorem 1.11 for identifiability.

Theorem 1.13. *Let $T : \mathcal{F} \to 2^{\mathsf{A}}$ be a functional and let $F_0, F_1 \in \mathcal{F}$ be such that there is a $t_0 \in T(F_0) \backslash T(F_1)$. If there is a $\lambda \in (0,1)$ such that $F_\lambda = \lambda F_1 + (1-\lambda)F_0 \in \mathcal{F}$ and $t_0 \in T(F_\lambda)$, then T is not identifiable.*

Proof. Let $\lambda \in (0,1)$ be as in the theorem and suppose that V is a strict identification function for T. Linearity of expectations and $t_0 \in T(F_\lambda)$ yield

$$0 = \bar{V}(t_0, F_\lambda) = \lambda \bar{V}(t_0, F_1) + (1-\lambda)\bar{V}(t_0, F_0) = \lambda \bar{V}(t_0, F_1),$$

which is a contradiction to $t_0 \notin T(F_1)$. □

A version of this theorem for single-valued functionals is presented in Fissler and Ziegel (2019b, Lemma B.1). The same technique is used in Dearborn and Frongillo (2020, Lemma 1) to show that the mode functional is not identifiable, a result which naturally extends to the modal interval, see also Section 3.5 and Theorem 3.10.

1.3.2 Invariant functionals

This section investigates functionals which are invariant under some transformations of the distribution. The starting point of these considerations is the special case of translation invariance. For a real-valued distribution function F and $z \in \mathbb{R}$ we define the translated distribution via $F_z(x) := F(x - z)$ and assume that the class $\mathcal{F}$ contains F_z for all $F \in \mathcal{F}$ and $z \in \mathbb{R}$. Then a functional $T : \mathcal{F} \to 2^{\mathsf{A}}$ is *translation invariant* if $T(F_z) = T(F)$ for all $z \in \mathbb{R}$ and $F \in \mathcal{F}$.

In order to generalize, let $\mathcal{F}$ be a class of distributions on an observation space $\mathsf{O} \subseteq \mathbb{R}^k$ and $g : \mathsf{O} \times \mathsf{O} \to \mathsf{O}$ a measurable function. The following definition formalizes a notion of invariance with respect to g.

Definition 1.14 (g-invariance)**.** The class $\mathcal{F}$ is *closed under g-transformations* if for all $z \in \mathsf{O}$ and $F \in \mathcal{F}$ the transformed distribution

$$F_{g(z,\cdot)}(x) := \int \mathbb{1}(g_1(z,y) \le x_1, \ldots, g_k(z,y) \le x_k) \, \mathrm{d}F(y) \tag{1.4}$$

is in $\mathcal{F}$. If $\mathcal{F}$ is closed under g-transformations, then a functional $T : \mathcal{F} \to 2^{\mathsf{A}}$ is *g-invariant* if $T(F_{g(z,\cdot)}) = T(F)$ for all $z \in \mathsf{O}$ and $F \in \mathcal{F}$.

We recover translation invariance by setting $g(z,y) = z + y$. The next result shows that non-constant g-invariant functionals cannot be elicitable if g is symmetric, i.e. $g(z,y) = g(y,z)$ for all $y, z \in \mathsf{O}$. A technical condition is necessary.

Condition 1.15. The scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is strictly consistent for $T : \mathcal{F} \to 2^{\mathsf{A}}$ relative to $\mathcal{F}$ and such that $(z,y) \mapsto S(x, g(z,y))$ is $(F \otimes G)$-integrable for all $x \in \mathsf{A}$ and $F, G \in \mathcal{F}$.

Theorem 1.16. *Let $g : \mathsf{O} \times \mathsf{O} \to \mathsf{O}$ be symmetric and measurable and let $\mathcal{F}$ be closed under g-transformations. If the functional $T : \mathcal{F} \to 2^{\mathsf{A}}$ is g-invariant and has a scoring function which satisfies Condition 1.15, then it is constant.*

Proof. Let $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a scoring function as specified in Condition 1.15. Strict consistency for T and g-invariance give

$$T(F) = T(F_{g(z,\cdot)}) = \arg\min_{x \in \mathsf{A}} \bar{S}(x, F_{g(z,\cdot)}) = \arg\min_{x \in \mathsf{A}} \int S(x, g(z,y))\,\mathrm{d}F(y),$$

for all $z \in \mathbb{R}$ and $F \in \mathcal{F}$. This implies that for all $z \in \mathsf{O}$ the function $S_z(x,y) := S(x, g(z,y))$ is a strictly consistent scoring function for T and due to Gneiting (2011a, Theorem 2), the same holds true for

$$S_F(x,y) := \int S(x, g(z,y))\,\mathrm{d}F(z), \tag{1.5}$$

where $F \in \mathcal{F}$. For any choice of $F, G \in \mathcal{F}$ Fubini's theorem and the symmetry of g now give

$$\begin{aligned} \bar{S}_G(x,F) &= \int\int S(x, g(z,y))\,\mathrm{d}G(z)\,\mathrm{d}F(y) \\ &= \int\int S(x, g(y,z))\,\mathrm{d}F(y)\,\mathrm{d}G(z) = \bar{S}_F(x,G) \end{aligned}$$

for all $x \in \mathsf{A}$. This yields $T(F) = T(G)$ for all $F, G \in \mathcal{F}$, so T is constant. □

Via the choice $g(z,x) = z + x$ in Theorem 1.16 we obtain that, subject to a technical regularity assumption, non-constant translation invariant functionals cannot be elicitable. Additionally, the choice $\mathsf{O} = (0,\infty)$ and $g(z,x) = zx$ shows that non-constant *scale invariant* functionals fail to be elicitable, too, if they are defined for strictly positive random variables. Other possible choices are $g(x,y) := \max(x,y)$ or $g(x,y) := \min(x,y)$ on $\mathsf{O} = \mathbb{R}$ which yield non-elicitability for some exceptional properties which only depend on the limiting behavior of the distribution for $|y| \to \infty$. Via these choices of g, Theorem 1.16 recovers several known results from the literature, for instance the non-elicitability of the variance (see Example 1.8 and Gneiting (2011a)) or, more generally, the non-elicitability of all forms of centered expectations $T(F) = \mathbb{E}_F h(Y - \mathbb{E}_F Y)$, e.g. centered moments. Moreover, Theorem 1.16 shows that certain tail properties cannot be elicitable, however, other techniques are better suited to address this problem, see Chapter 2.

It is well known that the variance is jointly elicitable with the mean, see also Example 1.8. An interesting further question is which other invariant functionals are jointly elicitable with an elicitable functional (e.g. the mean in case of translation invariance) which contains information on the transformation $g(z,\cdot)$ (e.g. the translation) of the distribution. The technique used in Theorem 1.16 does not seem suited to answer this question.

From a technical perspective, it is also possible to proceed in the proof of Theorem 1.16 without requiring g to be symmetric. We then obtain for any two independent random variables Y and Z, having distributions F and G, respectively, that the distribution of $g(Z,Y)$ has the same functional value as F. The functional T must thus be constant on some subset of the class $\mathcal{F}$, but how this subset depends on g is not obvious and possibly very complex.

Finally, we remark that Fissler and Ziegel (2019b) consider invariance as well, however, in their approach, invariance is a property of the consistent scoring function and not of the functional. They consider elicitable functionals which are π-equivariant, i.e. they satisfy $T(\mathcal{L}(\varphi(Y))) = (\pi\varphi)(T(\mathcal{L}(Y)))$ for all random variables Y, where $\mathcal{L}(Y)$ denotes the law of Y and φ, π are some transformations. This definition incorporates our notion of g-invariance by choosing $\varphi \in \Phi$, where $\Phi := \{g(z,\cdot) \mid z \in \mathsf{O}\}$ is a set of transformations and $\pi : \Phi \to \{\mathrm{id}_{2^\mathsf{A}}\}$ maps any φ on the identity of 2^A.

1.3.3 Connections to continuity

In this subsection we connect the non-elicitability result of Theorem 1.11 to continuity of single-valued functionals. A possible interpretation is that functionals have to be continuous in some sense in order to be elicitable. In general the connection between elicitability and continuity is intricate since there is no obvious concept of continuity for functionals. Possible choices are made in the characterization results of Lambert et al. (2008) and Steinwart et al. (2014). Here we follow Bellini and Bignozzi (2015) and focus on *mixture-continuity* of single-valued functionals $T : \mathcal{F} \to \mathsf{A}$, where A is a subset of some metric space (M, d). Throughout, we assume $\mathcal{F}$ to be convex.

Definition 1.17 (Mixture-continuity). A functional $T : \mathcal{F} \to \mathsf{A}$ is *mixture-continuous* if for all $F_0, F_1 \in \mathcal{F}$ the mapping

$$[0,1] \to \mathsf{A}, \quad \lambda \mapsto T(\lambda F_1 + (1-\lambda)F_0)$$

is a continuous function.

Many statistical properties are mixture-continuous, e.g. ratios of expectations, quantiles, and expectiles, see Fissler and Ziegel (2019b) for details. Fissler and Ziegel (2019b, Proposition 2.2) and Bellini and Bignozzi (2015, Proposition 3.4) show that under weak assumptions, an elicitable functional is mixture-continuous if it has a strictly consistent scoring function for which the expected score function $x \mapsto \bar{S}(x, F)$ is continuous for all $F \in \mathcal{F}$.

Bayes risk Motivated by Frongillo and Kash (2020) we call $T_e : \mathcal{F} \to \mathsf{A}$ a *Bayes risk functional* if it can be represented via $T_e(F) := \bar{S}(T(F), F)$ for a functional $T : \mathcal{F} \to \mathsf{A}$ and a scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ which is consistent for T relative to $\mathcal{F}$. Examples of Bayes risk functionals include the variance (see Example 1.8) and Expected Shortfall (ES). Although they fail to be elicitable in general, the vector valued property $(T_e, T)^\top$ becomes elicitable whenever T is, see Frongillo and Kash (2020, Theorem 1). However, we do not require elicitability for the following result.

Proposition 1.18. *The Bayes risk functional T_e is mixture-continuous.*

Proof. Let $F_0, F_1 \in \mathcal{F}$ and define $F_\lambda := \lambda F_1 + (1-\lambda)F_0$ for $\lambda \in [0,1]$. Let S be a consistent scoring function for the functional T in the representation $T_e(F) = \bar{S}(T(F), F)$. Consistency implies that the mapping $\lambda \mapsto T_e(F_\lambda)$ is concave and thus also continuous on $(0,1)$. It remains to consider the points $\{0,1\}$ and due to symmetry, it suffices to consider $\lambda \downarrow 0$ only. The first term on the right-hand side of

$$T_e(F_\lambda) = \lambda \bar{S}(T(F_\lambda), F_1) + (1-\lambda)\bar{S}(T(F_\lambda), F_0)$$

is bounded since consistency and Nau's inequality (Nau, 1985) give $\bar{S}(T(F_1), F_1) \leq \bar{S}(T(F_\lambda), F_1) \leq \bar{S}(T(F_0), F_1)$. Since $(F, y) \mapsto S(T(F), y)$ defines a proper scoring rule by Lemma 1.22, we obtain $T_e(F_\lambda) \to T_e(F_0)$ for $\lambda \downarrow 0$ from Lemma 1.24. □

This proposition extends mixture-continuity to the class of Bayes risk functionals, see Frongillo and Kash (2020) for a collection of examples. Remarkably, elicitability of T is not needed for this result.

Self-calibration We now show mixture-continuity for functionals which allow for self-calibration. This property of scoring functions is discussed in Fissler and Ziegel (2019b) from where we also take the definition.

Definition 1.19 (Self-calibration)**.** A scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is *$\mathcal{F}$-self-calibrated* for a functional $T : \mathcal{F} \to \mathsf{A}$ with respect to the metric d if for all $\varepsilon > 0$ and all $F \in \mathcal{F}$ there is a $\delta = \delta(\varepsilon, F) > 0$ such that for all $x \in \mathsf{A}$ and $t = T(F)$

$$\bar{S}(x, F) - \bar{S}(t, F) < \delta \quad \Rightarrow \quad d(x, t) < \varepsilon. \tag{1.6}$$

As stated in Fissler and Ziegel (2019b), self-calibration can be interpreted as continuity of the inverse of the expected score, which illustrates why it can be used to ensure convergence in M-estimation (Fissler and Ziegel, 2019b, Theorem 2.9). In particular, it implies that S is strictly consistent for T.

Proposition 1.20. *If the functional $T : \mathcal{F} \to \mathsf{A}$ has an $\mathcal{F}$-self-calibrated consistent scoring function, then it is mixture-continuous.*

Proof. Let $F_0, F_1 \in \mathcal{F}$, define $F_\lambda := \lambda F_1 + (1-\lambda) F_0$ and consider continuity at $\lambda = 0$ first. Let $\varepsilon > 0$ be given and set $x = T(F_\lambda)$ in (1.6). Due to Lemma 1.22 and 1.24 we obtain that for all $\delta = \delta(\varepsilon, F_0) > 0$ we can find a δ' such that $\bar{S}(T(F_\lambda), F_0) - \bar{S}(T(F_0), F_0) < \delta$ for all $\lambda \in [0, \delta']$. Self-calibration of S then implies $d(T(F_\lambda), T(F_0)) < \varepsilon$ for all $\lambda \in [0, \delta']$. Continuity in $\lambda = 1$ follows by symmetry and the case $\lambda \in (0, 1)$ follows via a re-parametrization argument: Consider some $\bar{\lambda} \in (0, 1)$ and define $F'_\mu := \mu F_1 + (1 - \mu) F_{\bar{\lambda}}$ with the re-parametrization $\mu = (\lambda - \bar{\lambda})/(1 - \bar{\lambda})$ for $\lambda \in [\bar{\lambda}, 1]$ and $\mu \in [0, 1]$. Then

$$\lim_{\lambda \downarrow \bar{\lambda}} T(F_\lambda) = \lim_{\mu \downarrow 0} T(F'_\mu) = T(F'_0) = T(F_{\bar{\lambda}})$$

and the same argument for $\lambda \uparrow \bar{\lambda}$ finishes the proof. □

This result is roughly the converse statement to Fissler and Ziegel (2019b, Proposition 2.8), which states that a mixture-continuous elicitable functional which admits continuous expected scores $\bar{S}(\cdot, F)$ must have a self-calibrated scoring function. Stated differently, we see that only mixture-continuous elicitable functionals can have self-calibrated scoring functions.

1.4 Proper scoring rules

In probabilistic forecasting, the whole distribution function instead of some statistical property is reported to the decision maker, i.e. the action space A is given by $\mathcal{F}$. Analogously to a scoring function, a *scoring rule* then assigns a score based on the reported distribution F and a realizing observation y. For recent reviews of the theory and application of proper scoring rules we refer to Dawid (2007), Gneiting and Raftery (2007), and Dawid and Musio (2014).

This section gives a short introduction to proper scoring rules for later reference and then turns to a novel construction principle. Our notation follows Gneiting and Raftery (2007), in particular we let $\bar{\mathbb{R}} := [-\infty, \infty]$ be the extended real line.

An extended real-valued function $S : \mathcal{F} \times \mathsf{O} \to \bar{\mathbb{R}}$ is a *scoring rule* if for all $F, G \in \mathcal{F}$ the integral $\bar{S}(F, G)$ is well-defined. The analogue to the concept of consistency (Definition 1.1) is usually called propriety.

Definition 1.21 (Propriety)**.** A scoring rule $S : \mathcal{F} \times \mathsf{O} \to \bar{\mathbb{R}}$ is *proper* if $\bar{S}(G, G) \leq \bar{S}(F, G)$ holds for all $F, G \in \mathcal{F}$. It is *strictly proper* if it is proper and for any $F, G \in \mathcal{F}$ the equality $\bar{S}(G, G) = \bar{S}(F, G)$ implies $G = F$.

In contrast to the definition of scoring functions, we merely require quasi-integrability for scoring rules such that expected scores can take values in $\bar{\mathbb{R}}$. Following Gneiting and Raftery (2007, Definition 1) we impose some mild restrictions by considering *regular* scoring rules only, i.e. S which satisfy $\bar{S}(F, G) > -\infty$ and $\bar{S}(G, G) \in \mathbb{R}$ for all $F, G \in \mathcal{F}$.

Proper scoring rules connect naturally to consistent scoring functions, as the following result (Gneiting, 2011a, Theorem 3) illustrates.

Lemma 1.22. *Let $T : \mathcal{F} \to \mathsf{A}$ be a functional and $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ a scoring function. If S is consistent for T, then the scoring rule $R : \mathcal{F} \times \mathsf{O} \to \mathbb{R}$ given by $R(F, y) := S(T(F), y)$ is proper.*

Note that this result cannot guarantee strict propriety, even if S is strictly consistent for T. We now introduce a continuity property of proper scoring rules, which is closely connected to the concept of mixture-continuity for functionals, see Definition 1.17. It is a useful tool for new results on functionals (Subsection 1.3.3) and proper scoring rules (Section 2.4).

Definition 1.23 (Diagonal-continuity)**.** Let $\mathcal{F}$ be convex. A scoring rule $S : \mathcal{F} \times \mathsf{O} \to \bar{\mathbb{R}}$ is *diagonal-continuous at G* if for all $F \in \mathcal{F}$

$$\bar{S}(\lambda F + (1 - \lambda)G, G) \to \bar{S}(G, G) \qquad \text{for } \lambda \downarrow 0.$$

Lemma 1.24. *Let $\mathcal{F}$ be convex. If $S : \mathcal{F} \times \mathsf{O} \to \bar{\mathbb{R}}$ is an $\mathcal{F}$-integrable proper scoring rule, then it is diagonal-continuous at each $G \in \mathcal{F}$.*

Proof. We proceed similar to the proof of Nau (1985, Proposition 3). Let $F, G \in \mathcal{F}$ and denote $F_\lambda := \lambda F + (1 - \lambda)G$ for $\lambda \in [0, 1)$. We obtain the inequality

$$\begin{aligned}(1 - \lambda)\bar{S}(F_\lambda, G) &= \bar{S}(F_\lambda, F_\lambda) - \lambda \bar{S}(F_\lambda, F) \\ &\leq \bar{S}(G, F_\lambda) - \lambda \bar{S}(F, F) \\ &= (1 - \lambda)\bar{S}(G, G) + \lambda\big(\bar{S}(G, F) - \bar{S}(F, F)\big),\end{aligned}$$

since S is a proper scoring rule. Rearranging leads to

$$|\bar{S}(\lambda F + (1 - \lambda)G, G) - \bar{S}(G, G)| \leq \frac{\lambda}{1 - \lambda}\left(\bar{S}(G, F) - \bar{S}(F, F)\right)$$

for $\lambda \in [0, 1)$ and the right hand side of this equation vanishes as $\lambda \downarrow 0$. □

The argument in the proof of Lemma 1.24 can be extended to all regular proper scoring rules as long as the expected score $\bar{S}(G,F)$ is finite.

1.4.1 Common choices of proper scoring rules

Various proper scoring rules have been proposed in the literature and this subsection collects some common choices for later reference. We start with proper scoring rules for densities and assume that μ is a σ-finite measure on $(\mathsf{O},\mathcal{O})$. The classes $\mathcal{L}_\alpha$ for $\alpha > 1$ are defined to contain all densities f of probability measures $F \in \mathcal{F}$ that are absolutely continuous with respect to μ and such that

$$\|f\|_\alpha := \left(\int_{\mathsf{O}} f(y)^\alpha \,\mathrm{d}\mu(y)\right)^{1/\alpha}$$

is finite. Moreover, for all $k \in \mathbb{N}$ we let $\mathcal{F}_k$ be the classes of distributions with finite k-th moment.

Logarithmic score The logarithmic score (Good, 1952) is one of the most common scoring rules since it connects to various fundamental statistical concepts, such as maximum-likelihood estimation, information criteria, or Bayes factors (Gneiting and Raftery, 2007). It is defined via

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

and it is strictly proper on $\mathcal{L}_1$. It is the central example of a scoring rule which takes values in $\bar{\mathbb{R}}$.

Pseudospherical score For any $\alpha > 1$ the pseudospherical score (Gneiting and Raftery, 2007) is defined via

$$\mathrm{PseudoS}(f,y) := -f(y)^{\alpha-1}/\|f\|_\alpha^{\alpha-1}$$

and it is strictly proper on $\mathcal{L}_\alpha$. After appropriate scaling the pseudospherical score converges to the logarithmic score as $\alpha \to 1$, see Gneiting and Raftery (2007) for details.

Hyvärinen score Let $\mathsf{O} = \mathbb{R}^d$ and let ∇ denote the gradient and Δ the Laplace operator. Define $\mathcal{L}^*$ as the class of densities on O which are twice differentiable, positive almost everywhere, and such that $\nabla \log(f(y))g(y) \to 0$ as $\|y\| \to \infty$ for all $f,g \in \mathcal{L}^*$. Then the Hyvärinen score (Hyvärinen, 2005) given by

$$\mathrm{HyvS}(f,y) := \Delta \log f(y) + \frac{1}{2}\|\nabla \log f(y)\|^2$$

is a strictly proper scoring rule on $\mathcal{L}^*$ if it is $\mathcal{L}^*$-integrable. The Hyvärinen score has the remarkable property that it is 0-homogeneous, i.e. $\mathrm{HyvS}(cf,y) = \mathrm{HyvS}(f,y)$ for all $c > 0$. It can thus be used in situations where only unnormalized models are available, see Hyvärinen (2005), Dawid et al. (2012), and Ehm and Gneiting (2012) for details.

Continuous ranked probability score (CRPS) For $\mathsf{O} = \mathbb{R}$ a popular choice is the continuous ranked probability score (CRPS) (Matheson and Winkler, 1976) defined via

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

for a cumulative distribution function (CDF) $F \in \mathcal{F}$. The CRPS is proper and it is even strictly proper on $\mathcal{F}_1$. Various modifications of the CRPS are available, in particular some regions of the observation domain can be emphasized in the evaluation: For a weight function $w : \mathbb{R} \to [0, \infty)$ the weighted CRPS (wCRPS) (Gneiting and Ranjan, 2011) is defined via

$$\mathrm{wCRPS}(F, y) := \int_{-\infty}^{\infty} w(x)(F(x) - \mathbb{1}(y \leq x))^2 \,\mathrm{d}x$$

and it is proper, but usually not strictly proper without additional assumptions, see Gneiting and Ranjan (2011) and Holzmann and Klar (2017).

Dawid-Sebastiani score For $\mathsf{O} = \mathbb{R}^d$ the Dawid-Sebastiani (DS) score (Dawid and Sebastiani, 1999) is defined via

$$\mathrm{DSS}(F, y) := \log\det\Sigma_F + (y - \mu_F)^\top \Sigma_F^{-1}(y - \mu_F)$$

on $\mathcal{F}_2$, where μ_F and Σ_F are the mean and the covariance matrix of the predictive distribution $F \in \mathcal{F}_2$. The DS score is proper, but not strictly proper, as distributions with the same first and second moments attain the same score.

Binary scoring rules If Y is a random variable taking values in $\{0, 1\}$, then we can identify the class $\mathcal{F}$ with the interval $[0, 1]$ via the success probability $p = \mathbb{P}_F(Y = 1)$. In this setting, there is a great variety of (strictly) proper scoring rules, which can be obtained via the mixture representation in Gneiting and Raftery (2007, Theorem 3). Common choices are the quadratic or Brier score (Brier, 1950) and the logarithmic score, defined via

$$S(p, y) = (p - y)^2 \qquad \text{and} \qquad S(p, y) = -y\log(p) - (1 - y)\log(1 - p),$$

respectively. Both are strictly proper scoring rules.

1.4.2 A construction principle

This subsection illustrates that the Dawid-Sebastiani (DS) score (see previous subsection) is a special case of a general result which constructs a proper scoring rule from an exponential family of densities under certain assumptions. To illustrate the idea, consider the DS score and notice that

$$\mathrm{DSS}(F, y) = -2\log(\varphi(y \mid \mu_F, \Sigma_F)) - d\log(2\pi) = 2\mathrm{LogS}(\varphi(\cdot \mid \mu_F, \Sigma_F), y) - d\log(2\pi),$$

where $\varphi(y \mid \mu, \Sigma)$ denotes the density of the multivariate normal distribution with mean μ and covariance matrix Σ. This connection raises the question, under which conditions it is possible to obtain a proper scoring rule from the logarithmic score combined with some parametric family of densities, e.g. the normal distributions.

To answer this question in a general setting let $\mathcal{E} := \{F_\theta \mid \theta \in \Theta\} \subseteq \mathcal{F}$ be a parametric family of distributions with parameter space Θ. Let ϕ given by $\phi : \mathcal{F} \to \mathcal{E}$, $F \mapsto F_\theta$ be a mapping onto $\mathcal{E}$ and write $\theta(F)$ for the parameter θ in $\phi(F) = F_\theta$. The following result specifies when a new proper scoring rule can be constructed from a given one.

Theorem 1.25. *Let $S : \mathcal{F} \times \mathsf{O} \to \bar{\mathbb{R}}$ be a proper scoring rule and $\phi : \mathcal{F} \to \mathcal{E}$ a mapping. If there is an $\mathcal{F}$-integrable function $H : \mathsf{O} \to \mathbb{R}$ such that for all $F, G \in \mathcal{F}$*

$$\bar{S}(\phi(F), G) + \bar{H}(G) = \bar{S}(\phi(F), \phi(G)) + \bar{H}(\phi(G)), \tag{1.7}$$

then the scoring rule

$$S^*(F, y) = S(\phi(F), y) = S(F_{\theta(F)}, y)$$

is proper.

Proof. For $F, G \in \mathcal{F}$ invoke Equation (1.7) two times to obtain

$$\begin{aligned}\bar{S}^*(F, G) = \bar{S}(\phi(F), G) &= \bar{S}(\phi(F), \phi(G)) + \bar{H}(\phi(G)) - \bar{H}(G)\\ &\geq \bar{S}(\phi(G), \phi(G)) + \bar{H}(\phi(G)) - \bar{H}(G) = \bar{S}(\phi(G), G) = \bar{S}^*(G, G),\end{aligned}$$

where the inequality stems from the propriety of S. □

Strict propriety is only possible for special choices of $\mathcal{E}$ and ϕ, rendering the mapping ϕ a bijection, but we omit these details here. Moreover, the existence of a mapping H is not necessary if we rewrite Theorem 1.25 in terms of equivalence of scoring rules, which is defined as in Definition 1.3.

We now turn to some examples of classes $\mathcal{E}$ and scoring rules S where Theorem 1.25 yields new proper scoring rules or recovers existing ones. We focus on the logarithmic and the Hyvärinen score (see Subsection 1.4.1) since both lead to explicit examples. We begin by defining an exponential family for a set of parameters Θ since it allows for several non-trivial parametric classes $\mathcal{E}$ which satisfy condition (1.7) in combination with the logarithmic or Hyvärinen score. A set of densities $\{f(\cdot \mid \theta) \mid \theta \in \Theta\}$ is an *exponential family* if any member can be represented via

$$f(y \mid \theta) = h(y) \exp\left(\eta(\theta)^\top t(y) - A(\theta)\right)$$

for $m \in \mathbb{N}$ and measurable functions $h : \mathsf{O} \to [0, \infty)$, $t : \mathsf{O} \to \mathbb{R}^m$, $\eta : \Theta \to \mathbb{R}^m$, and $A : \Theta \to \mathbb{R}$. The mapping A is often called *log-partition* function and t is a sufficient statistic for the parameter θ, see Barndorff-Nielsen (2014) for details.

Logarithmic score It is straightforward to see how the logarithmic score fits well with exponential families of distributions. In detail, let $\mathcal{E}$ be an exponential familiy and set $H(y) := \log h(y)$. Then Equation (1.7) for the logarithmic score reduces to

$$\bar{t}(G) = \bar{t}(\phi(G)), \tag{1.8}$$

such that only the expectation of t is involved. Such an expectation can often be calculated and expressed in terms of $\theta \in \Theta$ via the partial derivatives of the log-partition function A, giving simple conditions on the mapping $\phi : \mathcal{F} \to \mathcal{E}$.

Hyvärinen score Since the Hyvärinen score is based on the logarithm of the density, as well, it is natural to use it in Theorem 1.25. For simplicity assume that $\mathcal{E}$ is an exponential family of distributions on $\mathsf{O} = \mathbb{R}^d$ where the function h is constant and all densities satisfy the regularity conditions of the class $\mathcal{L}^*$ introduced above. If we define $W_\theta(y) := \eta(\theta)^\top t(y)$, then the Hyvärinen score on $\mathcal{E}$ depends on

$$\Delta W_\theta(y) = \sum_{i=1}^m \eta_i(\theta)\Delta t_i(y) \quad \text{and} \quad \nabla W_\theta(y) = \sum_{i=1}^m \eta_i(\theta)\nabla t_i(y)$$

only. As a consequence, Equation (1.7) holds if the derivatives of the sufficient statistic t satisfy

$$\mathbb{E}_G \Delta t_i(Y) = \mathbb{E}_{\phi(G)} \Delta t_i(Y) \quad \text{and} \quad \mathbb{E}_G \nabla t_i(Y)^\top \nabla t_j(Y) = \mathbb{E}_{\phi(G)} \nabla t_i(Y)^\top \nabla t_j(Y) \tag{1.9}$$

for $i, j = 1, \dots, m$, giving $m + m(m+1)/2$ identities. These equations are not necessary for Equation (1.7) to hold, but they provide simple conditions which can be checked to define a suitable mapping $\phi : \mathcal{F} \to \mathcal{E}$, see Example 1.29.

We close the section with several examples which construct proper scoring rules via Theorem 1.25 and the logarithmic or Hyvärinen score.

Example 1.26 (Normal distribution)**.** Let $\mathcal{E}$ consist of multivariate normal distributions with parameter $\theta = (\mu, \Sigma)$. Its exponential family representation implies $t(y) = (y, yy^\top)$. For the logarithmic score, Equation (1.8) yields that the mapping ϕ is determined by

$$(\mathbb{E}_G Y, \mathrm{Cov}_G(Y)) = \mathbb{E}_G t(Y) = \mathbb{E}_{\phi(G)} t(Y) = (\mu_G, \Sigma_G),$$

such that $\theta(F) = (\mathbb{E}_F Y, \mathrm{Cov}_F(Y))$ has to be computed from the predictive distribution. The resulting scoring function S^* from Theorem 1.25 is proper and equivalent to the DS score, as illustrated at the beginning of the subsection. $\diamond$

Example 1.27 (Laplace distribution)**.** Let $\mathcal{E}$ be the class of centered Laplace distributions with parameter $\nu > 0$, i.e. with densities $f(y \mid \nu) = (2\nu)^{-1} \exp(-|y|/\nu)$. This is an exponential family with $t(y) = |y|$ and for the logarithmic score Equation (1.8) becomes

$$\mathbb{E}_G|Y| = \mathbb{E}_G t(Y) = \mathbb{E}_{\phi(G)} t(Y) = \mathbb{E}_{\phi(G)}|Y| = \nu_G,$$

such that $\theta(F) = \mathbb{E}_F|Y|$ is computed from the predictive distribution. Theorem 1.25 implies that the scoring rule

$$S^*(F, y) = \log(2\nu_F) + \frac{|y|}{\nu_F},$$

where $\nu_F = \mathbb{E}_F|Y|$, is proper. One might wonder, whether it is possible to transfer the same arguments to the general class of Laplace distributions with parameters (μ, ν), i.e. to the situation of non-constant location parameter μ. In this case, Equation (1.7) reads

$$\mathbb{E}_G \left(\frac{|Y - \theta_1(F)|}{\theta_2(F)} + H(Y) \right) = \mathbb{E}_{\phi(G)} \left(\frac{|Y - \theta_1(F)|}{\theta_2(F)} + H(Y) \right),$$

which shows that the variable Y and the parameter $\theta_1(F)$ cannot be separated. It is thus unclear how to obtain a mapping ϕ which satisfies this identity for all $F, G \in \mathcal{F}$ if $\mathcal{F}$ is sufficiently large. As a consequence, Theorem 1.25 does not yield a new proper scoring rule. $\diamond$

Example 1.28 (Poisson distribution)**.** Let $\mathsf{O} = \mathbb{N}$ and $\mathcal{E}$ be the class of Poisson distributions with parameter $\lambda > 0$. We have $t(y) = y$ such that for the logarithmic score Equation (1.8) becomes

$$\mathbb{E}_G Y = \mathbb{E}_G t(Y) = \mathbb{E}_{\phi(G)} t(Y) = \mathbb{E}_{\phi(G)} Y = \lambda_G,$$

hence $\theta(F) = \mathbb{E}_F Y$ is reported. Theorem 1.25 implies that the scoring rule

$$S^*(F, y) = -y \log(\lambda_F) + \lambda_F + \log(y!),$$

where λ_F is the expectation of F, is proper. For an alternative derivation of an equivalent scoring rule, see Proposition 4.18. ◇

Example 1.29 (Normal distribution, continued)**.** Let $\mathcal{E}$ be as in Example 1.26. For the Hyvärinen score, the conditions in (1.9) reduce to equations which contain the moments $\mathbb{E}_G Y_i$ and mixed moments $\mathbb{E}_G Y_i Y_j$ for $i, j = 1, \ldots, m$, only. Hence, we obtain that the mapping ϕ of Example 1.26 with parameters $\theta(F) = (\mathbb{E}_F Y, \mathrm{Cov}_F(Y))$ satisfies these conditions. As a result we obtain a new Dawid-Sebastiani score given by

$$S^*(F, y) = -2 \operatorname{tr} \Sigma_F^{-1} + \|\Sigma_F^{-1}(y - \mu_F)\|^2 = -2 \operatorname{tr} \Sigma_F^{-1} + (y - \mu_F)^\top \Sigma_F^{-2}(y - \mu_F),$$

which is proper due to Theorem 1.25. ◇

1.5 Forecast comparison and score differences

When several reports of a functional T or a distribution are available, consistent scoring functions and proper scoring rules become useful statistical tools to address the question of relative forecast performance. This is because the concept of consistency, or equivalently propriety, suggests a natural way of comparing the quality of competing reports: Call a forecast superior to its competitor if it achieves a lower expected score. This principle allows for choosing between two forecasts based on their difference in expected scores, with only minimal assumptions on the data-generating process.

When this idea is put into practice we have to rely on estimates, as the true expected score differences are not available. This motivates an intuitive forecast comparison setting, see e.g. Nolde and Ziegel (2017) and Gneiting and Ranjan (2011). We give a brief introduction for consistent scoring functions, but the evaluation is analogous for proper scoring rules. Let $T : \mathcal{F} \to \mathsf{A}$ be an elicitable functional with strictly consistent scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ and consider a sequence of random variables $(Y_t)_{t \in \mathbb{N}}$ adapted to a filtration $(\mathcal{H}_t)_{t \in \mathbb{N}}$. Assume that forecasts of the functional T applied to the conditional distribution $Y_t \mid \mathcal{H}_{t-1}$ are given. These forecasts can be regarded as random sequences $R = (R_t)_{t \in \mathbb{N}}$ and $R^* = (R_t^*)_{t \in \mathbb{N}}$ and their performance can be compared via the *average score difference*

$$\Delta_n(R, R^*) := \frac{1}{n} \sum_{t=1}^n S(R_t, Y_t) - \frac{1}{n} \sum_{t=1}^n S(R_t^*, Y_t) = \frac{1}{n} \sum_{t=1}^n \left(S(R_t, Y_t) - S(R_t^*, Y_t) \right), \quad (1.10)$$

which is the natural estimator for the difference in expected scores. Based on the law of large numbers and the strict consistency of S, a positive value supports the conjecture that R^* is superior to R, while a negative value supports the opposite conjecture.

A key issue which remains is to assess whether the size of the observed average score difference provides enough evidence to favor one of the two forecast sequences. The Diebold-Mariano (DM) test (Diebold and Mariano, 1995) addresses this aspect by testing whether $\Delta_n(R, R^*)$ is significantly different from zero. In the simple situation of an i.i.d. sequence $(Y_t)_{t\in\mathbb{N}}$, the forecast sequences reduce to constants $r, r^* \in \mathsf{A}$ and we can perform such a test based on the asymptotic normality of the well-known t-statistic $t_n := \sqrt{n}\Delta_n(r, r^*)/\sqrt{\hat{\sigma}_n^2}$, where $\hat{\sigma}_n^2$ estimates the variance of $S(r, Y_t) - S(r^*, Y_t)$. For more general time series $(Y_t)_{t\in\mathbb{N}}$, $(R_t)_{t\in\mathbb{N}}$, and $(R_t^*)_{t\in\mathbb{N}}$ we refer to the evaluation setting worked out in Nolde and Ziegel (2017), where tests for equal forecast performance rely on suitable asymptotic results developed in Giacomini and White (2006).

Alternative testing schemes have been developed, e.g. tests for forecast encompassing (Giacomini and Komunjer, 2005; Clements and Harvey, 2010; Dimitriadis and Schnaitmann, 2020), which also allow to focus on conditional instead of unconditional performance. Moreover, if the sample size n is too low to justify asymptotic normality assumptions, the Wilcoxon signed-rank test provides an alternative to asymptotic normality, but relies on the assumption of symmetric score differences. Overall, a variety of testing approaches are available to leverage consistent scoring functions for comparative forecast evaluation.

2 | Elicitability of tail properties

This chapter addresses the question whether consistent scoring functions are available to perform comparative forecast evaluation for statistical properties of distribution tails. To this end, we introduce the concept of max-functionals which naturally arises from a key feature shared by the statistical functionals that are typically considered in extreme value theory. We demonstrate that max-functionals cannot be elicitable and even have infinite elicitation complexity. Consequently, we then turn from point to probabilistic forecasting and allow for reports of the entire distribution function. In this setting we show that it is an inherent property of all proper scoring rules that they cannot perfectly distinguish among different max-functional values.

The chapter is organized as follows. Section 2.1 recalls the notions of conditional elicitability and elicitation complexity and Section 2.2 introduces the class of max-functionals. We then show that they cannot be elicitable and that their elicitation complexity is infinite under mild assumptions. Section 2.3 provides examples of widely used max-functionals and Section 2.4 connects to proper scoring rules. We show that arbitrary large differences in tail behavior, either quantified by tail equivalence or max-functionals, can remain undetected by proper scoring rules. Section 2.5 concludes with a discussion of the results.

The material is based on Brehmer and Strokorb (2019), where I stated and proved the technical results, except for Proposition 2.11, and wrote the first draft. Kirstin Strokorb introduced the initial question of the paper. Both of us polished and edited text.

2.1 Conditional elicitability and elicitation complexity

The non-elicitability of certain functionals can often be addressed by requiring additional information from the forecasters before evaluation. The simplest instance of this idea is that a functional T can be *jointly elicitable*, meaning that there exists a functional $T' : \mathcal{F} \to \mathsf{A}'$ such that $(T, T')^\top$ becomes elicitable. Alternative, but closely connected, concepts are conditional elicitability (Emmer et al., 2015; Fissler and Ziegel, 2016) and elicitation complexity (Lambert et al., 2008; Frongillo and Kash, 2020). We start by giving an illustrating example.

Example 2.1 (Variance, continued)**.** Example 1.8 illustrates that the variance functional T_{var} is not elicitable, whereas the vector-valued functional $(T_1, T_{\text{var}})^T$ is. Hence, T_{var} is jointly elicitable with the first moment T_1. Another notable property is that on every subset of $\mathcal{F}$ where T_1 is constant, T_{var} reduces to a shifted version of the second moment T_2 and is thus elicitable on this subset. That is, T_{var} is conditionally elicitable given T_1 in the following sense. ◇

Definition 2.2 (Conditional elicitability)**.** Let $T : \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^n$ and $T' : \mathcal{F} \to \mathsf{A}' \subseteq \mathbb{R}^k$ be functionals and let T' be elicitable. For any $x \in \mathsf{A}'$ define the set

$$\mathcal{F}_x := \{F \in \mathcal{F} \mid T'(F) = x\}.$$

Then the functional T is called *conditionally elicitable given* T' if for any $x \in \mathsf{A}'$ its restriction to the class $\mathcal{F}_x$ is elicitable.

The concept of conditional elicitability was first introduced by Emmer et al. (2015) and motivated by a conditional backtesting approach for Expected Shortfall (ES) forecasts. A slight generalization was given by Fissler and Ziegel (2016). Our definition coincides with the one from Fissler and Ziegel (2016) except that we drop the condition that T' has elicitable components and only require it to be elicitable. This allows for a more convenient presentation of our results below.

Neither joint elicitability nor conditional elicitability imply elicitability, which follows from Example 2.1 with the variance functional serving as a counterexample. If a functional T is jointly elicitable with the functional T', and T' is elicitable, then it is conditionally elicitable given T'. Conversely, as discussed in Fissler and Ziegel (2016), it is unclear under which conditions a conditionally elicitable functional is jointly elicitable.

The definitions of joint elicitability and conditional elicitability both require a second elicitable functional T' accompanying the functional of interest. The distinction between both functionals is made more explicit in the concept of *elicitation complexity*. To illustrate this, recall Example 2.1 and note that the variance functional satisfies $T_{\text{var}} = f(T_1, T_2)$, where $f(x_1, x_2) = x_2 - x_1^2$. Since T_1 and T_2 are elicitable, we say that the variance functional has complexity 2. In general, T has elicitation complexity at most k if there is an elicitable functional $T' : \mathcal{F} \to \mathsf{A}' \subseteq \mathbb{R}^k$ such that $T = f(T')$ holds. Any f and T' satisfying this condition are then called *link function* and *intermediate functional*, respectively. The smallest dimension k for which such a representation is feasible is the elicitation complexity.

Definition 2.3 (Elicitation complexity)**.** For any set of distribution functions $\mathcal{F}$ the set of $\mathbb{R}^k$-valued elicitable functionals defined on $\mathcal{F}$ is denoted via $\mathcal{E}_k(\mathcal{F})$. For a functional $T : \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ and sets $\mathcal{C}_k \subseteq \mathcal{E}_k(\mathcal{F})$ the *elicitation complexity of* T *with respect to* $(\mathcal{C}_k)_{k\in\mathbb{N}}$ is defined via

$$\mathsf{elic}(T) := \min\{k \in \mathbb{N} \mid \exists\, T' \in \mathcal{C}_k : T = f \circ T' \text{ for some } f : T'(\mathcal{F}) \to \mathsf{A}\}.$$

If the minimum is not attained for any $k \in \mathbb{N}$, the elicitation complexity of T with respect to $(\mathcal{C}_k)_{k\in\mathbb{N}}$ is infinite and we write $\mathsf{elic}(T) = \infty$.

Elicitation complexity was introduced by Lambert et al. (2008) and further analyzed in Frongillo and Kash (2020), the latter motivated by its role in empirical risk minimization (ERM) algorithms in machine learning. Intuitively speaking, it replaces the question *whether* a functional is elicitable by the question *how complex* it is to elicit the functional.

If no regularity conditions are imposed on f or T', this can lead to small complexities without clear benefits in applications. More precisely, if f is arbitrary and $\mathcal{C}_k = \mathcal{E}_k(\mathcal{F})$ is chosen, pathological choices of f, like bijections from $\mathbb{R}^k$ to $\mathbb{R}$, cause all functionals to have complexity 1, as demonstrated by Frongillo and Kash (2020, Remark 4). To avoid such problems, it is standard to choose suitable subclasses $\mathcal{C}_k$ of intermediate functionals. One possible choice, which is used by Frongillo and Kash (2020) as well as Dearborn and Frongillo (2020), is $\mathcal{C}_k := \mathcal{I}_k(\mathcal{F}) \cap \mathcal{E}_k(\mathcal{F})$, where $\mathcal{I}_k(\mathcal{F})$ is the set of $\mathbb{R}^k$-valued identifiable functionals on $\mathcal{F}$. Another possibility, implicitly used by Lambert et al. (2008), is to define $\mathcal{C}_k$ to be a subclass of all functionals which have elicitable components.

Lastly, it is also possible to impose regularity conditions on the link function f, e.g. by requiring differentiability or continuity. Notably, joint elicitability can be understood as a version of elicitation complexity where the link function is the projection on the last component (Frongillo and Kash, 2020).

We need to be cautious when interpreting elicitation complexity, since imposing different regularity conditions via $(\mathcal{C}_k)_{k\in\mathbb{N}}$ can lead to different elicitation complexities for the same functional (Frongillo and Kash, 2020). In particular, an $\mathbb{R}^k$-valued functional might be elicitable and simultaneously have elicitation complexity strictly greater than k. Conversely, a functional can have elicitation complexity 1, although it is not itself elicitable, see Frongillo and Kash (2020, Remark 2).

We conclude this section with a lemma which considers the properties of a functional T if it is restricted to some subclass $\mathcal{F}^* \subseteq \mathcal{F}$. The first statement corresponds to the first part of Lemma 2.11 of Fissler and Ziegel (2015), the second and third statement are simple extensions. Their proofs are straightforward and therefore omitted.

Lemma 2.4. *Let $T : \mathcal{F} \to \mathsf{A}$ be a functional and let $\mathcal{F}^* \subseteq \mathcal{F}$ be non-empty.*

(i) If T is elicitable, then the restricted functional $T_{|\mathcal{F}^*}$ is elicitable.

(ii) If $\mathsf{elic}(T) = k$ with respect to $(\mathcal{C}_k)_{k\in\mathbb{N}}$ and we define $\mathcal{C}_k^* := \{T'_{|\mathcal{F}^*} \mid T' \in \mathcal{C}_k\}$, then $\mathsf{elic}(T_{|\mathcal{F}^*}) \leq k$ with respect to $(\mathcal{C}_k^*)_{k\in\mathbb{N}}$.

(iii) If $\mathsf{elic}(T) = k$ with respect to $(\mathcal{C}_k)_{k\in\mathbb{N}}$ and sets $(\mathcal{C}'_k)_{k\in\mathbb{N}}$ satisfy $\mathcal{C}_k \subseteq \mathcal{C}'_k$ for all $k \in \mathbb{N}$, then $\mathsf{elic}(T) \leq k$ with respect to $(\mathcal{C}'_k)_{k\in\mathbb{N}}$.

2.2 The elicitation complexity of max-functionals

This section introduces max-functionals, the central objects of our study, and investigates their elicitability as well as their elicitation complexity. Henceforth, let $\mathcal{F}$ always denote a *convex* class of distributions.

Definition 2.5 (Max-functional)**.** A functional $T : \mathcal{F} \to \mathbb{R}$ is a *max-functional* if

$$T(\lambda F_1 + (1-\lambda)F_0) = \max\left(T(F_0), T(F_1)\right)$$

holds for all $F_0, F_1 \in \mathcal{F}$ and $\lambda \in (0,1)$.

The essential feature of a max-functional is that its value on convex combinations of distributions is determined by the values attained on the extreme points. Equivalently, we can also define min-functionals and all results carry over with minor modifications. The constant functional is the simplest max-functional, but we will usually not be interested in this trivial case. Instead, Section 2.3 collects some non-trivial examples of max-functionals that are routinely considered in extreme value theory. Also note that, by definition, restrictions of max-functionals to a certain set of values are again max-functionals.

Lemma 2.6. *Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional and $A \subset \mathbb{R}$ a set. Set $\mathcal{F}_A := \{F \in \mathcal{F} \mid T(F) \in A\}$, then $\mathcal{F}_A$ is convex and the restricted functional $T : \mathcal{F}_A \to A \subset \mathbb{R}$ is also a max-functional.*

We start by proving that max-functionals cannot be elicitable. As remarked in Section 1.2, the usual way to show that a functional is not elicitable consists of applying Proposition 1.7, i.e. showing that it fails to have convex level sets. However, any max-functional has convex level sets by definition, so this approach is not feasible. Instead, we employ Theorem 1.11 to obtain the following result.

Theorem 2.7. *If $T : \mathcal{F} \to \mathbb{R}$ is a non-constant max-functional, then it is not elicitable.*

Turning from the elicitability to the elicitation complexity of max-functionals, the question of elicitation complexity is only meaningful in relation to a family of sets $(\mathcal{C}_k)_{k\in\mathbb{N}}$, where each set $\mathcal{C}_k \subset \mathcal{E}_k(\mathcal{F})$ is a collection of reasonably regular $\mathbb{R}^k$-valued elicitable functionals, cf. Section 2.1. Our major regularity requirement is mixture-continuity (see Definition 1.17) as in Bellini and Bignozzi (2015) and Fissler and Ziegel (2019b). As discussed in Subsection 1.3.3, an elicitable functional which is not mixture-continuous can have discontinuous expected scores. Moreover, it cannot have a self-calibrated consistent scoring function by Proposition 1.20. Both of these issues can lead to difficulties in forecast evaluation, estimation and regression.

To avoid further degenerate behavior, we impose a richness assumption on potential intermediate functionals T' in the sense that we require the image $T'(\mathcal{F}) \subseteq \mathbb{R}^k$ to have at least non-empty interior. This assumption is natural for large enough classes $\mathcal{F}$ and was, for instance, used by Fissler and Ziegel (2016, 2019b) when establishing results on consistent scoring functions for T'.

In addition to mixture continuity, we follow Lambert et al. (2008) and consider only functionals with elicitable components. Summarising, the first family of functionals which we consider in our complexity result is

$$\mathcal{U}_k := \left\{ T' \in \mathcal{E}_k(\mathcal{F}) \;\middle|\; \begin{array}{l} T' \text{ mixture-continuous with elicitable} \\ \text{components, } \operatorname{int}(T'(\mathcal{F})) \neq \emptyset \end{array} \right\},$$

where $\operatorname{int}(B)$ denotes the interior of a set $B \subseteq \mathbb{R}^k$. Alternatively, we require that the image $T'(\mathcal{F})$ of a potential intermediate functional T' has not only non-empty interior, but is itself an open set, i.e. we consider the family

$$\mathcal{V}_k := \left\{ T' \in \mathcal{E}_k(\mathcal{F}) \;\middle|\; \begin{array}{l} T' \text{ mixture-continuous with elicitable} \\ \text{components, } T'(\mathcal{F}) \text{ open} \end{array} \right\}.$$

We are now in position to consider the elicitation complexity of max-functionals with respect to these families.

Theorem 2.8. *Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional. Then the following hold true.*

(i) T has elicitation complexity ∞ with respect to $(\mathcal{U}_k)_{k\in\mathbb{N}}$ unless $T(\mathcal{F})$ contains its supremum.

(ii) T has elicitation complexity ∞ with respect to $(\mathcal{V}_k)_{k\in\mathbb{N}}$ unless T is constant.

Proof. Assume there is a $k \in \mathbb{N}$, a surjective functional $T' : \mathcal{F} \to \mathsf{A}'$ in $\mathcal{U}_k$ or $\mathcal{V}_k$ and a function $f : \mathsf{A}' \to \mathbb{R}$ such that $T = f \circ T'$. Without loss of generality, T' is surjective, hence its mixture-continuity together with the assumed convexity of $\mathcal{F}$ imply that A' is path-connected. Since it has non-empty interior, we can choose a hyperrectangle $Q := \prod_{i=1}^k [c_i, d_i] \subseteq \operatorname{int}(\mathsf{A}')$ and consider each component of T' isolated on Q. To do so, choose a component $j \in \{1, \dots, k\}$ and a $z_i \in [c_i, d_i]$ for all $i \in \{1, \dots, k\}\backslash\{j\}$. We can then obtain $F_{c_j,z}, F_{d_j,z} \in \mathcal{F}$ such that

$$\begin{aligned} T'(F_{c_j,z}) &= (z_1, \dots, z_{j-1}, c_j, z_{j+1}, \dots, z_k) \quad \text{and} \\ T'(F_{d_j,z}) &= (z_1, \dots, z_{j-1}, d_j, z_{j+1}, \dots, z_k). \end{aligned}$$

All components of T' are elicitable and thus have convex level sets by Proposition 1.7. Consequently, the i-th component, where $i \in \{1, \dots, k\}\backslash\{j\}$, equals z_i for all convex combinations of $F_{c_j,z}$ and $F_{d_j,z}$. If we define

$$\mathsf{A}'_{j,z} := \{(z_1, \dots, z_{j-1}, x, z_{j+1}, \dots, z_k) \mid x \in (c_j, d_j)\} \subseteq Q,$$

the fact that the j-th component has convex level sets and is mixture-continuous implies that for all $a \in \mathsf{A}'_{j,z}$ there exists a $\lambda \in (0,1)$ with $T'(\lambda F_{c_j,z} + (1-\lambda)F_{d_j,z}) = a$. The connection $T = f \circ T'$ now gives

$$\begin{aligned} f((z_1, \dots, z_{j-1}, x, z_{j+1}, \dots, z_k)) &= f(T'(\lambda F_{c_j,z} + (1-\lambda)F_{d_j,z}) \\ &= T(\lambda F_{c_j,z} + (1-\lambda)F_{d_j,z}) \\ &= \max(T(F_{c_j,z}), T(F_{d_j,z})) \end{aligned}$$

for all $x \in (c_j, d_j)$, implying that f has to be constant on the set $\mathsf{A}'_{j,z}$. Repeating this argument for any choice of $j \in \{1, \dots, k\}$ and $z_i \in [c_i, d_i]$ with $i \in \{1, \dots, k\}\backslash\{j\}$ shows that there is a $C \in \mathbb{R}$ such that $f(q) = C$ for all $q \in \operatorname{int}(Q)$.

Now fix $x_0 \in \operatorname{int}(Q)$. For any $x_1 \in \mathsf{A}'$ we can choose distributions $F_0, F_1 \in \mathcal{F}$ with $T'(F_0) = x_0$ and $T'(F_1) = x_1$. Since $x_0 \in \operatorname{int}(Q)$ and T' is mixture-continuous, there is a small $\mu \in (0,1)$ such that $T'(\mu F_1 + (1-\mu)F_0) \in \operatorname{int}(Q)$ holds. We thus obtain

$$\begin{aligned} C = f(T'(\mu F_1 + (1-\mu)F_0)) &= T(\mu F_1 + (1-\mu)F_0) \\ &= \max(T(F_0), T(F_1)) \\ &= \max(f(x_0), f(x_1)) = \max(C, f(x_1)), \end{aligned}$$

implying $f(x_1) \leq C$. Since x_1 was arbitrary, we have $f(x) \leq C$ for all $x \in \mathsf{A}'$, showing $C = \sup T(\mathcal{F})$ and proving statement (i).

Assume now that A' is open. Then for every $x_1 \in \mathsf{A}'$ there is a hyperrectangle $Q_1 \subseteq \mathsf{A}'$ such that $x_1 \in \operatorname{int}(Q_1)$. Arguing as in the beginning of the proof gives $f(q) = f(x_1)$ for all $q \in \operatorname{int}(Q_1)$. Letting $T'(F_1) = x_1$ as above, we obtain a $\nu \in (0,1)$ such that $T'(\nu F_1 + (1-\nu)F_0) \in \operatorname{int}(Q_1)$. This implies

$$\begin{aligned} C = f(T'(\mu F_1 + (1-\mu)F_0)) &= \max(T(F_0), T(F_1)) \\ &= f(T'(\nu F_1 + (1-\nu)F_0)) = f(x_1). \end{aligned}$$

Since x_1 was arbitrary, T must be constant, proving part (ii). □

Theorem 2.8 implies infinite elicitation complexity of max-functionals in a wide range of natural settings. Ultimately, our main interest lies in understanding the elicitation complexity with respect to the more general family $\mathcal{U}_k$, which imposes only very weak assumptions on potential intermediate functionals.

Corollary 2.9. *Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional and let one of the following conditions be satisfied.*

(i) T is unbounded.

(ii) T is surjective onto an open interval (a, b).

(iii) T is surjective onto a half-open interval $[a, b)$.

Then T has elicitation complexity ∞ with respect to $(\mathcal{U}_k)_{k\in\mathbb{N}}$.

Alternatively, considering elicitation complexity with respect to the family $(\mathcal{V}_k)_{k\in\mathbb{N}}$ amounts to requiring more regularity for a potential intermediate functional T' and, in this case, *all* non-constant max-functionals have infinite elicitation complexity. Lemma 2.4 further implies that the infinite elicitation complexity of max-functionals also extends to larger classes than the considered convex family of distribution functions $\mathcal{F}$ and is valid with respect to smaller families contained in $(\mathcal{U}_k)_{k\in\mathbb{N}}$ or $(\mathcal{V}_k)_{k\in\mathbb{N}}$.

Finally, by definition, any functional of finite elicitation complexity is conditionally elicitable, but it is unclear whether the reverse implication holds. We thus conclude with showing that max-functionals with infinite elicitation complexity can neither be conditionally elicitable nor jointly elicitable.

Theorem 2.10. *Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional such that $\mathsf{elic}(T) = \infty$ with respect to a family $(\mathcal{C}_k)_{k\in\mathbb{N}}$. Let $T' : \mathcal{F} \to \mathsf{A}'$ be a functional with $T' \in \mathcal{C}_m$ for some $m \in \mathbb{N}$. Then the following hold true.*

(i) T is not conditionally elicitable given T'.

(ii) T is not jointly elicitable with T'.

Proof. For the first part assume conversely, that there is an $m \in \mathbb{N}$ and a functional $T' \in \mathcal{C}_m$ such that T is conditionally elicitable given T'. That is, T is elicitable on the subclass $\mathcal{F}_x = \{F \in \mathcal{F} \mid T'(F) = x\}$ for any $x \in \mathsf{A}'$. By assumption, there is no link function f such that $T = f \circ T'$ holds. Consequently, there is at least one $z \in \mathsf{A}' \subseteq \mathbb{R}^m$ such that T is not constant on $\mathcal{F}_z$. If z defines such a class, then it is convex due to the elicitability of T' and moreover we can find $F_0, F_1 \in \mathcal{F}_z$ such that $T(F_0) \neq T(F_1)$ holds. Theorem 1.11 now implies that the restriction of T to $\mathcal{F}_z$ cannot be elicitable, a contradiction to the conditional elicitability of T.

For the second part note that, as remarked in Section 2.1 and in the discussion of Fissler and Ziegel (2016), the joint elicitability of T with an elicitable functional T' implies that T is conditionally elicitable given T'. Consequently, the first part of the proof implies the result. □

We conclude this section with a technical remark. In the spirit of Frongillo and Kash (2020), our complexity result (Theorem 2.8) employs regularity assumptions on the possible intermediate functionals. The main assumption is that they possess elicitable components. Why this is essential is illustrated by the use of the hyperrectangle Q in the proof. Intuitively, this assumption can be relaxed at the cost of more technical arguments. The main challenge hereby is to control the values of T' in a small hyperrectangle (or ball) around some $x_0 \in \operatorname{int}(\mathsf{A}')$. However, we did not pursue this approach further, since we believe that our setting covers many functionals of practical interest and at the same time illustrates the irregular behavior that will be inherent to any link function for a max-functional.

2.3 Examples of max-functionals

Prominent examples of max-functionals, to which the results of Section 2.2 apply, are routinely considered in extreme value theory and are key characteristics for the purpose of inference on the tail of a distribution.

Upper endpoint For a real-valued random variable with distribution function F, its upper endpoint is the supremum of its support

$$x^F := \sup\{x \in \mathbb{R} \mid F(x) < 1\}.$$

By definition, the upper endpoint can be interpreted as a real-valued max-functional on the convex class $\{F \in \mathcal{F} \mid x^F < \infty\}$. Bellini and Bignozzi (2015, Example 3.9) discuss the upper endpoint under the name *worst-case risk measure* and show that it is not elicitable, once further regularity conditions on the admissible scoring functions are imposed. In light of Theorem 2.7, the non-elicitability of the upper endpoint follows without any further assumptions. In addition, it has infinite elicitation complexity in the sense of Theorem 2.8 and Corollary 2.9.

Index of regular variation / Tail index When the upper endpoint is infinite, another key characteristic to describe the tail behavior of heavy-tailed distributions is the index of regular variation. A strictly positive measurable function f satisfying

$$\lim_{x\to\infty} \frac{f(xt)}{f(x)} = t^\rho$$

for $t > 0$ is called *regularly varying (at infinity) with index* $\rho(f) \in \mathbb{R}$. For a distribution F its index of regular variation is the respective index for its survival function $\overline{F} := 1-F$, that is, $T(F) := \rho(\overline{F})$. Its inverse $T(F)^{-1}$ is also called *tail index* in the risk management literature, cf. McNeil et al. (2015, Section 5.1). If the tail $\overline{F}$ is regularly varying with (a negative) index ρ, this means that $\overline{F}$ decays essentially like a power function with decay rate $1/\rho$. Since $\rho(f+g) = \max(\rho(f), \rho(g))$ (cf. e.g. de Haan and Ferreira (2006, Proposition B.1.9)), the index of regular variation T is naturally a max-functional, while the tail index T^{-1} is a min-functional.

Tail-separating functionals More generally, we can deduce that the property of *'being a max-functional' (or min-functional)* is in fact inherent to all *'tail-ordering indices'*. To make this precise, let us consider the following natural order on distribution tails.

For two distribution functions F and G with upper endpoints $x^F, x^G \in \mathbb{R} \cup \{\infty\}$ we say that G *has heavier tail than* F and write $F <_t G$ if

$$\text{either} \quad x^F < x^G \qquad \text{or} \qquad x^F = x^G = x^* \text{ and } \lim_{x\to x^*} \frac{\overline{F}(x)}{\overline{G}(x)} = 0.$$

We say that F and G are *tail equivalent* and write $F \sim_t G$ if they share the same upper endpoint $x^F = x^G = x^* \in \mathbb{R} \cup \{\infty\}$ and

$$\lim_{x\to x^*} \frac{\overline{F}(x)}{\overline{G}(x)} \in (0, \infty).$$

Note that "$<_t$" defines a strict partial order on any set of distribution functions $\mathcal{F}$ and that for tail equivalent F and G neither $F <_t G$ nor $G <_t F$ can hold. The following proposition shows that a functional which respects the tail order "$<_t$" is a max-functional.

Proposition 2.11. *Let $T : \mathcal{F} \to \mathbb{R}$ be a functional that satisfies for all $F, G \in \mathcal{F}$*

$$T(F) - T(G) \begin{cases} \leq 0 & \textit{if } F <_t G, \\ \geq 0 & \textit{if } G <_t F, \\ = 0 & \textit{else.} \end{cases}$$

Then T is a max-functional.

Proof. Let $F_0, F_1 \in \mathcal{F}$ and set $F_\lambda := \lambda F_1 + (1-\lambda)F_0$ for $\lambda \in (0,1)$. We distinguish three cases. If $F_0 <_t F_1$, we have $x^{F_\lambda} = x^{F_1} \geq x^{F_0}$ and the identity

$$\frac{\overline{F}_\lambda(x)}{\overline{F}_1(x)} = \lambda + (1-\lambda)\frac{\overline{F}_0(x)}{\overline{F}_1(x)}$$

for $x < x^{F_1}$ implies $F_\lambda \sim_t F_1$. Hence, neither $F_\lambda <_t F_1$ nor $F_1 <_t F_\lambda$ can be true. Together with $T(F_0) \leq T(F_1)$ we may conclude $T(F_\lambda) = T(F_1) = \max(T(F_0), T(F_1))$. By symmetry, the case $F_1 <_t F_0$ can be treated analogously. In the remaining case we have neither $F_0 <_t F_1$ nor $F_1 <_t F_0$, so $x^{F_1} = x^{F_0} = x^{F_\lambda} =: x^*$ must hold. Consequently,

$$\liminf_{x\to x^*} \frac{\overline{F}_\lambda(x)}{\overline{F}_1(x)} \geq \lambda > 0 \qquad \text{and} \qquad \limsup_{x\to x^*} \frac{\overline{F}_\lambda(x)}{\overline{F}_1(x)} < \infty,$$

where the latter follows as the tail of F_0 is not heavier than the tail of F_1. This implies that neither $F_1 <_t F_\lambda$ nor $F_\lambda <_t F_1$ can hold true, which gives $T(F_\lambda) = T(F_1) = \max(T(F_0), T(F_1))$ and concludes the proof. □

Another instance of a tail-ordering functional in the sense of Proposition 2.11 is the $\mathcal{M}$*-index* as introduced in Cadena and Kratz (2016). If it exists, it is the unique $\rho \in \mathbb{R}$ such that

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{x^{\rho+\varepsilon}} = 0 \quad \text{and} \quad \lim_{x\to\infty} \frac{\overline{F}(x)}{x^{\rho-\varepsilon}} = \infty \qquad \text{for all } \varepsilon > 0.$$

It is easily seen that the $\mathcal{M}$-index coincides with the index of regular variation for distribution functions F with regularly varying tail function $\overline{F}$. As it sorts survival functions according to their power law decay, Proposition 2.11 implies that the $\mathcal{M}$-index is a max-functional.

Extreme value index A central characteristic of extreme value theory is the extreme value index, which classifies the limiting behavior of rescaled maxima of growing samples from a distribution. More precisely, if there exist suitable location-scale normings $a_n > 0$, $b_n \in \mathbb{R}$ such that the distribution functions $F_n(x) := F^n(a_n x + b_n)$ converge weakly to a non-degenerate distribution function G, the limiting distribution function G is necessarily a *Generalized Extreme Value Distribution (GEV)*. This means that up to a location-scale normalization we have

$$G(x) = G_\gamma(x) = \exp\{-(1+\gamma x)_+^{-1/\gamma}\}$$

for some $\gamma = \gamma(F) \in \mathbb{R}$, where $G_0(x) = \exp\{-e^{-x}\}$ for $\gamma = 0$. The distribution F is said to be in the *max-domain of attraction* of $G = G_\gamma$ and the shape parameter $\gamma(F)$ is the *extreme value index (EVI) of F*, cf. e.g. the monographs Resnick (1987) and de Haan and Ferreira (2006) for further background.

Let $\mathcal{F}$ be the class of distribution functions which are in a max-domain of attraction for some GEV and consider first the EVI on the subclass of heavy-tailed distributions $\mathcal{F}_+ = \{F \in \mathcal{F} \mid \gamma(F) > 0\}$. It is well-known that a distribution $F \in \mathcal{F}$ has EVI $\gamma > 0$ if and only if $\rho(\overline{F}) = -\gamma^{-1}$, where ρ is the index of regular variation (cf. e.g. Resnick (1987, Proposition 1.11)). Consequently, the EVI γ is also a max-functional on $\mathcal{F}_+$.

When considering the class of light-tailed distributions, i.e. the case $\gamma(F) < 0$, we need to specify an upper endpoint first in order to make 'being a max/min-functional' meaningful for the EVI γ. To this end, let $\mathcal{F}_{x^*} = \{F \in \mathcal{F} \mid \gamma(F) < 0, x^F = x^*\}$. Again the EVI behavior is governed by regular variation, since $\gamma(F) = -\gamma(F_*)$ with $F_*(x) = F(x^* - x^{-1})$ (cf. e.g. Resnick (1987, Proposition 1.13)). This shows that the EVI γ is a min-functional on the class $\mathcal{F}_{x^*}$. Note that it is crucial to assume equal upper endpoints, because otherwise it is not the EVI that dominates the tail behavior, but the upper endpoint itself.

So far, we have looked at statistical indices that classify *univariate* tail behavior. However, similar issues arise when we want to quantify *joint* tail behavior in higher dimensions. Exemplary, let us consider the coefficient of tail dependence.

Coefficient of tail dependence In order to quantify the tail behavior of a bivariate distribution function, Ledford and Tawn (1996, 1997) introduced the coefficient of tail dependence. For a bivariate distribution function F of a random vector (X_1, X_2) let us write $\overline{F}_i(x) := \mathbb{P}(X_i > x)$, $i = 1, 2$ and $\overline{F}(x) := \mathbb{P}(X_1 > x, X_2 > x)$ for the associated survival functions. Suppose there is an $\alpha > 0$ such that both $\overline{F}_1$ and $\overline{F}_2$ are regularly varying with index $-\alpha$. If in addition the joint survival function $\overline{F}$ is regularly varying with index $-\alpha/\eta$ for some $\eta \in (0, 1]$, the coefficient $\eta = \eta(F)$ is called *coefficient of tail dependence (CTD)* of the bivariate distribution F. Let us consider the CTD η on the class of bivariate distributions

$$\mathcal{F}_\alpha = \{F \mid \rho(\overline{F}_1) = \rho(\overline{F}_2) = -\alpha,\ \rho(\overline{F}) = -\alpha/\eta \text{ for some } \eta \in (0, 1]\}.$$

Then it follows for $F, G \in \mathcal{F}_\alpha$ that $\rho(\lambda\overline{F} + (1-\lambda)\overline{G}) = -\alpha/\max(\eta(F), \eta(G))$ by the properties of the index of regular variation. Hence η is a max-functional on $\mathcal{F}_\alpha$.

2.4 Proper scoring rules and max-functionals

In light of the results of Section 2.2, the following approach may seem reasonable to someone seeking information about a max-functional: Instead of single values, distribution functions are reported and evaluated via proper scoring rules. Then the max-functionals are computed from the forecasted distributions.

If the max-functional of interest is a property of the tail, e.g. the extreme value index, one could expect this method to work well as long as the scoring rule shows a good performance in the tails. In order to emphasize specific regions of interests, in particular the tails, Gneiting and Ranjan (2011) and Diks et al. (2011) combined scoring rules with weight functions. Drawbacks and benefits of these weighted proper scoring rules were further studied in Lerch et al. (2017) and Holzmann and Klar (2017), where the latter propose general construction principles. A theoretical problem is pointed out by Taillardat et al. (2019), who show that the weighted continuous ranked probability score (wCRPS, see Subsection 1.4.1) cannot detect that two distributions are not tail equivalent. More precisely, Taillardat et al. (2019, Section 2) demonstrate that given a distribution G and $\varepsilon > 0$, it is always possible to construct a distribution F that is not tail equivalent to G and such that

$$|\mathbb{E}_G \operatorname{wCRPS}(G, Y) - \mathbb{E}_G \operatorname{wCRPS}(F, Y)| \leq \varepsilon.$$

This results shows that for any distribution G the tail can be modified while keeping the expected wCRPS ε-close to its minimum. As put by Taillardat et al. (2019), this means that the wCRPS is not a *tail equivalent score.*

In the following we show that *all* proper scoring rules fail to be tail equivalent in this sense. Moreover, we extend these findings to max-functionals, i.e. we show that no proper scoring rule is *max-functional equivalent.* These findings are motivated by the observation that tail equivalence and max-functionals lead to a similar kind of discontinuity on the convex combinations $\lambda F + (1 - \lambda)G$, which intuitively conflicts with the diagonal-continuity (Definition 1.23) of integrable proper scoring rules. This allows for an extension of the results of Taillardat et al. (2019). Recall the tail-ordering from Section 2.3 and that we assume $\mathcal{F}$ to be convex.

Theorem 2.12. *Let $S : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$ be an integrable proper scoring rule and $G \in \mathcal{F}$. Then the following are true.*

(i) If there is an $F \in \mathcal{F}$ with heavier tail than G, then for all $\varepsilon > 0$ there is an $F_\varepsilon \in \mathcal{F}$ that is not tail equivalent to G and such that

$$|\bar{S}(F_\varepsilon, G) - \bar{S}(G, G)| \leq \varepsilon.$$

(ii) Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional. If there is an $F \in \mathcal{F}$ with $T(F) > T(G)$, then for all $\varepsilon > 0$ there is an $F_\varepsilon \in \mathcal{F}$ such that $T(F_\varepsilon) = T(F) > T(G)$, while

$$|\bar{S}(F_\varepsilon, G) - \bar{S}(G, G)| \leq \varepsilon.$$

Proof. Fix $G \in \mathcal{F}$ and let S be an integrable proper scoring rule. For $F \in \mathcal{F}$ set $F_\lambda := \lambda F + (1 - \lambda)G$. Since $\mathcal{F}$ is convex, we have $F_\lambda \in \mathcal{F}$ for all $\lambda \in [0, 1]$. Moreover,

S is diagonal-continuous at G by Lemma 1.24, implying that for all $\varepsilon > 0$ and $F \in \mathcal{F}$ we can find a $\delta \in (0,1]$ such that $|\bar{S}(F_\lambda, G) - \bar{S}(G,G)| \le \varepsilon$ holds for all $\lambda \in [0,\delta]$. Now assume there is an $F \in \mathcal{F}$ with heavier tail than G. If $x^F > x^G$, we have $x^{F_\lambda} > x^G$ for all $\lambda \in (0,1]$. If on the other hand $x^F = x^G = x^*$ we have

$$\frac{\overline{F}_\lambda(x)}{\overline{G}(x)} = (1-\lambda) + \lambda\frac{\overline{F}(x)}{\overline{G}(x)}$$

for $x < x^*$ and the right-hand side goes to infinity as $x \to x^*$. Hence, in both cases the distributions F_λ cannot be tail equivalent to G for $\lambda \in (0,1]$, showing part (i). For the second part, let $F \in \mathcal{F}$ satisfy $T(F) > T(G)$. Since T is a max-functional, $T(F_\lambda) = T(F) > T(G)$ holds for $\lambda \in (0,1]$, proving part (ii). □

The first part of Theorem 2.12 shows that the lack of tail equivalence is not a flaw of the wCRPS, but inherent to *all* integrable proper scoring rules. The second part extends this non-equivalence of proper scoring rules to max-functionals. Loosely speaking, this means that there cannot only be pairs of not tail equivalent distributions, but also pairs of distributions with arbitrarily different max-functional values, and both having almost identical expected scores.

2.5 Discussion

This chapter shows that max-functionals do not only fail to be elicitable, but have in fact infinite elicitation complexity in a wide range of settings. This contrasts situations in which the non-elicitability can be alleviated by a finite elicitation complexity as, for instance, is the case for the variance or the Expected Shortfall (see Example 2.1 and Frongillo and Kash (2020); Fissler and Ziegel (2016)). Rather it bears resemblance to the mode, which is non-elicitable and has infinite elicitation complexity as well, see Heinrich (2014) and Dearborn and Frongillo (2020). Concerning probabilistic forecasts we demonstrate that integrable proper scoring rules do not lead to a satisfying comparison of max-functional values, either. This complements recent findings of Taillardat et al. (2019) and extends these from the wCRPS to all integrable proper scoring rules.

Collectively, these results cast doubt on the ability of expected scores to distinguish different tail regimes in the sense of max-functional values as they are routinely considered in extreme value theory. From an applied viewpoint this means that expected scores are not suitable to access such tail information for regression, M-estimation or comparative forecast evaluation. Thereby, our results provide a new perspective on the limitations of weighted scoring rules, adding to practical intricacies described in Lerch et al. (2017), Holzmann and Klar (2017) and Friederichs and Thorarinsdottir (2012). What might come to rescue though, is that the max-functionals themselves are often not the main concern in applications, but rather a tool to guide the extrapolation from intermediate order statistics to the functionals of interest. In practice, these functionals may include a high quantile or a tail expectation such as Expected Shortfall, which can be interpreted as tail properties 'less extreme' than max-functionals and with better elicitablity properties.

However, we would like to point out that non-elicitability is not the only problem in sound forecast evaluation. Even when elicitability is granted, there is no guarantee that the corresponding minimization problem will be well-posed. For instance, poorly behaved scoring functions may give rise to high variances of realized average scores, in which case practical sample sizes may be per se too low for an adequate assessment of competing forecasts.

3 | Scoring Interval Forecasts

In this chapter we consider different kinds of predictive intervals, ask whether they are elicitable, and study the available classes of consistent scoring functions. We provide results for three types of interval forecasts studied in the extant literature. After some general considerations, Section 3.3 begins with the equal-tailed interval. It is determined by the $\alpha/2$- and $(1-\alpha/2)$-quantile and is thus elicitable, with a rich class of consistent scoring functions. However, we show that subject to either translation invariance, or positive homogeneity and differentiability, the Winkler interval score becomes a unique choice. In Section 3.4 we then turn to the shortest interval, which has minimal length subject to a coverage probability of at least $1-\alpha$. We resolve a challenge raised by Askanazi et al. (2018), by showing that the shortest interval fails to be elicitable in practically relevant settings. Thirdly, Section 3.5 treats the modal interval, which has an interesting connection to the shortest interval and possesses a unique strictly consistent scoring function, up to equivalence. Section 3.6 concludes with a discussion.

The chapter is based on the preprint Brehmer and Gneiting (2020). Tilmann Gneiting conceptualized this work and formulated initial conjectures for the technical results. I developed the technical contents of the paper, including but not limited to all proofs and examples and wrote the first draft. Both of us polished and edited text.

3.1 Technical framework

Throughout this chapter, intervals will be defined via their lower and upper endpoints a and b, which are both elements of the observation domain O. Typically, O will either be the real line $\mathbb{R}$, or the set $\mathbb{N}_0$ of the nonnegative integers. Technically, we consider intervals which are elements of the action domain

$$\mathsf{A} = \mathsf{A}_{\mathsf{O}} := \{[a,b] \mid a, b \in \mathsf{O}, a \leq b\}.$$

This choice implies that the predictive intervals we consider are closed, with endpoints belonging to the observation domain O. The endpoint requirement leads to a natural and desirable reduction of the set of possible intervals for discrete data, such as in the case of count data, where the endpoints are required to be nonnegative integers. Closed intervals are compatible with the interpretation of the median as a '0% central prediction interval'. Moreover, in discrete settings an interval forecast might genuinely collapse to a single point, so closed intervals allow for a unified treatment of discrete and continuous distributions. Lastly, this setting is consistent with the extant literature, see e.g. Winkler (1972), Lambert and Shoham (2009, Section 7.6), and Askanazi et al. (2018). A more general treatment of interval forecasts could, for instance, allow forecasters to choose which type of interval (closed, half-open, open) they want to report. We do not pursue this approach, as it leads to further complexity without recognizable benefits.

The predictive intervals we consider below are rarely unique, which can be illustrated for the equal-tailed interval. As discussed above, its endpoints are given by quantiles, and as seen in Example 1.4, quantiles are inherently set-valued functionals whenever the CDFs are not strictly increasing. Hence, unless both quantiles reduce to single points, there are multiple equal-tailed intervals. This highlights that interval forecasts call for the set-valued framework of elicitability, which assumes functionals $T : \mathcal{F} \to 2^{\mathsf{A}}$.

We denote the length of an interval I as $\operatorname{len}(I)$, and if $\mathsf{A}' \subset \mathsf{A}$ is a set of intervals that all have the same length, we refer to this common length as $\operatorname{len}(\mathsf{A}')$. The left- and right-hand limits of a function $h : \mathbb{R} \to \mathbb{R}$ at x are denoted by $h(x-) := \lim_{y \uparrow x} h(y)$ and $h(x+) := \lim_{y \downarrow x} h(y)$, respectively.

3.2 Intervals with coverage guarantees

A standard principle for interval forecasts is that a correct report I contains (or covers) the outcome with specified nominal probability of at least $1 - \alpha$, where $\alpha \in (0, 1)$. A *guaranteed coverage interval* (GCI) at level α under the predictive distribution F is any element $[a, b] \in \mathsf{A}$ satisfying $F(b) - F(a-) \geq 1 - \alpha$, and for all $\varepsilon > 0$

$$F(b - \varepsilon) - F(a-) \leq 1 - \alpha \quad \text{and} \quad F(b) - F((a + \varepsilon)-) \leq 1 - \alpha. \tag{3.1}$$

A GCI thus contains just as much probability mass as necessary, but is not as short as possible. For continuous distributions this definition reduces to the intuitive requirement $F(b) - F(a) = 1 - \alpha$. We write $\mathrm{GCI}_\alpha(F)$ for the set of guaranteed coverage intervals at level α of F. An early theoretical treatment is in Proposition 7.6 of Lambert and Shoham (2009), according to which the GCI_α functional fails to be elicitable relative to the class of all distributions on the finite domain $\mathsf{O} = \{1, \ldots, n\}$. Frongillo and Kash (2019, Section 4.2) apply tools of convex analysis to extend this result to more general classes of distributions.

It is straightforward to recover these findings by showing that the GCI_α functional lacks the CxLS* property, see Proposition 1.7. Specifically, let $\alpha \in (0, 1)$ and consider continuous distributions F_0 and F_1 that satisfy $F_0(b') - F_0(a') = F_1(b') - F_1(a') = 1 - \alpha$ for some $a' < b'$, whereas

$$F_0(b) - F_0(a) > 1 - \alpha \quad \text{and} \quad F_1(b) - F_1(a) < 1 - \alpha$$

for some $a < b$. Then there is a $\lambda \in (0, 1)$ such that $[a, b] \in \mathrm{GCI}_\alpha(F_\lambda)$, even though $[a, b] \notin \mathrm{GCI}_\alpha(F_0) \cap \mathrm{GCI}_\alpha(F_1) \neq \emptyset$. Part (ii) of Proposition 1.7 thus implies that the GCI_α functional fails to be elicitable relative to convex classes $\mathcal{F}$ that contain distributions of the type used here. A similar construction for discrete distributions is immediate.

Fissler et al. (2020) introduce a concept of guaranteed coverage without the length restriction (3.1), i.e. they consider the class of intervals $[a, b] \in \mathsf{A}$ which satisfy $F(b) - F(a-) \geq 1 - \alpha$. Like the GCI_α, the corresponding set-valued functional fails to be elicitable, see Fissler et al. (2020, Corollary 4.7).

In addition to lacking elicitability, the GCI_α functional has the unattractive feature that it fails to be unique for very many distributions, including, but not limited to, all continuous distributions. This motivates the imposition of additional constraints on the predictive intervals, as discussed now. See also Fissler et al. (2020) for further types of intervals.

3.3 Equal-tailed interval (ETI)

A straightforward way to pick an interval with nominal coverage at least $1 - \alpha$ under F consists of choosing quantiles at levels $\beta \in (0, \alpha)$ and $\beta + 1 - \alpha$ as the lower and upper endpoint of the interval, respectively.

The ubiquitous choice is $\beta = \frac{\alpha}{2}$, such that under a continuous F the outcomes fall above or below the interval with equal probability of $\frac{\alpha}{2}$. In general, an *equal-tailed interval* (ETI) at level α of F is any member of

$$\mathrm{ETI}_\alpha(F) := \{[a,b] \in \mathsf{A} \mid a \in T_{\alpha/2}(F),\, b \in T_{1-\alpha/2}(F)\}, \tag{3.2}$$

where $T_\beta(F) := \{x \in \mathsf{O} \mid F(x-) \le \beta \le F(x)\}$ denotes the β-quantile functional. Some parts of the literature, e.g. Fissler et al. (2020), label it 'central prediction interval'. In the simplified situation where F is strictly increasing, all quantiles are unique and thus $\mathrm{ETI}_\alpha(F)$ reduces to a single interval.

In view of the definition via quantiles, forecasting equal-tailed intervals amounts to forecasting quantiles. Moreover, the ETI_α functional is elicitable, and we can construct consistent scoring functions for the ETI_α functional from the consistent scoring functions (1.2) for quantiles, as noted by Gneiting and Raftery (2007) and Askanazi et al. (2018). Specifically, if w_1, w_2 are nonnegative weights and $g_1, g_2 : \mathsf{O} \to \mathbb{R}$ are non-decreasing $\mathcal{F}$-integrable functions, then every $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ of the form

$$\begin{aligned} S([a,b],y) = {} & w_1 \left(\mathbb{1}(y \le a) - \frac{\alpha}{2}\right)(g_1(a) - g_1(y)) \\ & + w_2 \left(\mathbb{1}(y \le b) - \left(1 - \frac{\alpha}{2}\right)\right)(g_2(b) - g_2(y)) \end{aligned} \tag{3.3}$$

is a consistent scoring function for the ETI_α functional. Furthermore, S is strictly consistent if $w_1, w_2 \in (0, \infty)$ and g_1, g_2 are strictly increasing. It is no substantial loss of generality to restrict attention to the class in (3.3), since essentially all strictly consistent scoring functions for ETI_α are equivalent to this form. This results from the fact that ETI_α can be interpreted as a vector of two quantiles. Under suitable regularity conditions, all strictly consistent scoring functions for vectors of quantiles are equivalent to a sum of scoring functions of the form (1.2), see Fissler and Ziegel (2016, Proposition 4.2(ii)) and Fissler and Ziegel (2019a).

The choice $w_1 = w_2 = 2/\alpha$ and $g_1(x) = g_2(x) = x$ in (3.3) gives the classical *interval score* (IS) of Winkler (1972), namely,

$$\mathrm{IS}_\alpha([a,b],y) := (b-a) + \frac{2}{\alpha}(a-y)\mathbb{1}(y < a) + \frac{2}{\alpha}(y-b)\mathbb{1}(y > b), \tag{3.4}$$

which is strictly consistent for ETI_α relative to classes of distributions with finite first moment. This is the most commonly used scoring function for the ETI_α functional, and scaled or unscaled versions thereof have been employed implicitly or explicitly in highly visible, recent forecast competitions (Hong et al., 2016; Makridakis et al., 2020; M Open Forecasting Center, 2020).

The Winkler interval score (3.4) combines various additional, desirable properties of scoring functions on $\mathsf{O} = \mathbb{R}$, such as *translation invariance*, in the sense that for every $z, y \in \mathbb{R}$ and $a < b$

$$S([a-z, b-z], y-z) = S([a,b], y),$$

and *positive homogeneity* of order 1, in that for every $c > 0$, $y \in \mathbb{R}$, and $a < b$

$$S([ca, cb], cy) = cS([a,b], y).$$

Additionally, the IS applies the same penalty terms to values falling above or below the reported interval, such that it is *symmetric*, in the sense that

$$S([a,b], y) = S([-b,-a], -y)$$

for $y \in \mathbb{R}$ and $a < b$.

Our next two results concern scoring functions on $\mathsf{O} = \mathbb{R}$ that are of the form (3.3) and share one or more of these often desirable additional properties. In particular, the next theorem demonstrates that either translation invariance or positive homogeneity and differentiability, combined with symmetry, suffice to characterize the Winkler interval score (3.4), up to equivalence. To facilitate the exposition, assumption (ii) identifies the action domain $\mathsf{A} = \{[a,b] \mid a \le b\}$ with the respective subset $\{(a,b)^\top \in \mathbb{R}^2 \mid a \le b\}$ of the Euclidean plane.

Theorem 3.1. *Let S be of the form* (3.3) *with non-constant, non-decreasing functions g_1 and g_2. If S is either*

(i) translation invariant, or

(ii) positively homogeneous and differentiable with respect to $(a,b)^\top \in \mathbb{R}^2$, except possibly along the diagonal,

then g_1 and g_2 are linear. In particular, if S is symmetric and either (i) or (ii) applies, then S is equivalent to IS_α.

The first part of Theorem 3.1, which states the linearity of g_1 and g_2, continues to hold for asymmetric intervals, which are defined by choosing the endpoints $a \in T_\beta(F)$ and $b \in T_{\beta+1-\alpha}(F)$ for $\beta \in (0, \alpha)$ in (3.2). However, the second statement does not apply, as non-constant consistent scoring functions for such intervals cannot be symmetric.

Proof. Let S be a scoring function of the form (3.3). Let $y, z \in \mathbb{R}$, $a < b$ and choose $b = y$. Then translation invariance of S gives

$$\begin{aligned} -w_1\frac{\alpha}{2}(g_1(a) - g_1(y)) &= S([a,y], y) \\ &= S([a-z, y-z], y-z) \\ &= -w_1\frac{\alpha}{2}(g_1(a-z) - g_1(y-z)), \end{aligned}$$

and rearranging yields $g_1(a) - g_1(y) = g_1(a-z) - g_1(y-z)$ for $a, y, z \in \mathbb{R}$. Choose $y = 0$ and define $\tilde{g}(x) := g_1(x) - g_1(0)$ to obtain $\tilde{g}(a-z) = \tilde{g}(a) + \tilde{g}(-z)$ for $a, z \in \mathbb{R}$. Thus $\tilde{g}$

obeys Cauchy's functional equation, and since $\tilde{g}$ is non-constant and non-decreasing, we get $g_1(x) = \gamma x + g_1(0)$ for some $\gamma > 0$. For g_2 we apply the same arguments, to complete the proof of part (i).

Let $y \in \mathbb{R}$, $a < b$ and choose $b = y$. If S is positively homogeneous then for all $c > 0$

$$\begin{aligned}-w_1 c\frac{\alpha}{2}(g_1(a) - g_1(y)) &= cS([a, y], y)\\ &= S([ca, cy], cy) = -w_1\frac{\alpha}{2}(g_1(ca) - g_1(cy)),\end{aligned}$$

and thus $c(g_1(a) - g_1(y)) = g_1(ca) - g_1(cy)$ for $a, y \in \mathbb{R}$ and $c > 0$. Choose $y = 0$ and define $\tilde{g}(x) := g_1(x) - g_1(0)$ to obtain $c\tilde{g}(a) = \tilde{g}(ca)$ for $c > 0$ and $a \in \mathbb{R}$, as in Section C of the Supplementary Material for Nolde and Ziegel (2017). Since $\tilde{g}$ is non-constant, non-decreasing, and differentiable, $g_1(x) = \gamma x + g_1(0)$ for some $\gamma > 0$. Using the same arguments for g_2 we complete the proof of part (ii).

Now suppose S is also symmetric and $g_2(x) = \rho x + g_2(0)$ for some $\rho > 0$. Then the same reasoning as in the proof of Theorem 3.2 (see below) shows $w_1\gamma = w_2\rho$, which proves the equivalence to IS_α. □

If only symmetry is required in (3.3), then the class of possible scoring functions for the equal-tailed interval is much larger than just the interval score. To characterize these functions, take $\mathcal{I}$ to be the class of all non-decreasing functions $g : \mathbb{R} \to \mathbb{R}$ with the property that $g(x) = \frac{1}{2}(g(x-) + g(x+))$ for $x \in \mathbb{R}$. In a trivial deviation from Ehm et al. (2016) we define the *elementary quantile scoring function* as

$$S^{\mathrm{Q}}_{\alpha,\theta}(x, y) = (\mathbb{1}(y \le x) - \alpha)\left(\mathbb{1}(\theta < x) + \frac{1}{2}\mathbb{1}(\theta = x) - \mathbb{1}(\theta < y) - \frac{1}{2}\mathbb{1}(\theta = y)\right),$$

which is a special case of (1.2) with $g(x) = \mathbb{1}(\theta < x) + \frac{1}{2}\mathbb{1}(\theta = x)$. Given any $\theta \ge 0$, we now define

$$S_{\alpha,\theta}([a, b], y) = S^{\mathrm{Q}}_{\alpha/2,\theta}(a, y) + S^{\mathrm{Q}}_{1-\alpha/2,-\theta}(b, y)$$

and refer to $S_{\alpha,\theta}$ as the *elementary symmetric interval scoring function.* The following result shows that every symmetric scoring function of the form (3.3) arises as a mixture of elementary symmetric interval scoring functions. The Winkler interval score (3.4) emerges in the special case where the mixing measure μ is proportional to Lebesgue measure.

Theorem 3.2. *Let S be of the form* (3.3) *with non-constant, non-decreasing functions $g_1, g_2 \in \mathcal{I}$. If S is symmetric, then it is of the form*

$$S([a, b], y) = \int_{[0,\infty)} S_{\alpha,\theta}([a, b], y)\, \mathrm{d}\mu(\theta),$$

where μ is a Borel measure on $[0, \infty)$, defined via $\mathrm{d}\mu(\theta) = \mathrm{d}h(\theta)$ with $h(\theta) = w_1(g_1(\theta) - g_1(-\theta))$ for $\theta \in [0, \infty)$.

Proof. Let S be a scoring function of the form (3.3) and let $a, b, y \in \mathbb{R}$ with $a < b$ and $b = y$. Then the symmetry of S gives

$$\begin{aligned}-w_1\frac{\alpha}{2}(g_1(a) - g_1(y)) &= S([a, y], y)\\ &= S([-y, -a], -y) = w_2\frac{\alpha}{2}(g_2(-a) - g_2(-y)),\end{aligned}$$

Table 3.1: Properties of the four different intervals in $\mathrm{ETI}_\alpha(G)$, where $\alpha = 0.2$. The expected penalty for interval $[a, b]$ is given by $\mathbb{E}_G\left[\mathrm{IS}_\alpha([a, b], Y)\mathbb{1}(Y \notin [a, b])\right]$, so that the expected score decomposes into length plus penalty.

Interval	Coverage	Expected IS_α	Length	Expected Penalty
$[1, 2]$	0.8	3	1	2
$[0, 2]$	0.9	3	2	1
$[1, 3]$	0.9	3	2	1
$[0, 3]$	1.0	3	3	0

and rearranging yields $w_1(g_1(a) - g_1(y)) = w_2(g_2(-y) - g_2(-a))$ for $a, y \in \mathbb{R}$. For $x, y, \theta \in \mathbb{R}$, define the function

$$f(x, y, \theta) := \mathbb{1}(\theta < x) + \frac{1}{2}\mathbb{1}(\theta = x) - \mathbb{1}(\theta < y) - \frac{1}{2}\mathbb{1}(\theta = y),$$

which satisfies $f(-y, -x, \theta) = f(x, y, -\theta)$ for $x, y, \theta \in \mathbb{R}$. Moreover, for any $g \in \mathcal{I}$ and $y < x$

$$\int f(x, y, \theta)\,\mathrm{d}\mu_g(\theta) = \frac{1}{2}(g(x+) - g(y-)) + \frac{1}{2}(g(x-) - g(y+)) = g(x) - g(y),$$

where μ_g is the Borel measure on $\mathbb{R}$ induced by g. If we define the measures $\mu_1 = w_1\mu_{g_1}$ and $\mu_2 = w_2\mu_{g_2}$, then the first part of the proof implies

$$\begin{aligned}\int f(x, y, \theta)\,\mathrm{d}\mu_2(\theta) &= w_2(g_2(x) - g_2(y)) \\ &= w_1(g_1(-y) - g_1(-x)) \\ &= \int f(-y, -x, \theta)\,\mathrm{d}\mu_1(\theta) = \int f(x, y, -\theta)\,\mathrm{d}\mu_1(\theta)\end{aligned}$$

for $y < x$, and the proof is completed by defining μ via $\mu((y, x]) = \mu_1((y, x]) + \mu_1([-x, -y))$. □

The usual treatment of the ETI considers distributions $F \in \mathcal{F}$ with strictly increasing CDFs, such that all quantiles are unique. This ensures that the interval is truly equal-tailed, with $\mathrm{ETI}_\alpha(F) = [a, b]$ implying that $\mathbb{P}_F(Y < a) = \mathbb{P}_F(Y > b) = \frac{\alpha}{2}$. When F admits a Lebesgue density, but some quantiles are not unique, this property continues to hold.

However, care is needed when interpreting equal-tailed intervals for discrete distributions. As a simple example, let $\alpha = 0.2$ and consider the distribution G on $\mathbb{N}_0$ that assigns probability 0.1, 0.4, 0.4, and 0.1 to 0, 1, 2, and 3, respectively. Since neither the $\frac{\alpha}{2}$- nor the $(1 - \frac{\alpha}{2})$-quantile are unique, there are four possible equal-tailed intervals, as listed in Table 3.1. The distribution G illustrates that the coverage of an equal-tailed interval does not always equal $1-\alpha$, and may differ among the valid intervals. Moreover, $[0, 3]$ is not a guaranteed coverage interval in the sense of Section 3.2, as it is unnecessarily long. A natural idea is to issue recommendations for such cases, e.g. 'report the shortest available interval' or 'report the interval with the highest coverage'. However, consistent scoring functions for the ETI_α functional cannot be used to ensure that forecasters follow such further guidelines, since by the definition of consistency, any valid report attains the same expected score.

3.4 Shortest interval (SI)

Instead of defining an interval at the coverage level $1-\alpha$ via fixed quantiles, the shortest of these intervals is often sought. Specifically, a *shortest interval* (SI) at level α of F is any member of the set

$$\mathrm{SI}_\alpha(F) := \arg\min_{[a,b]\in \mathsf{A}} \{b-a \mid F(b)-F(a-)\geq 1-\alpha\}. \tag{3.5}$$

The shortest interval is never longer than an equal-tailed interval, and in general the two types of intervals differ from each other. To see this we follow Askanazi et al. (2018, Appendix) and consider a distribution F on $\mathsf{O}=[0,\infty)$ with strictly decreasing Lebesgue density, so that $\mathrm{SI}_\alpha(F)=[0,T_{1-\alpha}(F)]$, whereas $\mathrm{ETI}_\alpha(F)=[T_{\alpha/2}(F),T_{1-\alpha/2}(F)]$ with a lower endpoint that is strictly positive. However, for distributions with a symmetric, strictly unimodal Lebesgue density the two types of intervals are both unique and agree. If a distribution has multiple shortest intervals, then neither of them needs to be an equal-tailed interval.

As noted in Askanazi et al. (2018), loss functions that have been proposed for interval forecasts fail to be strictly consistent for the SI_α functional, since they are usually tailored to the ETI_α functional. The question whether the SI_α functional is elicitable thus remains unanswered, and Askanazi et al. (2018) formulate desiderata for possible scoring functions. A first result in this direction is discussed in Section 4.2 of Frongillo and Kash (2019), who show that the SI_α functional fails to be elicitable relative to classes $\mathcal{F}$ that contain piecewise uniform distributions. In the following we show non-elicitability for more general classes of distributions, and we also treat discrete distributions on $\mathbb{N}_0$. We start by studying level sets.

Proposition 3.3. *(i) The functional* SI_α *has the CxLS property.*

(ii) If the class $\mathcal{F}$ consists of distributions with continuous CDFs only, then SI_α *has the CxLS* property.*

Proof. Let $F_0,F_1\in\mathcal{F}$, and suppose that $[a,b]\in\mathrm{SI}_\alpha(F_0)\cap\mathrm{SI}_\alpha(F_1)$. Set $F_\lambda:=\lambda F_1+(1-\lambda)F_0$ and note that for all $\lambda\in(0,1)$ and all $s,t\in\mathbb{R}$ we have

$$F_\lambda(t)-F_\lambda(s-)=\lambda\left(F_1(t)-F_1(s-)\right)+(1-\lambda)\left(F_0(t)-F_0(s-)\right). \tag{3.6}$$

In particular, $F_\lambda(b)-F_\lambda(a-)\geq 1-\alpha$ and $[a,b]\in\mathrm{SI}_\alpha(F_\lambda)$, as otherwise (3.6) yields a contradiction to our initial assumption. This proves part (i).

Now let $F_0,F_1\in\mathcal{F}$ have continuous CDFs. Since $(s,t)\mapsto F(t)-F(s)$ is a continuous function for all $F\in\mathcal{F}$, we must have $F(b)-F(a)=1-\alpha$ for every $[a,b]\in\mathrm{SI}_\alpha(F)$. Suppose $[a',b']\in\mathrm{SI}_\alpha(F_0)\cap\mathrm{SI}_\alpha(F_1)$, as otherwise there is nothing to show, and let $\lambda\in(0,1)$ and $[a,b]\in\mathrm{SI}_\alpha(F_\lambda)$ be given. By the first part of the proof

$$\mathrm{len}(\mathrm{SI}_\alpha(F_1))=\mathrm{len}(\mathrm{SI}_\alpha(F_0))=b'-a'=b-a. \tag{3.7}$$

Furthermore, $F_\lambda(b) - F_\lambda(a) = 1 - \alpha$, and we see from (3.6) that $F_0(b) - F_0(a) \geq 1 - \alpha$ or $F_1(b) - F_1(a) \geq 1 - \alpha$ must hold. Suppose the first of these inequalities is satisfied. Then equality must hold since the strict inequality $F_0(b) - F_0(a) > 1 - \alpha$ is a contradiction to (3.7). This yields $[a, b] \in \mathrm{SI}_\alpha(F_0)$ and via (3.6) we obtain $F_1(b) - F_1(a) = 1 - \alpha$. The same reasoning applies if the second inequality holds. Taken together this gives $[a, b] \in \mathrm{SI}_\alpha(F_1) \cap \mathrm{SI}_\alpha(F_0)$, which proves part (ii). □

The subsequent example shows that the CxLS* property can be violated for discrete distributions.

Example 3.4. Let $\alpha \in (0, \frac{1}{3})$, and let $k \geq 1$ be an integer. Let $\varepsilon \in (0, \frac{\alpha}{3})$ and $\delta \in (0, \varepsilon)$. Let F_0 and F_1 be distributions on $\mathbb{N}_0$ that assign mass $\varepsilon + \delta$ to $k - 1$ and mass $1 - \alpha - \varepsilon$ to k. Furthermore, F_0 and F_1 assign mass $\varepsilon + \delta$ and $\varepsilon - \delta$, respectively, to $k + 1$. Then $\mathrm{SI}_\alpha(F_0) = \{[k - 1, k], [k, k + 1]\}$, $\mathrm{SI}_\alpha(F_1) = \{[k - 1, k]\}$, and for $\lambda \in [0, \frac{1}{2}]$ we have $\mathrm{SI}_\alpha(F_\lambda) = \mathrm{SI}_\alpha(F_0) \supsetneq \mathrm{SI}_\alpha(F_1)$. Therefore, SI_α does not have the CxLS* property relative to any convex class $\mathcal{F}$ that includes F_0 and F_1. ◇

As the construction extends to all $\alpha \in (0, 1)$, we obtain the following result.

Theorem 3.5. *Let $k \geq 1$ be an integer, and let $\mathcal{F}$ be a class of probability measures on $\mathbb{N}_0$ that contains all unimodal distributions with mode k. Then the SI_α functional is not elicitable relative to $\mathcal{F}$.*

We turn to classes of distributions with Lebesgue densities, so that the SI_α functional has the CxLS* property, and a more refined analysis proves useful. First we take up an example in Section 4.2 of Frongillo and Kash (2019).

Example 3.6. Given $\alpha \in (0, \frac{3}{5})$, we define distributions F_0 and F_1 via the piecewise uniform densities

$$f_0(x) = (1 - \alpha)\mathbb{1}_{[0,1]}(x) + \frac{\alpha}{3}\mathbb{1}_{[2,5]}(x) \quad \text{and} \quad f_1(x) = \frac{1 - \alpha}{2}\mathbb{1}_{[0,2]}(x) + \frac{\alpha}{3}\mathbb{1}_{[2,5]}(x),$$

so that $\mathrm{SI}_\alpha(F_0) = [0, 1]$ and $\mathrm{SI}_\alpha(F_1) = [0, 2]$, respectively. As $\mathrm{SI}_\alpha(F_\lambda) = [0, 2]$ for all $\lambda \in (0, 1)$, we conclude from Theorem 1.11 that the SI_α functional fails to be elicitable relative to convex classes of distributions that contain F_0 and F_1. ◇

As noted, Example 3.6 applies in situations where the class $\mathcal{F}$ includes all distributions with piecewise constant densities. As this assumption may be too restrictive in practice, we proceed to demonstrate non-elicitability based on substantially more flexible criteria.

Condition 3.7. The distribution F admits a Lebesgue density, and there are numbers $a < b$ and $\varepsilon > 0$ such that $\mathrm{SI}_\alpha(F) = [a, b]$, $F(b) = F(b + \varepsilon)$, and if $\beta < \alpha$, then $\mathrm{len}(\mathrm{SI}_\beta(F)) > \mathrm{len}(\mathrm{SI}_\alpha(F)) + \frac{1}{2}\varepsilon$.

Loosely speaking, this condition requires that there are ‘gaps’ on the right- and left-hand side of the shortest interval at level α, while every shortest interval for a level $\beta < \alpha$ is notably longer than the one at level α.

Theorem 3.8. *If the class $\mathcal{F}$ contains the location-scale family of a distribution satisfying Condition 3.7, along with its finite mixtures, then the* SI_α *functional is not elicitable relative to $\mathcal{F}$.*

Proof. We proceed by constructing suitable convex combinations as in Example 3.6. Specifically, let F_0 satisfy Condition 3.7, and without loss of generality assume that $\mathrm{SI}_\alpha(F_0) = [0, b]$ for some $b > 0$. Define F_1 via

$$F_1(x) := F_0\left(\frac{b}{b + \frac{1}{2}\varepsilon}\, x\right)$$

and set $F_\lambda := \lambda F_1 + (1-\lambda)F_0$. We proceed to show that $[0, b+\frac{1}{2}\varepsilon] \in \mathrm{SI}_\alpha(F_\lambda)$ for all $\lambda \in (0, 1]$, which allows us to apply Theorem 1.11 and conclude non-elicitability.

Clearly, $\mathrm{SI}_\alpha(F_1) = [0, b+\frac{1}{2}\varepsilon]$, and since $F_0(b) = F_0(b+\varepsilon)$ it holds that $F_\lambda(b+\frac{1}{2}\varepsilon) - F_\lambda(0) = 1-\alpha$ for $\lambda \in (0,1)$. For a contradiction, suppose there are $\lambda \in (0,1)$ and $a_\lambda \le b_\lambda$ with $F_\lambda(b_\lambda) - F_\lambda(a_\lambda) \ge 1-\alpha$ and $b_\lambda - a_\lambda < b + \frac{1}{2}\varepsilon$. Since $\mathrm{SI}_\alpha(F_1) = [0, b+\frac{1}{2}\varepsilon]$ it cannot be true that $F_1(b_\lambda) - F_1(a_\lambda) \ge 1-\alpha$ and so $F_0(b_\lambda) - F_0(a_\lambda) > 1-\alpha$ must hold, for a contradiction to the final part of Condition 3.7. Consequently, $\mathrm{SI}_\alpha(F_\lambda) = [0, b+\frac{1}{2}\varepsilon]$ for all $\lambda \in (0,1]$, and the proof is complete. □

A comparable result showing the non-elicitability of SI_α is given in Theorem 4.16(i) of Fissler et al. (2020). The main difference to Theorem 3.8 is that they consider a different class $\mathcal{F}$ and allow for scoring functions which take values in the extended real numbers $\mathbb{R} \cup \{-\infty, \infty\}$.

Although Condition 3.7 might seem technical, suitable distributions F can be constructed under rather weak assumptions. For instance, assume $\alpha < \frac{1}{2}$, and let the class $\mathcal{F}$ contain some compactly supported distribution, along with the respective location-scale family, and all finite mixtures thereof. Then constructing an F that satisfies Condition 3.7 is straightforward. A more restrictive requirement is the identity $F(b) = F(b+\varepsilon)$, as it rules out distributions with strictly positive densities. The existence of strictly consistent scoring functions relative to classes of distributions of this type, including but not limited to the important case of the finite mixture distributions with Gaussian components, remains an open problem.

3.5 Modal interval (MI)

In stark contrast to shortest and equal-tailed intervals, we turn to a type of interval that seeks to maximize coverage, subject to constraints on length.

Specifically, given any $c > 0$, a *modal interval* (MI) of length $2c$ of F is any member of the set

$$\mathrm{MI}_c(F) = \arg\max_{[a,b]\in \mathsf{A}} \{F(b) - F(a-) \mid b - a \le 2c\}. \tag{3.8}$$

If F has a strictly unimodal Lebesgue density, then the modal interval shrinks towards the mode as $c \to 0$. For distributions on $\mathbb{N}_0$ the modal interval even agrees with the mode if $c < \frac{1}{2}$. These connections highlight the fact that the MI_c functional is a location statistic, whereas the shortest and equal-tailed intervals contain information on both location and spread simultaneously.

In what follows, separate discussions for classes $\mathcal{F}$ of continuous and discrete distributions will be warranted. For distributions on $\mathbb{N}_0$, the length of the modal interval will effectively be $\lfloor 2c \rfloor$, since expanding it further cannot add probability mass. In this situation, it is convenient to consider $c \geq 0$, substitute $2c = k$ where $k \in \mathbb{N}_0$, and encode the interval via its *lower endpoint* functional l_k, so that $\mathrm{MI}_{k/2}(F) = \{[x, x+k] \mid x \in l_k(F)\}$. Then

$$S(x,y) = -\mathbb{1}(x \leq y \leq x+k) \tag{3.9}$$

is a strictly consistent scoring function for the functional l_k on the class of all distributions on $\mathbb{N}_0$. In particular, the l_k and $\mathrm{MI}_{k/2}$ functionals are elicitable. In the special case $k = 0$, l_0 is the mode functional and (3.9) becomes $S(x,y) = -\mathbb{1}(x = y)$, the familiar zero-one or misclassification loss. Lambert and Shoham (2009) and Gneiting (2017) demonstrate that for distributions with finitely many outcomes, zero-one loss is essentially the only consistent scoring function for the mode functional. We extend this result to all integers $k \geq 0$, showing that k-zero-one-loss (3.9) is essentially the only strictly consistent scoring function for the l_k and $\mathrm{MI}_{k/2}$ functionals.

Theorem 3.9. *Let $k \geq 0$ be an integer, and let $\mathcal{F}$ be a class of probability measures on $\mathbb{N}_0$ that contains all distributions with finite support. Then any scoring function that is strictly consistent for the l_k functional relative to the class $\mathcal{F}$ is equivalent to k-zero-one-loss* (3.9)*.*

Proof. Let $k \geq 0$ be an integer, and suppose that S is a strictly consistent scoring function for the functional l_k relative to $\mathcal{F}$. To facilitate the presentation, we introduce the alternative notation $S(M, y)$ for $S(x_M, y)$, where $x_M \in \mathbb{N}_0$ denotes the lower endpoint of an interval $M \in \mathsf{A}$, with $\mathsf{A} = \{[x, x+k] \mid x \in \mathbb{N}_0\}$. We proceed in three steps.

Step 1 We show that S is of the form

$$S(x,y) = g(x,y)\mathbb{1}(x \leq y \leq x+k) + h(y) \tag{3.10}$$

for functions $g : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{R}$ and $h : \mathbb{N}_0 \to \mathbb{R}$.

To this end, let $M_0, M_1 \in \mathsf{A}$ such that $M_0 \cap M_1 = \emptyset$. For a contradiction, suppose that the mapping $\varphi : \mathbb{N}_0 \to \mathbb{R}$ defined via $\varphi(y) = S(M_0, y) - S(M_1, y)$ is non-zero on $U := (M_0 \cup M_1)^c \cap \mathbb{N}_0$. We first treat the case where $\varphi(y) = c$ for all $y \in U$ and some $c \in \mathbb{R}\backslash\{0\}$. If $c > 0$ let F_0 be the uniform distribution on M_0 and for all $n \in \mathbb{N}$ let F_n be the uniform distribution on some set $U_n \subset U$ with $|U_n| = 2nk$. If we define $G_n := \frac{1}{n}F_0 + (1 - \frac{1}{n})F_n$, then $\mathrm{MI}_{k/2}(G_n) = \mathrm{MI}_{k/2}(F_0) = M_0$ for all $n \in \mathbb{N}$. Since $\int \varphi(y)\,\mathrm{d}G_n(y) \to c > 0$ for $n \to \infty$, we obtain a contradiction to the strict consistency of S. A similar argument applies if $c < 0$. Consequently, φ cannot be constant on U, i.e. there are $i_0, i_1 \in U$ such that $\varphi(i_0) \neq \varphi(i_1)$.

Now set $I := \{i_0, i_1\}$. As the class $\mathcal{F}$ contains all distributions with finite support, we can find probability measures $F_0, F_0', F_1 \in \mathcal{F}$ that satisfy the following three conditions:

(i) There exists a $\lambda^* \in (0,1)$ such that for $F_\lambda := \lambda F_1 + (1-\lambda)F_0$ and $F_\lambda' := \lambda F_1 + (1-\lambda)F_0'$

$$\mathrm{MI}_{k/2}(F_\lambda) = \mathrm{MI}_{k/2}(F_\lambda') = \begin{cases} M_0, & \lambda < \lambda^*, \\ M_1, & \lambda > \lambda^*. \end{cases}$$

(ii) F_0 and F_0' coincide outside of I.

(iii) $\int_I \varphi(y)\,\mathrm{d}F_0(y) \neq \int_I \varphi(y)\,\mathrm{d}F_0'(y)$.

To see this, define F_0 and F_0' via the probabilities $F_0(\{j\}) = F_0'(\{j\}) = 1/(k+2)$ for $j \in M_0$ and

$$F_0(\{i_0\}) = F_0'(\{i_1\}) = \frac{1}{2(k+2)} + \varepsilon, \quad \text{and} \quad F_0(\{i_1\}) = F_0'(\{i_0\}) = \frac{1}{2(k+2)} - \varepsilon \tag{3.11}$$

for some $\varepsilon \in (0, 1/(2(k+2)))$. Condition (ii) is immediate and (iii) follows from the fact that $\varphi(i_0) \neq \varphi(i_1)$. Moreover, letting F_1 be the uniform distribution on M_1 ensures (i).

Consider the integrated score difference

$$\Delta(F, G, \lambda) := \int S(M_0, y) - S(M_1, y)\,\mathrm{d}(\lambda G + (1-\lambda)F)(y),$$

which is linear in $\lambda \in [0, 1]$. The strict consistency of S in concert with (i) yields $\Delta(F_0, F_1, 0) < 0$, $\Delta(F_0', F_1, 0) < 0$, and $\Delta(F_0, F_1, 1) = \Delta(F_0', F_1, 1) > 0$. As $\Delta(F_0, F_1, 0)$ $\neq \Delta(F_0', F_1, 0)$ by (ii) and (iii), the two linear mappings $\lambda \mapsto \Delta(F_0, F_1, \lambda)$ and $\lambda \mapsto \Delta(F_0', F_1, \lambda)$ must have distinct roots. This implies that one of the two mappings does not vanish at λ^*, in contradiction to the consistency of S. Consequently, $\varphi = 0$ on U such that we can conclude $S(M_0, y) = S(M_1, y)$ for all $y \in (M_0 \cup M_1)^c$. By varying the disjoint intervals $M_0, M_1 \in \mathsf{A}$, we obtain that for all $y \in \mathbb{N}_0$ the values $S(M, y)$ are the same for all $M \in \mathsf{A}$ with $y \notin M$. This yields that there exists a function $h : \mathbb{N}_0 \to \mathbb{R}$ such that S is of the form (3.10).

Step 2 Now we prove that $y \mapsto g(x, y)$ is constant on $[x, x+k]$. As before, we use the notation $g(M, y)$ for $g(x_M, y)$, where $x_M \in \mathbb{N}_0$ is the lower endpoint of $M \in \mathsf{A}$. For $k = 0$ there is nothing to show, so let $k > 0$. For a contradiction, suppose there is an $M_0 \in \mathsf{A}$ such that $y \mapsto g(M_0, y)$ is not constant on M_0, i.e. there are $i_2, i_3 \in M_0$ such that $g(M_0, i_2) \neq g(M_0, i_3)$. This ensures that we can choose an interval $M_1 \in \mathsf{A}$, with $M_1 \cap M_0 = \emptyset$, and distributions $F_0, F_0', F_1 \in \mathcal{F}$ that satisfy conditions (i), (ii), and (iii) in Step 1, for $I = \{i_2, i_3\}$. For example, we can choose F_0 and F_0' by using the uniform distribution on M_0 and modifying it at i_2 and i_3 as in (3.11), while ensuring M_1 is separated from M_0 by a sufficiently large gap. As in Step 1 we obtain $\Delta(F_0, F_1, 0) \neq \Delta(F_0', F_1, 0)$ such that the mappings $\lambda \mapsto \Delta(F_0, F_1, \lambda)$ and $\lambda \mapsto \Delta(F_0', F_1, \lambda)$ have distinct roots. This is a contradiction to the consistency of S and proves that $y \mapsto g(M_0, y)$ is constant on M_0. We can thus replace $g(x, y)$ in (3.10) by $\tilde{g}(x)$ for some function $\tilde{g} : \mathbb{N}_0 \to \mathbb{R}$.

Step 3 It remains to be shown that $\tilde{g}$ reduces to a negative constant. To this end, consider $M_0 \in \mathsf{A}$ and $M_1 \in \mathsf{A}$ and assume that $\tilde{g}(M_0) < \tilde{g}(M_1)$. Due to the specific form of (3.10) we have

$$\mathbb{E}_F\left[S(M_0, Y) - S(M_1, Y)\right] = \tilde{g}(M_0)\mathbb{P}_F(Y \in M_0) - \tilde{g}(M_1)\mathbb{P}_F(Y \in M_1)$$

for all $F \in \mathcal{F}$. However, due to the strict consistency of S this expression must be negative if $M_0 \in \mathrm{MI}_{k/2}(F)$ and positive if $M_1 \in \mathrm{MI}_{k/2}(F)$, for the desired contradiction. Therefore $\tilde{g}$ reduces to a constant, and using once more the consistency of S, we see that this constant is negative. The proof is complete. □

For distributions with Lebesgue densities we encode MI_c via its *midpoint* functional m_c so that $\mathrm{MI}_c(F) = \{[x-c, x+c] \mid x \in m_c(F)\}$, where $c > 0$. Under this convention,

$$S(x,y) := -\mathbb{1}(x - c \le y \le x + c) \tag{3.12}$$

is a strictly consistent scoring function for m_c on the class of distributions with Lebesgue densities, whence m_c and MI_c are elicitable. In the limit as $c \to 0$, the scoring function (3.12) becomes zero almost everywhere and thus cannot be strictly consistent for any functional. Heinrich (2014) shows that there are no alternative scoring functions, so the mode fails to be elicitable relative to sufficiently rich classes of distributions with densities. Further aspects are treated in Dearborn and Frongillo (2020).

The following theorem demonstrates, perhaps surprisingly, that c-zero-one-loss (3.12) is essentially the only strictly consistent scoring function for the m_c and MI_c functionals.

Theorem 3.10. *Let $c > 0$, and let $\mathcal{F}$ be a class of probability measures on $\mathbb{R}$ that contains all distributions having Lebesgue densities with compact support. Then any scoring function that is strictly consistent for the m_c functional relative to $\mathcal{F}$ is almost everywhere equal to a scoring function which is equivalent to c-zero-one-loss* (3.12)*.*

Proof. We sketch this proof only, as it proceeds in the very same three steps as the proof of Theorem 3.9. Specifically, let $c > 0$, and let S be a strictly consistent scoring function for the functional m_c relative to $\mathcal{F}$. In Step 1, we show that S is almost everywhere of the form

$$S(x,y) = g(x,y)\mathbb{1}(x - c \le y \le x + c) + h(y)$$

for $\mathcal{F}$-integrable functions $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$. In Step 2 we prove that g reduces to a function $\tilde{g}$ in the variable x only, and in Step 3 we demonstrate that $\tilde{g}$ reduces to a negative constant. The technical details are analogous to those in the above proof of Theorem 3.9, with the only difference that the set I is now an interval and the statements hold Lebesgue almost everywhere. □

We complete this section by connecting modal and shortest intervals. While these are conceptually different types of intervals, a comparison of (3.5) and (3.8) shows that the SI_α and MI_c functionals relate via their defining optimization problems. Specifically, the $\mathrm{SI}_\alpha(F)$ functional is a solution to the constrained optimization problem

$$\min_{[a,b]\in\mathsf{A}} (b - a) \quad \text{such that} \quad F(b) - F(a-) \ge 1 - \alpha,$$

while the MI_c functional is a solution to

$$\max_{[a,b]\in\mathsf{A}} (F(b) - F(a-)) \quad \text{such that} \quad b - a \le 2c.$$

Consequently, if either $\mathrm{len}(\mathrm{SI}_\alpha(F)) = 2c$ or $\mathbb{P}_F(Y \in \mathrm{MI}_c(F)) = 1 - \alpha$, one condition implies the other, and $\mathrm{MI}_c(F) = \mathrm{SI}_\alpha(F)$ holds. It remains unclear whether this connection can be exploited to construct strictly consistent scoring functions for the SI_α functional on suitably restrictive, special classes of distributions.

3.6 Discussion

Of the three types of predictive intervals presented in this chapter, the equal-tailed and modal intervals are elicitable, and we have discussed the corresponding classes of strictly consistent scoring functions. In contrast, the shortest interval functional fails to be elicitable relative to classes of distributions of practical relevance. We thus provide a negative answer to the questions raised by Askanazi et al. (2018) concerning the existence of suitable loss functions for the shortest interval. Importantly, there is yet to find a way of setting incentives for forecasters to report their true shortest intervals. Equal-tailed intervals are preferable due to their elicitability, in concert with other considerations, such as the intuitive connection to quantiles and equivariance under strictly monotone transformations (Askanazi et al., 2018, p. 961).

The modal interval admits a unique strictly consistent scoring function relative to comprehensive classes of both discrete and continuous distributions, up to equivalence. This appears to be a rather special situation, as functionals studied in the extant literature either fail to be elicitable, or admit rich classes of genuinely distinct consistent scoring functions (Gneiting, 2011a; Steinwart et al., 2014; Fissler and Ziegel, 2016; Frongillo and Kash, 2019). It would be of great interest to gain an understanding of conditions under which consistent scoring functions are essentially unique.

As illustrated, interval forecasts are best suited for continuous distributions, and may exhibit counter-intuitive properties in discrete settings. In particular, in the discrete case it may be unavoidable that the coverage probability of a perfect forecast exceeds the nominal level $1-\alpha$. This raises problems when assessing interval calibration with the methods of Christoffersen (1998), since asymptotically the null hypothesis of frequency calibration will then be rejected even under perfectly correct forecasts. Modifying the null hypothesis to nominal coverage greater than or equal to $1-\alpha$ is not a remedy, since such a test does not have any power against forecast intervals with too high coverage. Consequently, tests for correct forecast specification as in Christoffersen (1998) can be problematic when data fail to be well-approximated by continuous distributions, such as in the case of retail sales. Fortunately, comparative evaluation via consistent scoring functions remains valid and unaffected (Czado et al., 2009; Kolassa, 2016).

In general, we agree with Askanazi et al. (2018) that probabilistic forecasts in the form of predictive distributions are preferable to interval forecasts. They constitute the gold standard of forecasting as they contain all the available distributional information and thereby allow for optimal decision making. Additionally, well-understood and powerful evaluation methods, e.g. proper scoring rules, are available (Dawid, 1986; Gneiting and Raftery, 2007; Gneiting et al., 2007; Gneiting and Katzfuss, 2014). Probabilistic forecasts can be issued in a number of distinct formats, ranging from the use of parametric distributions, such as in the Bank of England Inflation Report (Clements, 2004), to Monte Carlo samples from predictive models. However, in many applications the full potential of the probabilistic framework remains unexplored, owing to established conventions or technical and methodological difficulties. In these settings, reporting a collection of predictive intervals, which amounts to a collection of quantiles in case of ETIs, can be a reasonable alternative, which is already commonly used, e.g. in the Global Energy Forecasting Competition 2014 (Hong et al., 2016) or the COVID-19 Forecasting Hub (Bracher et al., 2020). The findings of this chapter support forecast evaluation methods at this intermediate stage in the transition from point to probabilistic forecasting.

4 Scoring functions for point process characteristics

This chapter transfers the idea of comparative forecast evaluation via consistent scoring functions to the point process setting. More precisely, we show elicitability, and derive consistent scoring functions, for a variety of common point process characteristics such as the intensity or the K-function. This complements a similar approach of Heinrich et al. (2019) by considering a more general setting, which leads to a variety of novel results. We discuss two existing methods for model comparison in statistical seismology and find that our results can be interpreted as generalizations of these ideas.

We begin with a brief discussion of the point process evaluation setting in Section 4.1 and then rigorously introduce scoring functions for point process characteristics in Section 4.2. Section 4.3 explains how some of these scoring functions connect to existing methods for model assessment and Section 4.4 presents simulation results that illustrate finite sample behavior. Section 4.5 concludes with an outlook towards further refinements and a discussion.

4.1 Different point process scenarios

This section discusses how forecast evaluation via scoring functions, as lined out in Section 1.5, transfers to the point process setting. The key difference to the usual framework is the treatment of time, since points of a (spatio-)temporal point process occur at random times instead of fixed measurement dates. This feature has to be addressed in order to use the full information of any sampled point pattern. For clarity of presentation, we distinguish three different point process scenarios, based on common applications:

Scenario A (purely spatial) In this scenario the process is defined on either a single domain (Scenario A1), or non-overlapping subdomains with no (or little) dependence between them (Scenario A2). Examples include the points which an observer fixates in an image (Barthelmé et al., 2013) or the locations of trees in a forest (Stoyan and Penttinen, 2000). Stationarity is a common simplifying assumption in this context.

Scenario B (purely temporal) In this scenario, there is no spatial component and the process consists of points in time only. Examples are the arrival times of e-mails (Fox et al., 2016) or times of infections with a disease (Schoenberg et al., 2019). In this special setting the directional character of time allows for a distinct interpretation and treatment, as detailed below.

Scenario C (spatio-temporal) Additional to the spatial component, processes in this scenario possess a temporal component, which could be discrete (Scenario C1) or continuous (Scenario C2). Examples are locations of crime hotspots in a city (Mohler et al., 2011) or earthquakes over time in a specific region (Ogata, 1998; Zhuang et al., 2002).

In order to compare forecasts in each of these scenarios, the ideas of Section 1.5 can be used as follows. Let Φ be a point process and S a scoring function such that $S(a, \Phi)$ is the score of the report $a \in \mathsf{A}$. Moreover, assume that S is strictly consistent for a statistical property of point processes, e.g. the intensity. Section 4.2 discusses which properties and scoring functions are available, as well as technical details. In this situation, two forecasts a and a^* can be compared based on the sign of the expectation $\mathbb{E}\left[S(a, \Phi) - S(a^*, \Phi)\right]$, where, due to the consistency of S, negative values provide evidence that report a is superior to a^*, while positive values support the opposite conclusion.

Given point process realizations $\Phi_1, \ldots, \Phi_n$ the difference in expected scores can be accessed via the average score difference as in (1.10). Although the idea of score differences is the same for all three scenarios, the detailed estimation may vary among them. In particular, if we want to estimate the uncertainty inherent in the realized score differences this task depends on whether the process has a continuous or a discrete time component.

Discrete time Assume that the point process data is sampled at fixed points in time, i.e. it can be modeled by a sequence $(\Phi_t)_{t\in\mathbb{N}}$ adapted to a filtration $(\mathcal{H}_t)_{t\in\mathbb{N}}$. Moreover, let $(R_t)_{t\in\mathbb{N}}$ and $(R_t^*)_{t\in\mathbb{N}}$ be two forecast sequences. This setting includes the special case of i.i.d. realizations and relates to Scenario C1 as well as variants of Scenario A. Since the score differences $(S(R_t, \Phi_t) - S(R_t^*, \Phi_t))_{t\in\mathbb{N}}$ form a sequence of real-valued random variables, the ideas of Section 1.5, particularly the comparative testing framework of Nolde and Ziegel (2017) and its asymptotic results, are directly applicable. This yields an approach to forecasts evaluation and model selection which is suitable if dependence is non-existent or not of central interest for forecast evaluation.

Continuous time If we consider point processes in Scenario C2 or Scenario B, then temporal dependence between the points becomes an essential feature of the process and can also be object of the forecast. Apart from that, it has to be accounted for in estimation and testing, since it will affect asymptotic results. To see this, assume for simplicity that Φ is a purely temporal process observed over a time period $[0, T]$ with $0 < t_1 < \cdots < t_k < T$ denoting the corresponding arrival times. In a stepwise forecast setting, where the reports R_i and R_i^* can adapt to previous arrivals $t_1, \ldots, t_{i-1}$ (see e.g. Subsection 4.2.5), this yields a realized score difference

$$\Delta_T(R, R^*) = \sum_{i=1}^{n(T)} \left(S_i(R_i, t_i) - S_i(R_i^*, t_i)\right), \tag{4.1}$$

where $n(T) := \Phi((0, T])$ is the random number of points in $[0, T]$. The score difference $\Delta_T(R, R^*)$ is a sum of a random number of random variables, usually called a random sum. This perspective connects score estimation for temporal point processes to the theory of total claim amount in insurance, see e.g. Mikosch (2009) and Embrechts et al. (1997). An example illustrates possible results and difficulties.

Example 4.1 (Independently marked renewal process)**.** Let Φ be a renewal process (Daley and Vere-Jones, 2003, Chapter 4) with i.i.d. marks $(X_i)_{i\in\mathbb{N}}$ following a distribution $F \in \mathcal{F}$. By definition, the time between the i-th and $(i+1)$-th point is given by Y_i, where $(Y_i)_{i\in\mathbb{N}}$ is an i.i.d. sequence of strictly positive random variables with distribution G and finite first moment. Suppose furthermore that the marks are unpredictable, i.e. the sequences $(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ are independent. An observation φ of Φ on $[0,T]$ then consists of the points $\{t_1,\dots,t_n\}$ and the corresponding marks $\{X_1,\dots,X_n\}$.

Assume two forecasters are asked to report their beliefs F^1, F^2 and G^1, G^2 concerning the mark distribution F and the inter-arrival distribution G, respectively. A strictly consistent scoring function for this pair of distributions is then given by

$$S((F,G),\varphi) = \sum_{i=1}^{n} S_1(F,X_i) + S_2(G,t_i - t_{i-1}),$$

where S_1, S_2 are strictly proper scoring rules and $n = n(T) = \varphi((0,T])$ is the number of points. The difference in average scores of the two competing forecasters is $\Delta_T := S((F^1,G^1),\varphi) - S((F^2,G^2),\varphi)$ and it can be interpreted as the sum of the marks of a new process Φ^* which has the same ground process as Φ and marks

$$M_i = S_1(F^1,X_i) - S_1(F^2,X_i) + S_2(G^1,t_i - t_{i-1}) - S_2(G^2,t_i - t_{i-1}).$$

In general the marks of Φ^* are no longer independent of the ground process, although the mark sequence $(M_i)_{i\in\mathbb{N}}$ is i.i.d. Suppose we assume that the forecasters possess equal predictive ability, in the sense that their scores are equal in expectation. This implies that $(M_i)_{i\in\mathbb{N}}$ has mean zero and we further assume that it has strictly positive variance σ^2. If we combine $\Phi^*((0,T])/t \to 1/\mathbb{E}Y_1$, the elementary renewal theorem, with a special case of a result due to Anscombe (see Rényi (1957, Theorem 1) and Gut (2012, Theorem 2.3)), then we obtain that

$$\frac{\Delta_T}{\sigma\sqrt{n(T)}}$$

is asymptotically standard normal as $T \to \infty$. This asymptotic result allows for DM tests as in Section 1.5 in this simplified renewal process setting. ◇

The assumptions on the underlying process Φ in the previous example are too restrictive for most real-world applications. Hence, more general asymptotic results for the score difference (4.1) for $T \to \infty$ are desirable to assess the uncertainty of the forecast evaluation task. One possible approach to this problem relies on limit theorems for randomly indexed processes due to Anscombe (1952), in particular random central limit theorems: If the number of points $n(T)$ satisfies a weak law of large numbers, then under Anscombe's condition, we only have to ensure that the sequence $(S_i(R_i,t_i) - S_i(R_i^*,t_i))_{i\in\mathbb{N}}$ satisfies a central limit theorem in order to obtain asymptotic normality of (4.1) after suitable scaling. Random central limit theorems in this spirit are available for strong mixing (Lee, 1997), ψ-weakly dependent (Hwang and Shin, 2012), and m-dependent (Shang, 2012) sequences.

4.2 Consistent scoring functions for point patterns

This section explores which consistent scoring functions are available if the observations are finite point patterns in a set $\mathcal{X} \subset \mathbb{R}^d$. We start with general principles which connect to existing theory and then derive scoring functions for a variety of popular point process characteristics, which relate to Scenarios A, B, and C (see Section 4.1).

4.2.1 Technical context

A finite *point process* Φ is a random element in the space $\mathbb{M}_0 = \mathbb{M}_0(\mathcal{X})$ of finite counting measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra of $\mathcal{X}$, cf. Daley and Vere-Jones (2003) for details. Hence, our observation domain is $\mathbb{M}_0$ in this context and we shall denote a set of probability measures on $\mathbb{M}_0$ by $\mathcal{P}$ and the distribution of Φ by P_Φ. As in Section 1.1, a *functional* is a mapping $\Gamma : \mathcal{P} \to \mathsf{A}$, where A is an action domain. A mapping $S : \mathsf{A} \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ is a *scoring function* if

$$\mathbb{E}_P S(a, \Phi) = \bar{S}(a, P) = \int_{\mathbb{M}_0} S(a, \varphi) \, \mathrm{d}P(\varphi)$$

is well-defined for all $a \in \mathsf{A}$ and $P \in \mathcal{P}$, i.e. we drop the requirement of integrability used in the previous chapters. The reason for this is that the results of this chapter often rely on proper scoring rules, which are by definition $\bar{\mathbb{R}}$-valued functions, see Section 1.4. *Elicitability* of Γ as well as *(strict) consistency* of S are defined as in Section 1.1 via inequality (1.1).

For ease of presentation and practical implementation, we will usually state how the score of a realization $\varphi = \sum_{i=1,\ldots,n} \delta_{y_i} \in \mathbb{M}_0$ is computed from an enumeration of its points, i.e. from the set $\{y_1, \ldots, y_n\}$ if $n = |\varphi|$. For $n = 0$ no points occurred, so the set is empty. To make this meaningful, we will ensure that for spatial processes all scoring functions are independent of the enumeration of points (see also Daley and Vere-Jones (2003, Section 5.3)). For temporal processes we use the natural enumeration which orders the points from smallest to largest.

In light of the results of Section 1.2, constructing simple elicitable functionals of point processes is straightforward: Since point processes induce real-valued random variables in many ways, the expectations of these random variables (provided they are well-defined) will be elicitable functionals.

Example 4.2 (Expected number of points)**.** Given a set $B \in \mathcal{B}(\mathcal{X})$, the ($\mathbb{N}_0$-valued) random variable $\Phi(B)$ denotes the random number of points of Φ in B. If the functional $\Gamma_B : \mathcal{P} \to \mathbb{R}$ given by $\Gamma_B(P) = \mathbb{E}_P \Phi(B)$ is well-defined, Theorem 1.6 shows that it is elicitable with Bregman scoring function

$$S_B(x, \varphi) = b(x, \varphi(B)) = -f(x) - \nabla f(x)^\top (\varphi(B) - x),$$

where $f : [0, \infty) \to \mathbb{R}$ is a strictly convex function. ◇

This construction is not limited to the expected number of points in a set, but works for any combination of elicitable functional (e.g. expectation) and point process feature (e.g. number of points). More precisely, let O be an observation domain, $g : \mathbb{M}_0 \to \mathsf{O}$ a measurable mapping and $g(\mathcal{P}) := \{P \circ g^{-1} \mid P \in \mathcal{P}\}$. Due to the transformation principle

(Proposition 1.10), the functional $\Gamma(P) := T(P \circ g^{-1})$ is elicitable whenever $T : g(\mathcal{P}) \to \mathsf{A}$ is elicitable. We recover Example 4.2 by choosing $T(F) = \mathbb{E}_F Y$ and $g(\varphi) = \varphi(B)$. The elicitability of other "simple" properties like finite-dimensional distributions and void probabilities is a straightforward consequence of Proposition 1.10.

Example 4.3 (Void probability)**.** For any fixed set $B \in \mathcal{B}(\mathcal{X})$ the functional Γ defined via $\Gamma(P) = \mathbb{P}_P(\Phi(B) = 0)$ is elicitable. $\diamond$

Example 4.4 (Point process integrals)**.** Fix some measurable functions $f_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, m$ for $m \in \mathbb{N}$. Define $g : \mathbb{M}_0 \to \mathbb{R}^m$ via

$$g(\varphi) = \left(\int f_1 \, \mathrm{d}\varphi, \ldots, \int f_m \, \mathrm{d}\varphi \right)^\top = \left(\sum_{x_i \in \varphi} f_1(x_i), \ldots, \sum_{x_i \in \varphi} f_m(x_i) \right)^\top$$

and let $T = \mathrm{id}_{g(\mathcal{P})}$. Then the finite-dimensional distribution functional $\Gamma_{f_1,\ldots,f_m}(P) = T(P \circ g^{-1})$ is an elicitable property of the point process Φ. $\diamond$

As illustrated by the previous examples, different choices for T and g in Proposition 1.10 lead to a wide variety of different functionals and consistent scoring functions. In Heinrich et al. (2019) the main idea is to choose T as the identity on $g(\mathcal{P})$ (see also Example 4.4). Two distributional models $P, Q \in \mathcal{P}$ of the process Φ can then be compared based on realizations by comparing $P \circ g^{-1}$ and $Q \circ g^{-1}$ via a proper scoring rule. The mapping $g : \mathbb{M}_0 \to \mathsf{O}$ is selected to be an estimator of some quantity of interest, e.g. a kernel-based intensity estimator. Since the distributions of such estimators will usually not be explicitly available, approximating the values of the scoring rule via simulations becomes necessary. Moreover, as different $P \in \mathcal{P}$ may lead to the same law $P \circ g^{-1}$, this approach hinges on the ability of g to discriminate between two distributions P and Q.

Instead of following this approach, we develop strictly consistent scoring functions in order to compare certain point process characteristics $\Gamma : \mathcal{P} \to \mathsf{A}$. Important examples for Γ include the point process distribution and the intensity measure. This allows for a direct comparison of the characteristic Γ which includes distributional models $P \in \mathcal{P}$ as a special case. In contrast, comparison in Heinrich et al. (2019) always depends on specific aspects of the distributions in $\mathcal{P}$ which are determined via the 'estimator choice' g. The discrimination ability of our approach depends on how similar the property values $\Gamma(P)$ and $\Gamma(Q)$ (e.g. the intensity measures) are. In particular, knowledge of the distribution P is not needed as long as $\Gamma(P)$ is available. In cases where Γ can be computed explicitly, this avoids possibly high computational costs owing to point process simulations.

4.2.2 Density and distribution – general processes

This subsection constructs consistent scoring functions for the whole distribution P_Φ of the finite point process Φ, which corresponds to the identity functional $\Gamma = \mathrm{id}_{\mathcal{P}}$. We obtain two results, corresponding to distinct representations of the distribution P_Φ. Although applicable to general processes, the main focus is Scenario A and related results for temporal point processes are given in Subsection 4.2.5.

For our first approach we use the fact that the law P_Φ of a finite point process on $\mathcal{X}$ can be equivalently represented by two sequences $(\Pi_k)_{k \in \mathbb{N}}$ and $(p_k)_{k \in \mathbb{N}_0}$. Each p_k specifies the probability of finding k points in a realization. The Π_k are symmetric probability measures on $\mathcal{X}^k$ which describe the distribution of any ordering of points, given k points are realized, see Daley and Vere-Jones (2003, Chapter 5.3) for details. We can thus construct scoring functions for the distribution of the process by combining scoring functions for all Π_k with a scoring function for the distribution $(p_k)_{k\in\mathbb{N}_0}$.

To state this result, we introduce the notion of *symmetric* scoring functions/rules, where $S : \mathsf{A} \times \mathbb{R}^n \to \bar{\mathbb{R}}$ is called symmetric if $S(a, y_1, \ldots, y_n) = S(a, y_{\pi(1)}, \ldots, y_{\pi(n)})$ holds for all $a \in \mathsf{A}$, $y \in \mathbb{R}^n$ and permutations π. Via this definition we ensure that the scoring functions of this subsection are independent of the enumeration of the realization of Φ.

Proposition 4.5. *Let $\mathcal{P}$ be a class of distributions of finite point processes, with $Q \in \mathcal{P}$ decomposed into $(\Pi_k^Q)_{k\in\mathbb{N}}$ and $(p_k^Q)_{k\in\mathbb{N}_0}$. Set $\mathcal{F}_k := \{\Pi_k^Q \mid Q \in \mathcal{P}\}$ and let $S_k : \mathcal{F}_k \times \mathcal{X}^k \to \bar{\mathbb{R}}$ be a symmetric (strictly) proper scoring rule for all $k \in \mathbb{N}$. Let S_0 be a (strictly) proper scoring rule on $\mathbb{N}_0$. Then the function $S : \mathcal{P} \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ defined via*

$$S(((\Pi_k^Q)_{k\in\mathbb{N}}, (p_k^Q)_{k\in\mathbb{N}_0}), \{y_1, \ldots, y_n\}) = S_n(\Pi_n^Q, y_1, \ldots, y_n) + S_0((p_k^Q)_{k\in\mathbb{N}_0}, n)$$

for $n \in \mathbb{N}$ and $S(((\Pi_k^Q)_{k\in\mathbb{N}}, (p_k^Q)_{k\in\mathbb{N}_0}), \emptyset) = S_0((p_k^Q)_{k\in\mathbb{N}_0}, 0)$ is a consistent scoring function for the distribution of the point process Φ. It is strictly consistent if S_0 and $(S_k)_{k\in\mathbb{N}}$ are strictly proper.

Proof. The result follows by decomposing the expectation $\bar{S}(Q, P_\Phi)$ into expectations on the sets $\{\Phi = n\}$ for $n \in \mathbb{N}$ and using (strict) propriety of S_n on each set. □

Although Proposition 4.5 allows for a general choice of scoring rules for the distributions $(\Pi_k)_{k\in\mathbb{N}}$, probability densities are often more convenient vehicles, especially when multivariate distributions are of interest. Subsection 1.4.1 holds several common choices of scoring rules for density forecasts.

To define the density of a point process we follow Daley and Vere-Jones (2003) and let P_0 denote the distribution of the Poisson point process with unit rate on some bounded domain $\mathcal{X} \subset \mathbb{R}^d$. If $P \in \mathcal{P}$ is absolutely continuous with respect to P_0, then the Radon-Nikodým density $\mathrm{d}P/\mathrm{d}P_0$ exists and can be regarded as the density of P. It can be computed via the identity $\mathrm{d}P/\mathrm{d}P_0(\varphi) = \exp(|\mathcal{X}|) j_k(y_1, \ldots, y_k)/k!$, where $|\mathcal{X}|$ denotes the Lebesgue measure of $\mathcal{X}$, $y_1, \ldots, y_k$ are the points of $\varphi \in \mathbb{M}_0$ and the (symmetric) function j_k defined via

$$j_k(x_1, \ldots, x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_k = k! p_k \,\mathrm{d}\Pi_k(x_1, \ldots, x_k) \tag{4.2}$$

is the k-th *Janossy density* of Φ. For $k = 0$ this is interpreted as $j_0 = p_0$. The value $j_k(x_1, \ldots, x_k)$ can be understood as the *likelihood* of k points materializing, one of them in each of the distinct locations $x_1, \ldots, x_k \in \mathcal{X}$. We refer to Daley and Vere-Jones (2003, Chapter 7.1 and 5.3) for further details.

In principle, the previous discussion and Subsection 1.4.1 allow us to obtain scoring functions for the point process distribution P based on its densities $(j_k^P)_{k\in\mathbb{N}_0}$. However, two important difficulties have to be addressed in the point process setting. Firstly, explicit expressions for $(j_k)_{k\in\mathbb{N}_0}$ are usually hard to determine and known only for some models, see Daley and Vere-Jones (2003, Chapter 7.1) and Examples 4.6 and 4.7 below. Secondly, even if explicit expressions are available, calculating the realized scores can pose major challenges. For instance, the use of the pseudospherical score relies on the norm $\|\mathrm{d}P/\mathrm{d}P_0\|_\alpha$ which necessitates computing $(k!)^{-1}\int j_k(x_1,\dots,x_k)^\alpha\,\mathrm{d}x_1\dots\mathrm{d}x_k$ for all $k\in\mathbb{N}$. We will thus only discuss two choices of scoring functions, the logarithmic and the Hyvärinen score, see Subsection 1.4.1 for their definitions.

Logarithmic score Assume that for all distributions $Q\in\mathcal{P}$ the Janossy densities $(j_k^Q)_{k\in\mathbb{N}_0}$ are well-defined. Due to the strict consistency of the logarithmic score (Gneiting and Raftery, 2007), the function $S:\mathcal{P}\times\mathbb{M}_0\to\bar{\mathbb{R}}$ defined via

$$S((j_k^Q)_{k\in\mathbb{N}_0},\{y_1,\dots,y_n\}) = -\log(j_n^Q(y_1,\dots,y_n)) \tag{4.3}$$

for $n\in\mathbb{N}$ and $S((j_k^Q)_{k\in\mathbb{N}_0},\emptyset) := -\log(j_0^Q)$ is a strictly consistent scoring function for the distribution of the point process Φ. The term $-|\mathcal{X}|+\log(n!)$ can be omitted, since it is independent of the report $(j_k^Q)_{k\in\mathbb{N}_0}$. This choice recovers the log-likelihood of the distribution Q from the perspective of consistent scoring functions.

Example 4.6 (Poisson point process)**.** Let Φ be an inhomogeneous Poisson point process with intensity $\lambda:\mathcal{X}\to[0,\infty)$. It is well-known that Φ admits the density

$$j_{n,\lambda}(y_1,\dots,y_n) = \Big(\prod_{i=1}^n \lambda(y_i)\Big)\exp\Big(-\int_{\mathcal{X}}\lambda(y)\,\mathrm{d}y\Big)$$

for $n\in\mathbb{N}$, see e.g. Daley and Vere-Jones (2003, Chapter 7). In case $n=0$ the product is interpreted as one. When reporting the Poisson point process distribution P_Φ, (4.3) gives the score

$$S(P_\Phi,\{y_1,\dots,y_n\}) = -\sum_{i=1}^n \log\lambda(y_i) + \int_{\mathcal{X}}\lambda(y)\,\mathrm{d}y \tag{4.4}$$

for $n\in\mathbb{N}$ and $S(P_\Phi,\emptyset)=|\lambda|$, where $|\lambda|=\int_{\mathcal{X}}\lambda(y)\,\mathrm{d}y$. In the context of Proposition 4.5, the definition of Poisson point processes implies that the distribution of Φ can be decomposed via

$$\mathrm{d}\Pi_n(y_1,\dots,y_n) = \prod_{i=1}^n \frac{\lambda(y_i)}{|\lambda|}\,\mathrm{d}y_1\cdots\mathrm{d}y_n \quad\text{and}\quad p_n = \frac{|\lambda|^n}{n!}e^{-|\lambda|}.$$

When choosing all the scoring rules $(S_n)_{n\in\mathbb{N}}$ and S_0 as the logarithmic score, the scoring function S in Proposition 4.5 simplifies and agrees with (4.4). ◇

The previous example illustrates how the scoring functions for $(\Pi_k,p_k)_{k\in\mathbb{N}_0}$ and $(j_k)_{k\in\mathbb{N}_0}$ are connected in the case of Poisson point processes. In general, identity (4.2) shows that choosing the logarithmic score for $(S_n)_{n\in\mathbb{N}}$ and S_0 leads to identical scoring functions in (4.3) and Proposition 4.5. For other choices, relating both expressions is an open problem.

Hyvärinen score Apart from the Poisson point process, some other models admit explicit expressions for $(j_k)_{k\in\mathbb{N}_0}$ or the densities of $(\Pi_k)_{k\in\mathbb{N}}$, however often only up to an unknown normalizing constant. In this situation 0-homogeneous proper scoring rules for densities can be of use, as they allow for consistent evaluation of a density f on $\mathbb{R}^k$

without the normalizing constant being known. The most relevant scoring rule of this type is the *Hyvärinen score* given by

$$\mathrm{HyvS}(f, y) = \Delta \log f(y) + \frac{1}{2} \|\nabla \log f(y)\|^2,$$

where f is a twice differentiable density, see Hyvärinen (2005) and Subsection 1.4.1 for details.

Similar to the logarithmic score, we can transfer the Hyvärinen score to the point process setting. Since it is not obvious how differentiation of the density $\mathrm{d}P/\mathrm{d}P_0$, i.e. differentiation in the space $\mathbb{M}_0$, can be performed, we apply the score to $(j_k)_{k\in\mathbb{N}_0}$ directly. Hence, assume that for all $Q \in \mathcal{P}$ and $k \in \mathbb{N}$, j_k^Q satisfies the regularity conditions of the Hyvärinen score, as stated in Subsection 1.4.1. Then the function $S : \mathcal{P} \times \mathbb{M}_0 \to \mathbb{R}$ defined via

$$S((j_k^Q)_{k\in\mathbb{N}_0}, \{y_1, \dots, y_n\}) = -\mathrm{HyvS}(j_n^Q, y_1, \dots, y_n) \tag{4.5}$$

for $n \in \mathbb{N}$ and $S((j_k^Q)_{k\in\mathbb{N}_0}, \emptyset) := 0$ is a consistent scoring function for the distribution of the point process Φ. Observe that we cannot achieve strict consistency for S, since the probability of $|\Phi| = n$ is proportional to j_n (see identity (4.2)) and thus not accessible to the Hyvärinen score.

Example 4.7 (Gibbs point process)**.** Stemming from theoretical physics, Gibbs processes are a popular tool to model particle interactions. They are defined via their Janossy densities

$$j_n(y_1, \dots, y_n) = C(\theta) \exp\left(-\theta U(y_1, \dots, y_n)\right),$$

where U represents the point interactions, θ is a parameter relating to the temperature, and C is the partition function, which ensures that the collection $(j_k)_{k\in\mathbb{N}_0}$ is properly normalized, see e.g. Daley and Vere-Jones (2003, Chapter 5.3) and Chiu et al. (2013, Chapter 5.5). It is in general difficult to find expressions for C or even approximate it, hence the Hyvärinen score might seem attractive to evaluate models based on $(j_k)_{k\in\mathbb{N}_0}$. Plugging j_n into (4.5) gives

$$S((j_k)_{k\in\mathbb{N}_0}, \{y_1, \dots, y_n\}) = \theta \left(-\Delta U(y_1, \dots, y_n) + \frac{\theta}{2} \|\nabla U(y_1, \dots, y_n)\|^2\right)$$

for $n \in \mathbb{N}$, where the derivatives are computed with respect to the coordinates of the vector $(y_1, \dots, y_n) \in (\mathbb{R}^d)^n$. The simplest choice for interactions is to restrict U to first- and second-order terms

$$U(y_1, \dots, y_n) = \sum_{i=1}^{n} l(y_i) + \sum_{i,j=1}^{n} \psi\left(\|y_i - y_j\|^2\right)$$

for $l : \mathbb{R}^d \to \mathbb{R}$ and $\psi : [0, \infty) \to [0, \infty)$ with $\psi(0) = 0$, see e.g. Daley and Vere-Jones (2003, Chapter 5.3). To apply the Hyvärinen score in this setting, l and ψ have to satisfy some regularity conditions, detailed in Subsection 1.4.1 and Hyvärinen (2005), and in particular admit second order derivatives almost everywhere. The soft-core models for

ψ introduced in Ogata and Tanemura (1984) satisfy this condition, while their hard-core model for ψ is not even continuous. An additional regularity issue is that all possible pairs of densities f_1, f_2 have to satisfy $(\partial_i \log f_1(y))f_2(y) \to 0$, whenever $\|y\| \to \infty$. Since Ogata and Tanemura (1984) consider a constant l, their processes cannot satisfy this condition. Choosing l such that the j_n decay fast enough for $\|y\| \to \infty$ can be a solution to this problem. ◇

4.2.3 Intensity and moment measures

As one of the key characteristics of point processes, the intensity measure, or more general moment measures, can be interpreted as analogons to the moments of a univariate random variable. To construct scoring functions for these measures, let $\mathcal{X} \subset \mathbb{R}^d$ be bounded and $\mathcal{M}_f = \mathcal{M}_f(\mathcal{X})$ a set of finite measures on $\mathcal{X}$. We call $\Lambda^* := \Lambda/|\Lambda|$, where $|\Lambda| := \Lambda(\mathcal{X})$, the *normalized measure* of a finite measure $\Lambda \in \mathcal{M}_f$.

Intensity measure The intensity measure defined via $B \mapsto \mathbb{E}\Phi(B)$, and usually denoted via Λ, quantifies the expected number of points in a set $B \in \mathcal{B}(\mathcal{X})$, see e.g. Daley and Vere-Jones (2003) and Chiu et al. (2013). It is one of the central tools to describe average spatial point process behavior and thereby relates to applications in Scenario A or Scenario C. For a fixed Borel set B, the expected number of points was already discussed in Example 4.2, thus here we focus on scoring functions for the full measure. We begin by considering the normalized intensity measure, since it is a probability measure.

Proposition 4.8. *Set $\mathcal{F} := \{\Lambda^* \mid \Lambda \in \mathcal{M}_f\}$ and let $S' : \mathcal{F} \times \mathcal{X} \to \bar{\mathbb{R}}$ be a (strictly) proper scoring rule. The scoring function $S : \mathcal{F} \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ defined via*

$$S(\Lambda^*, \{y_1, \dots, y_n\}) = \sum_{i=1}^{n} S'(\Lambda^*, y_i)$$

for $n \in \mathbb{N}$ and $S(\Lambda^*, \emptyset) = 0$ is consistent for the normalized intensity measure. It is strictly consistent if S' is strictly proper.

Proof. Let $Q \in \mathcal{M}_f$ and Φ be a point process with intensity measure $\Lambda \in \mathcal{M}_f$ and distribution $P_\Phi \in \mathcal{P}$. Using Campbell's theorem, the difference in expected scores is

$$\begin{aligned} \bar{S}(Q^*, P_\Phi) - \bar{S}(\Lambda^*, P_\Phi) &= \int \sum_{x_i \in \varphi} \big(S'(Q^*, x_i) - S'(\Lambda^*, x_i)\big) \,\mathrm{d}P_\Phi(\varphi) \\ &= \int S'(Q^*, x) - S'(\Lambda^*, x) \,\mathrm{d}\Lambda(x) \\ &= |\Lambda| \left(\bar{S}'(Q^*, \Lambda^*) - \bar{S}'(\Lambda^*, \Lambda^*)\right) \geq 0, \end{aligned}$$

where the inequality follows from the propriety of S'. If the difference is zero and S' is strictly proper, this gives $Q^* = \Lambda^*$, showing that S is strictly consistent for the normalized intensity measure. □

In principle, it is possible to define scoring rules for non-normalized measures, as well. Hendrickson and Buehler (1971) use a constant extension of scoring rules to the cone induced by a set of probability measures, in order to connect to homogeneous convex functions. However, as the proof of Proposition 4.8 illustrates, information concerning the intensity measure can be accessed only after normalization. Since the total mass $|\Lambda| = \mathbb{E}\Phi(\mathcal{X})$ is an elicitable property of Φ (see Example 4.2), combining this information with Λ^* leads to a consistent scoring function for the (unnormalized) intensity. This follows from an application of the revelation principle (Proposition 1.9).

Corollary 4.9. *Set $\mathcal{F} := \{\Lambda^* \mid \Lambda \in \mathcal{M}_f\}$ and let $S' : \mathcal{F} \times \mathcal{X} \to \bar{\mathbb{R}}$ be a (strictly) proper scoring rule. Let $b : [0, \infty) \times [0, \infty) \to \mathbb{R}$ be a (strictly) consistent Bregman function, as defined in* (1.3)*. The scoring function $S : \mathcal{M}_f \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ defined via*

$$S(\Lambda, \{y_1, \dots, y_n\}) = \sum_{i=1}^{n} S'(\Lambda^*, y_i) + cb\big(|\Lambda|, n\big)$$

for $n \in \mathbb{N}$ and $S(\Lambda, \emptyset) = cb(|\Lambda|, 0)$ for $c > 0$ is consistent for the intensity measure. It is strictly consistent if S' is strictly proper and b is strictly consistent.

Example 4.10. As an important special case, assume that each $\Lambda \in \mathcal{M}_f$ admits a density λ with respect to Lebesgue measure. Using the quadratic score for b and the logarithmic score for S' (see Section 1.4.1), the strictly consistent scoring function of Corollary 4.9 becomes

$$S(\Lambda, \{y_1, \dots, y_n\}) = -\sum_{i=1}^{n} \log(\lambda(y_i)) + n \log |\Lambda| + c\,(|\Lambda| - n)^2$$

for some $c > 0$. If we choose the logarithmic score for b and $c = 1$, then S leads to the same score as obtained in (4.4) for Poisson point process reports. This is further discussed in Subsection 4.3.3. Simulation experiments in Subsection 4.4.1 illustrate how S compares different intensity forecasts. ◇

The choice of the constant $c > 0$ in Corollary 4.9 is irrelevant for (strict) consistency of the scoring function S. However, since S evaluates a mixture of shape and normalization of the intensity, where c balances the magnitudes of the scoring components, a careful choice will likely be crucial in applications. An alternative is to compare the realized scores for all c in some interval of suitable values.

Moment Measures If the intensity is interpreted as the first moment of a point process, moment measures generalize this notion to higher moments. Thereby, they are useful tools to quantify point interactions for processes occurring in Scenario A or C. Strictly consistent scoring functions for these measures can be constructed analogously. For $n \in \mathbb{N}$, let $\mathcal{M}_f^n = \mathcal{M}_f(\mathcal{X}^n)$ be the set of finite Borel measures on $\mathcal{X}^n$. For positive measurable functions $f : \mathcal{X}^n \to (0, \infty)$ the *n-th moment measure* $\mu^{(n)}$ and the *n-th factorial moment measure* $\alpha^{(n)}$ are defined via the relations

$$\mathbb{E}\left[\sum_{x_1,\dots,x_n \in \Phi} f(x_1, \dots, x_n)\right] = \int_{\mathcal{X}^n} f(x_1, \dots, x_n)\, \mathrm{d}\mu^{(n)}(x_1, \dots, x_n)$$

$$\text{and} \quad \mathbb{E}\left[\sum_{x_1,\dots,x_n \in \Phi}^{\neq} f(x_1, \dots, x_n)\right] = \int_{\mathcal{X}^n} f(x_1, \dots, x_n)\, \mathrm{d}\alpha^{(n)}(x_1, \dots, x_n),$$

respectively, see e.g. Chiu et al. (2013) and Daley and Vere-Jones (2003). Here $\Sigma^{\neq}$ denotes summation over all n-tuples that contain distinct points of Φ. Using the notion of *factorial product* defined via

$$m^{[n]} := \begin{cases} m(m-1)(m-2)\cdots(m-n+1) & , m \geq n \\ 0 & , m < n \end{cases}$$

for $m, n \in \mathbb{N}$ we obtain the concise representations $\mu^{(n)}(B^n) = \mathbb{E}\Phi(B)^n$ and $\alpha^{(n)}(B^n) = \mathbb{E}\Phi(B)^{[n]}$ for $B \in \mathcal{B}(\mathcal{X})$, see e.g. Daley and Vere-Jones (2003, Chapter 5). The next result follows by using the same arguments as in the proof of Proposition 4.8 together with the revelation principle (Proposition 1.9).

Proposition 4.11. *Set $\mathcal{F}^n := \{P^* \mid P \in \mathcal{M}_f^n\}$, let $S : \mathcal{F}^n \times \mathcal{X}^n \to \bar{\mathbb{R}}$ be a (strictly) proper scoring rule and $b : [0,\infty) \times [0,\infty) \to \mathbb{R}$ a (strictly) consistent Bregman function.*

(i) The function $S_1 : \mathcal{M}_f^n \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ defined via

$$S_1(\mu, \{y_1, \dots, y_m\}) = \sum_{x_1,\dots,x_n \in \{y_1,\dots,y_m\}} S(\mu^*, x_1, \dots, x_n) + cb(\mu(\mathcal{X}^n), m^n)$$

for $m \in \mathbb{N}$ and $S_1(\mu, \emptyset) = cb(\mu(\mathcal{X}^n), 0)$ for $c > 0$ is a consistent scoring function for the n-th moment measure.

(ii) The function $S_2 : \mathcal{M}_f^n \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ defined via

$$S_2(\alpha, \{y_1, \dots, y_m\}) = \sum^{\neq}_{x_1,\dots,x_n \in \{y_1,\dots,y_m\}} S(\alpha^*, x_1, \dots, x_n) + cb(\alpha(\mathcal{X}^n), m^{[n]})$$

for $m \geq n$ and $S_2(\alpha, \{y_1, \dots, y_m\}) = cb(\alpha(\mathcal{X}^n), 0)$ for $m < n$ and with $c > 0$ is a consistent scoring function for the n-th factorial moment measure.

Both S_1 and S_2 are strictly consistent if S is strictly proper and b is strictly consistent.

In many cases of interest $\alpha^{(n)}$ is absolutely continuous with respect to Lebesgue measure on $\mathcal{X}^n$ and its density $\varrho^{(n)}$ is the *product density*, see e.g. Chiu et al. (2013). A (strictly) consistent scoring function for $\varrho^{(n)}$ can be obtained from Proposition 4.11 (ii) by choosing S to be a strictly proper scoring rule for densities, as in the next example.

Example 4.12. Let $n = 2$ and consider the product density $\varrho^{(2)}$ of a stationary and isotropic point process. In this situation, $\varrho^{(2)}$ depends on the point distances only, i.e. it can be represented via $\varrho^{(2)}(x_1, x_2) = \varrho_0^{(2)}(\|x_1 - x_2\|)$ for some $\varrho_0^{(2)} : [0,\infty) \to [0,\infty)$. Analogous to Example 4.10, we can use the quadratic score for b and the logarithmic score for S in Proposition 4.11 (ii). This gives the strictly consistent scoring function

$$\begin{aligned} S(\varrho^{(2)}, \{y_1, \dots, y_m\}) = &- \sum^{\neq}_{x_1,x_2 \in \{y_1,\dots,y_m\}} \log(\varrho_0^{(2)}(\|x_1 - x_2\|)) \\ &+ m^{[2]} \log|\varrho^{(2)}| + c\,(|\varrho^{(2)}| - m^{[2]})^2, \end{aligned}$$

where $c > 0$ is some scaling constant. Simulation experiments in Subsection 4.4.2 show how S compares different product density forecasts. ◇

4.2.4 Summary statistics

Summary statistics of point processes are central tools to quantify point interactions such as clustering or inhibition, hence they are typically used in Scenario A or Scenario C1. This subsection constructs strictly consistent scoring functions for the most frequently used statistics, the K- and L-function. Throughout we assume that Φ is a *stationary* point process on $\mathbb{R}^d$, i.e. any translation of the process by $x \in \mathbb{R}^d$, which we denote via Φ_x, has the same distribution as Φ. This implies that the intensity measure of Φ is a multiple of Lebesgue measure and can be represented via some $\lambda > 0$, see e.g. Chiu et al. (2013, Section 4.1).

A common way to describe a stationary point process is to consider its properties in the neighborhood of $x \in \mathbb{R}^d$, given that x is a point in Φ. Due to stationarity, the location of x is irrelevant and thus it is usually referred to as the "typical point" of Φ. The technical tool to describe the behavior around this point is the *Palm distribution* of Φ, denoted via $\mathbb{P}_0$ for probabilities and $\mathbb{E}_0$ for expectations. It satisfies the defining identity

$$\lambda\,|W|\,\mathbb{E}_0 f(\Phi) = \mathbb{E}\left(\sum_{x\in\Phi\cap W} f(\Phi_{-x})\right)$$

for all measurable $f : \mathbb{M}_0 \to \mathbb{R}$ such that the expectations are finite and it is independent of the observation window $W \in \mathcal{B}(\mathbb{R}^d)$, see e.g. Illian et al. (2008, Chapter 4). Denote the d-dimensional ball of radius $r > 0$ around zero via $B_r = B(0, r)$. The *K-function* of Φ is defined via

$$K : (0,\infty) \to [0,\infty), \quad r \mapsto \frac{\mathbb{E}_0\Phi\left(B_r\backslash\{0\}\right)}{\lambda},$$

and it quantifies the mean number of points in a ball around the "typical point" of Φ, see e.g. Chiu et al. (2013, Chapter 4) and Illian et al. (2008, Chapter 4). As pointed out by Heinrich et al. (2019) it is unclear whether it is possible to express $K(r)$ as an elicitable property in order to employ Proposition 1.10. However, we show that we can proceed as for the intensity measure (see Subsection 4.2.3) and construct a strictly consistent scoring function for reports consisting of both, the K-function and the intensity (see also Example 4.2.) Our point process property of interest is thus $\Gamma(P) := (\lambda_P, K_P)$, where the subscript denotes the dependence of the quantities on the distribution $P \in \mathcal{P}$ of the process Φ.

To derive consistent scoring functions let us fix some $r > 0$ and assume for now that λ_P is known and that instead of data we directly observe the Palm distribution of Φ. In this simplified situation, $K_P(r)$ is just an expectation with respect to $\mathbb{P}_0$, hence "consistent scoring functions" for it are of the Bregman form

$$b(x,\varphi) = -f(x) - f'(x)(\varphi(B_r\backslash\{0\}) - \lambda_P x) \tag{4.6}$$

for a convex function $f : (0,\infty) \to \mathbb{R}$, see Theorem 1.6 and Example 4.2. This is because $\mathbb{E}_0^P b(x,\Phi) \geq \mathbb{E}_0^P b(K_P(r),\Phi)$ holds for all $x \geq 0$ and $P \in \mathcal{P}$. To arrive at a strictly consistent scoring function for the functional Γ three steps remain: Firstly, we have to include a consistent scoring function for the first component of Γ, i.e. the intensity. Moreover, we need to integrate (4.6) with respect to $r > 0$ in order to evaluate the entire K-function. Finally, we have to account for the fact that we are not observing $\mathbb{P}_0$, but only points of Φ on some closed and bounded observation window $W \subset \mathbb{R}^d$. Hence, we need to compute the expected score $\mathbb{E}_0 b(x, \Phi)$ via an expectation of Φ on W. Such problems lead to edge corrections, i.e. additional terms to account for the fact that (unobserved) points outside of W affect estimation near the boundary of W, see e.g. Chiu et al. (2013, Chapter 4.7) for details. Since (4.6) is linear in φ, edge corrections for the expected score are equivalent to edge corrections for the expectation $\mathbb{E}_0\Phi(B_r\backslash\{0\})$, which are well-known in the context of K-function estimation. Taken together we obtain the following result.

Proposition 4.13. *Let $b_1, b_2 : [0,\infty) \times [0,\infty) \to \mathbb{R}$ be (strictly) consistent Bregman functions and $w : (0,\infty) \to [0,\infty)$ a weight function. Define $\mathcal{C} := \{K_P \mid P \in \mathcal{P}\}$, a set of possible K-functions and let the function κ satisfy $\mathbb{E}_P\kappa(B_r, \Phi) = \lambda_P \mathbb{E}_0^P \Phi(B_r\backslash\{0\})$ for all $P \in \mathcal{P}$. Then the function $S : ((0,\infty) \times \mathcal{C}) \times \mathbb{M}_0 \to \mathbb{R}$ defined via*

$$S((\lambda, K), \varphi) = b_1(\lambda, \varphi(W)|W|^{-1}) + \int_0^\infty b_2(\lambda^2 K(r), \kappa(B_r, \varphi))w(r)\,\mathrm{d}r$$

is consistent for the point process property Γ as long as the expectation of the integral is finite. It is strictly consistent if b_1 and b_2 are strictly consistent and w is strictly positive.

Proof. Using Theorem 1.6, the Fubini-Tonelli theorem, and

$$\mathbb{E}_P\kappa(B_r, \Phi) = \lambda_P \mathbb{E}_0^P \Phi(B_r\backslash\{0\}) = \lambda_P^2 K_P(r),$$

standard arguments show that the scoring function

$$S'((\lambda, h), \varphi) = b_1(\lambda, \varphi(W)|W|^{-1}) + \int_0^\infty b_2(h(r), \kappa(B_r, \varphi))w(r)\,\mathrm{d}r,$$

where $h : (0,\infty) \to (0,\infty)$ is an increasing function, is (strictly) consistent for the property $\Gamma'(P) := (\lambda_P, \lambda_P^2 K_P(r))$. An application of the revelation principle (Proposition 1.9) gives (strict) consistency for Γ. □

Similar to Corollary 4.9, this result blends two scoring components, namely the expected number of points and their distances. Hence, choosing suitable Bregman functions b_1 and b_2 in applications, again leads to issues of balancing the magnitudes of different scoring components.

A similarly delicate question is the choice of κ. Relevant choices result from the construction of estimators for the K-function, which are often based on dividing κ by an estimator for λ^2, see e.g. Chiu et al. (2013). A common choice is

$$\kappa_{\mathrm{st}}(B_r, \varphi) := \sum_{x_1, x_2 \in \varphi \cap W}^{\neq} \frac{\mathbb{1}_{B_r}(x_2 - x_1)}{|W_{x_1} \cap W_{x_2}|},$$

where $W_z := \{x + z \mid x \in W\}$ is the shifted observation window and r is such that $|W \cap W_z|$ is positive for all $z \in B_r$, see e.g. Illian et al. (2008, Chapter 4.3) and Chiu et al. (2013, Chapter 4.7). An alternative arises via minus-sampling, i.e. by reducing the observation window W in order to reduce edge effects. It is given by

$$\kappa_{\text{minus}}(B_r, \varphi) := \frac{1}{|W|} \sum\nolimits^{\neq}_{x_1,x_2 \in \varphi \cap W,\, x_2 \in W \ominus r} \mathbb{1}_{B_r}(x_2 - x_1),$$

where $W \ominus r := \{x \mid B(x,r) \subset W\}$ is the reduced observation window. For other choices of κ, most notably for isotropic point processes, see Chiu et al. (2013, Chapter 4.7).

Practitioners usually rely on the L-function, a modification of the K-function, which is defined via $L(r) = \sqrt[d]{K(r)/b_d}$ for $r \geq 0$, where $b_d := |B_1|$. It satisfies $L(r) = r$ for the Poisson point process, and thus normalizes the K-function such that it is independent of the dimension d for a Poisson point process, see e.g. Chiu et al. (2013). A (strictly) consistent scoring function for the L-function follows immediately from Proposition 4.13 and another application of the revelation principle. The explicit formula follows by replacing the first component of b_2 by $\lambda^2 L(r)^d b_d$ in Proposition 4.13. The idea underlying the construction of scoring functions for the K- and L-function presented here can be transferred to other summary statistics for stationary point processes.

4.2.5 Density and distribution – temporal processes

This subsection turns to consistent scoring functions for temporal point processes and thereby relates to Scenario B, however, suitable adaptions to spatio-temporal point processes (Scenario C2) are straightforward. The key feature of temporal (point) processes is that the dimension "time" possesses a natural ordering, which allows for an intuitive conditioning on the past that can be used to obtain explicit temporal point process models.

In order to construct such models, the main tool is the probability of a new point in the process *conditional* on past points. The instantaneous rate of points occurring in the point process Φ is usually described via the *conditional intensity*

$$\lambda^*(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}\left[\Phi((t, t+\Delta t)) \mid \mathcal{H}_t\right]}{\Delta t}, \tag{4.7}$$

where $(\mathcal{H}_t)_{t \in \mathbb{R}}$ is the filtration generated by the history of Φ. Although $\lambda^*(t)$ is random, it is known conditional on Φ, hence a measurable mapping linking it to Φ allows for modeling as well as evaluation via consistent scoring functions. This mapping is usually based on the concept of *hazard functions* which also reflects a fruitful perspective in applications (Harte, 2015; Reinhart, 2018). We turn to an illustrative example and refer to Daley and Vere-Jones (2003, Chapter 7) and Daley and Vere-Jones (2008, Chapter 14) for further details.

Example 4.14 (Hawkes process)**.** This basic *self-exciting* point process model was proposed by Hawkes (1971) and is defined via the conditional intensity

$$\lambda^*(t) = \nu + \sum_{t_i < t} g(t - t_i),$$

where $\nu \geq 0$ is the *background rate*, $g : (0,\infty) \to [0,\infty)$ is the *triggering function*, and $(t_i)_{i \in \mathbb{N}}$ comprises the points of the point process Φ. For a review of its applications see for instance Reinhart (2018). ◇

Let Φ be a point process on $\mathbb{R}_+$ and consider an observation window $\mathcal{X} := [0, T]$ for some $T > 0$. Given a realization $0 < t_1 < \ldots < t_n$ of Φ the realized values of the conditional intensity can be computed for all $t \in \mathcal{X}$. More precisely, for a $t \in \mathcal{X}$ with $t_1 < \ldots < t_i \leq t < t_{i+1}$ we denote the realized value of λ^* at t via $\lambda^*(t \mid t_1, \ldots, t_i)$. Whenever there is no need to emphasize the dependence on $t_1, \ldots, t_i$ we use the simpler notation $\lambda^*(t)$. To ensure uniqueness, we follow Daley and Vere-Jones (2003) and assume that a left-continuous version of λ^* exists and is used.

Since the collection of all mappings $t \mapsto \lambda^*(t \mid t_1, \ldots, t_i)$ for all $i = 0, \ldots, n$ and all possible realizations $t_1, \ldots, t_n$ uniquely determines the distribution of Φ (Daley and Vere-Jones, 2003), comparing forecasts for the conditional intensity is equivalent to a comparison of forecasts for the distribution. The connection is the representation of the likelihood of $t_1, \ldots, t_n$ occurring in $[0, T]$ via

$$j_n(t_1, \ldots, t_n) = \Big(\prod_{i=1}^{n} \lambda^*(t_i) \Big) \exp\left(- \int_0^T \lambda^*(u) \, \mathrm{d}u \right), \tag{4.8}$$

where the product is interpreted as one if no points occur. Consequently, (strictly) consistent scoring functions for the conditional intensity can be obtained via the same arguments as in Subsection 4.2.2, as illustrated by the following example.

Example 4.15. As in Subsection 4.2.2, the logarithmic score is the most intuitive choice of strictly proper scoring rules for densities. Plugging (4.8) into (4.3) we see that the scoring function S given by

$$S(\lambda^*, \{t_1, \ldots, t_n\}) = - \sum_{i=1}^{n} \log\left(\lambda^*(t_i)\right) + \int_0^T \lambda^*(u) \, \mathrm{d}u$$

is strictly consistent for the conditional intensity. This recovers the log-likelihood of a temporal point process, see for instance Daley and Vere-Jones (2003) and Reinhart (2018). If Φ is a Poisson point process on $\mathbb{R}_+$ with intensity λ, its conditional intensity agrees with λ and S coincides with (4.4). Simulation experiments in Subsection 4.4.3 illustrate how S compares different conditional intensities of Hawkes processes. $\diamond$

Instead of using the likelihood (4.8) as a scoring function, an alternative is to proceed stepwise and rely on the distribution of the next point conditional on all previous points. After doing this for all points the resulting distribution is then evaluated via proper scoring rules. The conditional probability density and distribution function of the point t_i, given $t_1, \ldots, t_{i-1}$ are

$$f_i(t \mid t_1, \ldots, t_{i-1}) = \lambda^*(t \mid t_1, \ldots, t_{i-1}) \exp\left(- \int_{t_{i-1}}^{t} \lambda^*(u \mid t_1, \ldots, t_{i-1}) \, \mathrm{d}u \right) \quad \text{and}$$

$$F_i(t \mid t_1, \ldots, t_{i-1}) = 1 - \exp\left(- \int_{t_{i-1}}^{t} \lambda^*(u \mid t_1, \ldots, t_{i-1}) \, \mathrm{d}u \right),$$

where $t \in (t_{i-1}, \infty)$. For $i = 0$ the functions are unconditional and we use the convention $t_0 = 0$. These distributions give an equivalent characterization of the point process distribution P_Φ, see Daley and Vere-Jones (2003, Chapter 7.2) for details. Let $\mathcal{D}$ be the set of left-continuous positive mappings on $[0, T)$. Adding the scores for all intervals gives the following result.

Proposition 4.16. *Let $S_i : \mathcal{F} \times [0,\infty) \to \bar{\mathbb{R}}$ for $i \in \mathbb{N}$ be (strictly) proper scoring rules for densities and $S' : [0,1] \times \{0,1\} \to \bar{\mathbb{R}}$ a (strictly) proper scoring rule for Bernoulli distributions. Then the scoring function $S : \mathcal{D} \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ defined via*

$$S(\lambda^*, \{t_1, \ldots, t_n\}) = \sum_{i=1}^{n} S_i(f_i(\cdot \mid t_1, \ldots, t_{i-1}), t_i) + S'(1 - F_{n+1}(T \mid t_1, \ldots, t_n), 1)$$

for $n \in \mathbb{N}$ and $S(\lambda^*, \emptyset) = S'(1 - F_1(T), 1)$ is consistent for the conditional intensity restricted to $[0,T]$. It is strictly consistent, if all $(S_i)_{i\in\mathbb{N}}$ and S' are strictly proper.

Proof. By the tower property of conditional expectations, the expectation of S_i is the mean of the conditional expectation given $t_1, \ldots, t_{i-1}$ for every $i = 2, \ldots, n$. Hence, the reported conditional distributions are compared to the true conditional distributions in expectation. An analogous argument for S' shows (strict) consistency. □

The scoring function S' evaluates the forecast probability of an empty interval occurring after t_n. Since it is possible to choose a different scoring rule S_i for every point t_i, Proposition 4.16 leads to a greater variety of scoring functions for temporal point processes compared to the likelihood (4.8). Choosing the logarithmic score for all $(S_i)_{i\in\mathbb{N}}$ and S' recovers the scoring function of Example 4.15 and thereby connects both approaches.

Another benefit of Proposition 4.16 becomes apparent when considering marked point processes. In this setting, points t_i and corresponding marks κ_i are observed and the conditional intensity admits the decomposition $\lambda^*(t,\kappa) = \lambda_g^*(t)\psi^*(\kappa \mid t)$, where λ_g^* is the conditional intensity of the ground process and $\psi^*(\cdot \mid t)$ is the conditional density of the marks, see Daley and Vere-Jones (2003, Chapter 7.3) for details. If both functions are reported, a consistent scoring function for λ^* emerges by using S as given in Proposition 4.16 for λ_g^* and adding scores for the marks. More precisely, such a scoring function is of the form

$$\tilde{S}((\lambda_g^*, \psi^*), \{t_1, \ldots, t_n\}, \{\kappa_1, \ldots, \kappa_n\}) = S(\lambda_g^*, \{t_1, \ldots, t_n\}) + \sum_{i=1}^{n} S_i^{\mathrm{m}}(\psi^*(\cdot \mid t_i), \kappa_i),$$

where $(S_i^{\mathrm{m}})_{i\in\mathbb{N}}$ are (strictly) proper scoring rules for densities. These additional scores evaluate each density $\psi^*(\cdot \mid t_i)$ given a point t_i. A decomposed scoring function such as $\tilde{S}$ might be beneficial in applications where different scoring rules for the points and marks are reasonable. For instance, only some distributional properties of ψ^*, e.g. the tails, might be of interest and the $(S_i^{\mathrm{m}})_{i\in\mathbb{N}}$ can be tailored to emphasize this.

4.3 Review of extant methods for model comparison

This section discusses existing techniques for the evaluation of point process models and how they relate to consistent scoring functions. We focus on two model evaluation approaches common in statistical seismology, and discuss how they can be interpreted as scoring function-based comparison methods. Due to this emphasis, other concepts which do not primarily aim at comparative evaluation, e.g. diagnostic tools, are omitted.

A standard approach to model comparison consists of using the log-likelihood of a model, i.e. its log-density evaluated at the observations. Most prominently, information criteria such as the AIC or BIC build on this idea to assess relative quality of competing models, and they can also be used for point process models, as long as likelihoods are available. Information criteria connect naturally to proper scoring rules through their goodness-of-fit component which usually consists of a log-likelihood and can thus be interpreted as evaluating the logarithmic score (see Subsection 1.4.1) for the given model. The penalty component, which is computed from the number of fitted parameters, highlights the difference to proper scoring rules. It is a necessary correction for information criteria, as their comparison is in-sample, i.e. it relies on the same data which is used for model fitting. In contrast, comparative forecast evaluation via scoring functions/rules is ideally performed out-of-sample, i.e. using new data which was not previously used to fit models or issue forecasts, see also Gneiting and Raftery (2007, Section 7).

In the Bayesian setting a standard approach to model comparison is the use of *Bayes factors*, which indicate whether there is sufficient evidence for one model to be more likely than a competitor. Like information criteria they are closely connected to the logarithmic score, as discussed in Gneiting and Raftery (2007, Section 7). Marzocchi et al. (2012) employ Bayes factors to compare point process models in the setting of earthquake likelihood model testing (Subsection 4.3.2).

A further variant of the likelihood principle for point processes can be obtained from a combination with residual methods. In general, point process residuals form an empirical process arising from fitting a conditional intensity (see Subsection 4.2.5) to data. They can be used to assess goodness-of-fit and especially indicate in which regions a model fits well or poorly, see e.g. Bray and Schoenberg (2013) for a review. Clements et al. (2011) propose the use of deviance residuals to graphically compare two competing models for the conditional intensity of a spatio-temporal process. The method plots the log-likelihood ratio of two models for every set of a partition of the spatial domain and can thus be interpreted as a visualization of local differences in logarithmic score.

4.3.1 Information gain

Closely connected to log-likelihood methods is the information gain approach, introduced by Vere-Jones (1998) as a tool to compare temporal point processes (Scenario B). The main idea is to assess a model based on the event probabilities that it induces for a collection of intervals. Although the term 'information gain' is sometimes used in the context of spatial point processes, too (Rhoades et al., 2011; Strader et al., 2017), we focus on the temporal case and consider a spatial analogon in the next subsection.

The basis for the information gain is a partition of the interval $[0, T]$ into n subintervals with length δ_i for $i = 1, \ldots, n$. A distributional model for a point process Φ, i.e. a distribution $P \in \mathcal{P}$ can then be used to generate (conditional) probabilities for the event that at least one point occurs in interval i. Assume two collections of such probabilities $(p_i)_{i=1,\ldots,n}$ and $(q_i)_{i=1,\ldots,n}$ are given, where the former is computed from the model under consideration and the latter corresponds to a reference model, usually a homogeneous Poisson process. The time-normalized log-likelihood ratio between $(p_i)_{i=1,\ldots,n}$ and $(q_i)_{i=1,\ldots,n}$ is

$$\bar{\rho}_T = \frac{1}{T} \sum_{i=1}^{n} X_i \log\left(\frac{p_i}{q_i}\right) + (1 - X_i) \log\left(\frac{1 - p_i}{1 - q_i}\right), \tag{4.9}$$

where the X_i are binary random variables indicating whether or not interval i contains a point of Φ, see Daley and Vere-Jones (2003, Chapter 7.6). This term is called *mean information gain per unit time* by Vere-Jones (1998) and positive values are assumed to indicate improved forecast performance of the model $(p_i)_{i=1,\ldots,n}$ compared to the reference model. Consequently, (4.9) resembles differences of consistent scoring functions. A rigorous analysis of this connection is presented in Subsection 4.3.3.

When using the information gain method, the choice of suitable subintervals is crucial, as this will influence the performance measured by $\bar{\rho}_T$. The impact of different choices is difficult to assess, however, Vere-Jones (1998) and Daley and Vere-Jones (2003, Chapter 7.6) show that the maximal performance is independent of the choice of intervals. More precisely, let Φ be a stationary temporal point process with conditional intensity function λ^* and X_i Bernoulli variables with parameters p_i for $i = 1, \ldots, n$. If q_i is computed from a homogeneous Poisson point process with rate $\bar{\lambda}$ given by the mean intensity $\bar{\lambda} := \mathbb{E}[\lambda^*(0)]$, i.e. via $q_i = 1 - \exp(-\bar{\lambda}\delta_i)$, then $\mathbb{E}\bar{\rho}_T$ is bounded by

$$\mathcal{I} := \mathbb{E}[\lambda^*(0) \log(\lambda^*(0))] - \bar{\lambda} \log(\bar{\lambda}),$$

the *entropy gain per unit time* of the process. Moreover, $\mathbb{E}\bar{\rho}_T \to \mathcal{I}$ for a refining sequence of partitions of $[0, T]$. Due to this result, $\mathcal{I}$ can be interpreted as the 'predictability' of the process Φ by measuring the potential for concentration in contrast to the homogeneous Poisson process with rate equal to the mean intensity, see e.g. Vere-Jones (1998) and Daley and Vere-Jones (2004). This quantification of predictability via $\mathcal{I}$ can also be used for a goodness-of-fit criterion, where $\mathcal{I}$ is estimated from the data via the information gain (4.9) and compared to the true value for a given model, see Daley and Vere-Jones (2004) and Harte and Vere-Jones (2005) for details.

4.3.2 Earthquake likelihood model testing

When considering spatial point processes, as in our Scenario A, explicitly computable likelihoods are often not available. A major exception are Poisson point processes, which motivates a model evaluation approach by Kagan and Jackson (1995) and Schorlemmer et al. (2007), to which we refer as 'earthquake likelihood model testing'. Together with further conceptual and computational improvements due to Zechar et al. (2010) and Rhoades et al. (2011) this method is used in the RELM initiative, where a collection of earthquake forecasts underwent several prospective testing procedures, see Schorlemmer and Gerstenberger (2007) for details.

Earthquake likelihood model testing represents each earthquake by a point in $\mathbf{S} \times \mathbf{M}$, where $\mathbf{S} \subset \mathbb{R}^k$ is some region in space and $\mathbf{M} \subset \mathbb{R}^d$ is a set of marks, representing earthquake features such as magnitude. The set $\mathbf{S} \times \mathbf{M}$ is partitioned into bins $B_1, \ldots, B_N$ for some $N \in \mathbb{N}$ and the values $x_1, \ldots, x_N \in \mathbb{N}_0$ count the numbers of earthquakes falling in each bin. A forecast or 'model' is determined by values $\lambda_1, \ldots, \lambda_N \in (0, \infty)$ and its 'log-likelihood' (Schorlemmer et al., 2007) is defined as a sum of Poisson likelihoods

$$\ell(\lambda_1, \ldots, \lambda_N, x_1, \ldots, x_N) = \sum_{i=1}^{N} (x_i \log \lambda_i - \log(x_i!) - \lambda_i) . \tag{4.10}$$

This terminology is motivated by the fact that, if Φ is a Poisson point process with intensity measure Λ such that $\Lambda(B_i) = \lambda_i$ for $i = 1, \ldots, N$, then (4.10) is the log-likelihood of the realization $x_1, \dots, x_N$. Based on this idea, Kagan and Jackson (1995) and Schorlemmer et al. (2007) propose different tests to evaluate forecasts.

To assess the absolute performance of a forecast, they introduce the *L-test*, which compares the realized value $z := \ell(\lambda_1, \dots, \lambda_N, x_1, \dots, x_N)$ to the distribution of the random variable $Z := \ell(\lambda_1, \dots, \lambda_N, X_1, \dots, X_N)$, where $X_1, \dots, X_N$ are independent Poisson random variables with parameters $\lambda_1, \dots, \lambda_N$. The model is rejected if the realization z lies in the tail of the distribution of Z. To determine the latter, we can either rely on simulations or approximate the CDF of Z, as proposed in Rhoades et al. (2011).

The *R-test*, or ratio test, compares two forecasts A and B specified by their bin intensities λ_i^A and λ_i^B for $i = 1, \dots, N$, and aims to check whether model A is at least as good as model B. Naturally, the analogous formulation with reversed roles is possible, as well. The R-test considers the 'log-likelihood ratio' based on (4.10), i.e.

$$R(A, B, x_1, \dots, x_N) = \ell(\lambda_1^A, \dots, \lambda_N^A, x_1, \dots, x_N) - \ell(\lambda_1^B, \dots, \lambda_N^B, x_1, \dots, x_N), \quad (4.11)$$

and then proceeds analogously to the L-test, i.e. it compares the realized value $z := R(A, B, x_1, \dots, x_N)$ to the distribution of $Z := R(A, B, X_1, \dots, X_N)$, where $X_1, \dots, X_N$ are independent Poisson random variables with parameters λ_i^A for $i = 1, \dots, N$. If z lies in the lower tail of the distribution of Z, then model A is deemed worse than model B.

As the distributional assumptions on $X_1, \dots, X_N$ demonstrate, there is an asymmetry inherent in the R-test: If model A is tested against model B, then the X_i are assumed to have parameters λ_i^A and if B is tested against A, then λ_i^B are the parameters of the X_i. As noted by Rhoades et al. (2011) this implies that the R-test is not really a comparative test, but rather a goodness-of-fit test such as the L-test. This explains seemingly contradictory results observed in practice, where R-tests deem A worse than B and vice versa, see Rhoades et al. (2011) and Bray and Schoenberg (2013) for details and references.

Motivated by this asymmetry, Rhoades et al. (2011) propose two modifications of the R-test which avoid such contradictory results. Instead of assuming a distribution for (4.11), they find a different representation of R and assume a normal distribution for this test statistic. Large positive realizations of R then support model A, while large negative values support model B. If too few data are available they propose to test whether R has zero median by employing the Wilcoxon signed-rank test. Both testing ideas can be interpreted as variants of Diebold-Mariano (DM) tests (see Diebold and Mariano (1995) and Section 1.5) and the connection to scoring functions is detailed in Subsection 4.3.3. Note that Rhoades et al. (2011) use the term 'information gain' to refer to their test statistic, see Subsection 4.3.1.

As pointed out by Harte (2015), earthquake likelihood model testing suffers from several drawbacks. Firstly, relying on binning leads to a loss of information, since the behavior of models inside bins will not affect the evaluation. Moreover, assuming independence among bins as well as a Poisson distribution leads to a likelihood misspecification when assessing general point process models. This prohibits the testing of model characteristics other than bin expectations, since by reporting $(\lambda_i)_{i=1,\dots,n}$, every forecast is converted to a Poisson point process. The Collaboratory for the Study of Earthquake Predictabiliy (CSEP), which succeeds RELM, will address these problems by considering more complex forecasts, which include distributional features or correlations, see Schorlemmer et al. (2018) for details. However, as mentioned by Bray and Schoenberg (2013), it is unclear how big the impact of the Poisson assumption is on the testing results. By viewing the testing methods from the perspective of consistent scoring functions, the next subsection gives new insights into this question.

4.3.3 Connections to scoring theory

We now explain how information gains and earthquake likelihood model testing connect to scoring functions and the results from Section 4.2. In a nutshell, both approaches can be interpreted as special choices of consistent scoring functions which compare forecasts and realizations for each set in a partition $\mathcal{T}$ of the domain $\mathcal{X}$.

Information gain We start with Scenario B and consider a temporal point process on an interval $[0,T]$ and a partition $\mathcal{T}_n := \{(a_1,b_1],\ldots,(a_{k_n},b_{k_n}]\}$ of $(0,T]$ into k_n subintervals. Motivated by the information gain approach (Subsection 4.3.1) we define the *interval scoring function* $S_{\text{int}}^{\mathcal{T}_n} : [0,1]^{k_n} \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ via

$$S_{\text{int}}^{\mathcal{T}_n}(p_1,\ldots,p_{k_n},\varphi) = \sum_{i=1}^{k_n} \left[-\mathbb{1}\big(\varphi\big((a_i,b_i]\big) > 0\big)\log(p_i) - \mathbb{1}\big(\varphi\big((a_i,b_i]\big) = 0\big)\log(1-p_i)\right] \tag{4.12}$$

for each partition $\mathcal{T}_n$, $n \in \mathbb{N}$. The summands of (4.9) are equal to $S(q_i,X_i) - S(p_i,X_i)$, where $X_i = \mathbb{1}(\Phi((a_i,b_i]) > 0)$ and $S : [0,1] \times \{0,1\} \to \bar{\mathbb{R}}$ defined via

$$S(p,y) = -y\log(p) - (1-y)\log(1-p)$$

is the binary logarithmic score (see Subsection 1.4.1). Since S is a strictly proper scoring rule (Gneiting and Raftery, 2007), we obtain that $S_{\text{int}}^{\mathcal{T}_n}$ is strictly consistent for the collection of probabilities $\mathbb{P}(X_i = 1) = \mathbb{P}(\Phi((a_i,b_i]) > 0)$ with $(a_i,b_i] \in \mathcal{T}_n$. Hence, we can loosely speak of *negative* information gains as being strictly consistent for the collection of probabilities that interval i contains at least one point of Φ. This holds for unconditional as well as conditional probabilities alike. In case of conditional probabilities, forecasters need to report instructions on how to calculate the probabilities from past observations. Simulation experiments in Subsection 4.4.3 illustrate how this approach compares different conditional intensity forecasts.

If conditional probabilities can be computed from a conditional intensity model λ^*, then $S_{\text{int}}^{\mathcal{T}_n}$ connects naturally to the scoring functions derived in Subsection 4.2.5. To make this precise, we follow Daley and Vere-Jones (2003, Definition A1.6.I) and call a sequence of partitions $(\mathcal{T}_n)_{n\in\mathbb{N}}$ *dissecting* if it is nesting and asymptotically separates every pair of points. Then the following approximation result holds and Subsection 4.4.3 studies the quality of this approximation via simulations.

Proposition 4.17. *Let λ^* be a conditional intensity and $(\mathcal{T}_n)_{n\in\mathbb{N}}$ a dissecting system of partitions of $(0,T]$, consisting of intervals. Let $P_0 \in \mathcal{P}$ be the distribution of the unit rate Poisson point process on $[0,T]$ and define conditional probability reports*

$$p_i^{(n)} = 1 - \exp\left(-\int_{a_i^{(n)}}^{b_i^{(n)}} \lambda^*\big(t \mid t_j < a_i^{(n)}\big)\,\mathrm{d}t\right)$$

for all $i = 1, \dots, k_n$, $(a_i^{(n)}, b_i^{(n)}] \in \mathcal{T}_n$, *and* $n \in \mathbb{N}$. *Then*

$$S_{\text{int}}^{\mathcal{T}_n}(p_1^{(n)}, \dots, p_N^{(n)}, \varphi) + \sum_{i=1}^{k_n} \mathbb{1}\big(\varphi\big((a_i^{(n)}, b_i^{(n)}]\big) > 0\big) \log\big(b_i^{(n)} - a_i^{(n)}\big) \longrightarrow S(\lambda^*, \varphi)$$

for P_0*-a.e.* $\varphi \in \mathbb{M}_0([0,T])$ *as* $n \to \infty$, *where* S *is the scoring function from Example 4.15.*

Proof. Denote a point process realization via $\varphi = \{t_1, \dots, t_m\}$ for $m \in \mathbb{N}_0$ and let n be large enough so that each interval contains at most one point. Let $I_0^{(n)} \subset \{1, \dots, k_n\}$ denote the indices of intervals which do not contain a point and $I_1^{(n)}$ the indices of intervals which contain a point. The score $S_{\text{int}}^{\mathcal{T}_n}(p_1^{(n)}, \dots, p_N^{(n)}, \varphi)$ can now be divided into two sums with respect to the indices $I_0^{(n)}$ and $I_1^{(n)}$. For the first sum we obtain

$$\sum_{i \in I_0^{(n)}} -\mathbb{1}\big(\varphi\big((a_i^{(n)}, b_i^{(n)}]\big) = 0\big) \log(1 - p_i^{(n)}) = \sum_{i \in I_0^{(n)}} \int_{a_i^{(n)}}^{b_i^{(n)}} \lambda^*\big(t \mid t_j < a_i^{(n)}\big)\, \mathrm{d}t.$$

For the second sum we add the correction term and use the fact that $|\log(1-\exp(-x)) - \log(x)| \to 0$ for $x \to 0$. This yields

$$\begin{aligned}
&\sum_{i \in I_1^{(n)}} -\mathbb{1}\big(\varphi\big((a_i^{(n)}, b_i^{(n)}]\big) > 0\big) \log(p_i^{(n)}) + \sum_{i=1}^{k_n} \mathbb{1}\big(\varphi\big((a_i^{(n)}, b_i^{(n)}]\big) > 0\big) \log\big(b_i^{(n)} - a_i^{(n)}\big) \\
&\quad = \sum_{i \in I_1^{(n)}} -\log\left(\big(b_i^{(n)} - a_i^{(n)}\big)^{-1} \int_{a_i^{(n)}}^{b_i^{(n)}} \lambda^*(t \mid t_j < a_i^{(n)})\, \mathrm{d}t\right) + o(1) \\
&\quad \longrightarrow -\sum_{j=1}^{m} \log(\lambda^*(t_j \mid t_1, \dots, t_{j-1}))
\end{aligned}$$

for $n \to \infty$ and P_0-a.e. $\varphi \in \mathbb{M}_0([0,T])$. The convergence follows from suitable approximation results for λ^*, see e.g. Daley and Vere-Jones (2003, Lemma A1.6.III). Combined with the first equation this gives the result. □

Earthquake likelihood model testing In earthquake likelihood model testing (Subsection 4.3.2) forecasts consist of positive values $\lambda_1, \dots, \lambda_n$, and the corresponding Poisson distributions are compared via the logarithmic score. To formalize this, consider a bounded spatial domain $\mathcal{X}$ which is partitioned into k_n bins $\mathcal{T}_n = \{B_1, \dots, B_{k_n}\}$. Based on (4.10) and (4.11) we define the *bin scoring function* $S_{\text{bin}}^{\mathcal{T}_n} : (0, \infty)^{k_n} \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ via

$$S_{\text{bin}}^{\mathcal{T}_n}(\lambda_1, \dots, \lambda_{k_n}, \varphi) = \sum_{i=1}^{k_n} -\varphi(B_i) \log(\lambda_i) + \lambda_i \tag{4.13}$$

for each partition $\mathcal{T}_n$, $n \in \mathbb{N}$. The following result establishes that $S_{\text{bin}}^{\mathcal{T}_n}$ is strictly consistent for the collection of bin expectations $\mathbb{E}\Phi(B_i)$, $B_i \in \mathcal{T}_n$, see also Example 4.2.

Proposition 4.18. *If $\mathcal{N}$ is a set of probability measures on $\mathbb{N}_0$ with finite first moments, then the scoring function $S : [0,\infty) \times \mathbb{N}_0 \to \bar{\mathbb{R}}$ defined via*

$$S(\lambda, y) = -y \log(\lambda) + \lambda \tag{4.14}$$

is strictly consistent for the expectation.

Proof. For all $\lambda_1, \lambda_2 > 0$ we have

$$S(\lambda_1, y) - S(\lambda_2, y) = \lambda_1 - \log(\lambda_1)y - \lambda_2 + \log(\lambda_2)y = b(\lambda_1, y) - b(\lambda_2, y),$$

where $b(x, y) = x - \log(x)y$ is the Bregman function corresponding to the strictly convex function $f(\lambda) = \lambda(\log(\lambda) - 1)$. Hence, strict consistency for the expectation follows from Theorem 1.6. □

The scoring function (4.14) can be interpreted as a discrete analogon to the David-Sebastiani (DS) score (see Subsection 1.4.1 and Dawid and Sebastiani (1999)), but with the normal distribution replaced by the Poisson distribution, see also Example 1.28. The term $\log(y!)$ which appears in (4.10) can be omitted, since it does not depend on the report λ.

As noted by Harte (2015) the implications of using Poisson likelihoods for earthquake likelihood model testing are not completely clear. Via Proposition 4.18 we now see that the Poisson assumption leads to a sound comparison of bin expectations, since the true expectations obtain minimal expected score. This conclusion holds even if the data do not follow a Poisson point process. Hence, the Poisson assumption does not imply that the corresponding tests are only valid for Poisson point process data. It rather means that the tests are sensitive to the bin expectations only, since they rely on strictly consistent scoring functions for the expectation. As a consequence, the symmetric modifications of the R-test due to Rhoades et al. (2011), which assume a normal distribution for the log-likelihood ratio (4.11), can be seen as DM tests in the spirit of Section 1.5, based on the scoring function $S_{\text{bin}}^{\mathcal{T}_n}$.

Just as the Poisson distribution gives rise to the scoring function (4.14) for the expectations, the Poisson point process can be used to obtain a scoring function for the intensity. The reason is that every intensity report induces a Poisson point process with this intensity and these processes can then be compared via the logarithmic score (4.3), which attains the value (4.4) for Poisson densities. Using the notation from Subsection 4.2.3 we obtain strict consistency, analogous to Proposition 4.18.

Proposition 4.19. *Let all elements of $\mathcal{M}_f$ admit densities with respect to Lebesgue measure. Then the scoring function $S : \mathcal{M}_f \times \mathbb{M}_0 \to \bar{\mathbb{R}}$ defined via*

$$S(\Lambda, \{y_1, \ldots, y_n\}) = -\sum_{i=1}^{n} \log \lambda(y_i) + \int_{\mathcal{X}} \lambda(y)\, \mathrm{d}y \tag{4.15}$$

for $n \in \mathbb{N}$ and $S(\Lambda, \emptyset) = \int \lambda(y)\, \mathrm{d}y$ is a strictly consistent scoring function for the intensity.

Proof. The scoring function (4.15) corresponds to an S from Corollary 4.9 when choosing the logarithmic score for S', the Bregman function b as in the proof of Proposition 4.18, and $c = 1$. Since S' is strictly proper and b is strictly consistent, S is strictly consistent for the intensity. □

Like (4.14) the scoring function (4.15) can be interpreted as a point process analogon to the DS score. While the DS score relies on first and second moments, this score depends on the intensity only.

Proposition 4.19 shows that a straightforward generalization of earthquake likelihood model testing, which compares general processes via their Poisson point process counterparts, leads to a consistent comparison of intensities. In particular, we can conclude that binning is not necessary for this approach. Intensities can be modeled without relying on bins and they can be compared via scoring functions for intensities (see Subsection 4.2.3) with Proposition 4.19 giving one possible choice.

However, in some situations binning might be desirable, e.g. when no explicit expression for λ is available. The next result shows that under weak conditions $S_{\text{bin}}^{\mathcal{T}_n}$ can be used as an approximation to the scoring function (4.15) in this situation.

Proposition 4.20. *Let $\lambda : \mathcal{X} \to [0, \infty)$ be an intensity and $(\mathcal{T}_n)_{n\in\mathbb{N}}$ a dissecting system of partitions of $\mathcal{X}$, consisting of rectangles. Let $P_0 \in \mathcal{P}$ be the distribution of the unit rate Poisson point process on $\mathcal{X}$ and define reports*

$$\lambda_i^{(n)} = \int_{B_i^{(n)}} \lambda(y)\,\mathrm{d}y,$$

for all $i = 1, \dots, k_n$, $B_i^{(n)} \in \mathcal{T}_n$, and $n \in \mathbb{N}$. Then

$$S_{\text{bin}}^{\mathcal{T}_n}\big(\lambda_1^{(n)}, \dots, \lambda_{k_n}^{(n)}, \varphi\big) + \sum_{i=1}^{k_n} \mathbb{1}(\varphi(B_i^{(n)}) > 0) \log(|B_i^{(n)}|) \longrightarrow S(\Lambda, \varphi)$$

for P_0-a.e. $\varphi \in \mathbb{M}_0(\mathcal{X})$ as $n \to \infty$, where S is the scoring function (4.15).

Proof. Let $\varphi = \{y_1, \dots, y_m\}$ with $m \in \mathbb{N}_0$ be a point process realization. For a large enough $n \in \mathbb{N}$ every bin $B_i^{(n)}$ contains at most one point of φ so let $i_n(j)$ denote the index of the bin such that $y_j \in B_{i_n(j)}^{(n)}$ for all $j = 1, \dots, m$. This yields

$$\begin{aligned}
&S_{\text{bin}}^{\mathcal{T}_n}(\lambda_1^{(n)}, \dots, \lambda_{k_n}^{(n)}, \varphi) + \sum_{i=1}^{k_n} \mathbb{1}(\varphi(B_i^{(n)}) > 0) \log(|B_i^{(n)}|) \\
&\quad = -\sum_{i=1}^{k_n} \left[\varphi(B_i^{(n)}) \log\Big(\int_{B_i^{(n)}} \lambda(y)\,\mathrm{d}y \Big) - \mathbb{1}(\varphi(B_i^{(n)}) > 0) \log(|B_i^{(n)}|) - \int_{B_i^{(n)}} \lambda(y)\,\mathrm{d}y \right] \\
&\quad = -\sum_{j=1}^{m} \log\left(|B_{i_n(j)}^{(n)}|^{-1} \int_{B_{i_n(j)}^{(n)}} \lambda(y)\,\mathrm{d}y \right) + \int_{\mathcal{X}} \lambda(y)\,\mathrm{d}y \\
&\quad \longrightarrow -\sum_{j=1}^{m} \log(\lambda(y_j)) + \int_{\mathcal{X}} \lambda(y)\,\mathrm{d}y
\end{aligned}$$

for $n \to \infty$ and P_0-a.e. $\varphi \in \mathbb{M}_0(\mathcal{X})$. The last line follows from a result for the approximation of Radon–Nikodým derivatives applied to λ, see Daley and Vere-Jones (2003, Lemma A1.6.III). □

Table 4.1: Fraction of times the 'row forecast' was preferred over the 'column forecast' by a standard DM test with level $\alpha = 0.05$, where Φ is a Poisson point process (left) or a Gaussian determinantal point process (right)

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.46	0.84	0.66	1	1
f_1	0		0.4	0.64	0.99	1
f_2	0	0		0.29	0.96	1
f_3	0	0	0		0.49	1
f_4	0	0	0	0		1
f_5	0	0	0	0	0	

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.57	0.89	0.76	1	1
f_1	0		0.45	0.73	0.99	1
f_2	0	0.01		0.38	0.97	1
f_3	0	0	0		0.55	1
f_4	0	0	0	0		1
f_5	0	0	0	0	0	

This result can be proved for more general partitions, as long as the family of sets $(\mathcal{T}_n)_{n\in\mathbb{N}}$ generates the Borel σ-algebra on $\mathcal{X}$. However, in most cases of interest the partitions arise from binning each coordinate into intervals, giving rectangles. Subsection 4.4.1 studies the speed of convergence of this binning approach via simulation experiments.

4.4 Simulation examples

This section investigates finite sample properties of scoring function-based model evaluation, by illustrating the behavior of average score differences and Diebold-Mariano (DM) tests for different models and scenarios. We begin with spatial point processes and consider the intensity and product density (Subsection 4.2.3). We compare different forecasts for both characteristics based on $n \in \mathbb{N}$ realizations of the point process, where n could reflect the number of locations of measurement (Scenario A), or different points in time (Scenario C1), e.g. $n = 52$ for one year of weekly data. We then turn to temporal processes (Scenario B) and compare forecasts of the triggering properties of linear Hawkes processes (Subsection 4.2.5) based on one realization of the process in an interval $[0, T]$.

All simulations are performed with the free software R (R Core Team, 2020). We use the `spatstat` package (Baddeley and Turner, 2005; Baddeley et al., 2015) for the spatial point processes and Ogata's thinning method (Ogata, 1981) for the temporal point processes.

4.4.1 Intensity

This subsection compares different intensity reports based on average scores for a point process Φ on the window $[0, 1]^2$, which corresponds to Scenario A. We draw $n = 20$ i.i.d. samples φ_i from Φ and compare the average score $\bar{s}_j := \frac{1}{n}\sum_{i=1}^{n} S(f_j, \varphi_i)$ for different forecast intensities f_j. We consider four different data-generating processes for Φ all of which have (approximate) intensity $\lambda(x, y) = 30\sqrt{x^2 + y^2}$. The simulations are repeated $m = 500$ times to assess the variation in average scores.

Six different intensity forecasts are compared below, namely the perfect forecast

$f_0 = \lambda$ and

$$\begin{aligned} f_1(x,y) &= 40\sqrt{(x-0.2)^2+(y-0.1)^2} \\ f_2(x,y) &= 11.78(x+3y) \\ f_3(x,y) &= 45\sqrt{(x-0.2)^2+(y-0.1)^2} \\ f_4(x,y) &= 9.5\left(\frac{1}{\sqrt{1.2-x}}+2(1-y)\right) \\ f_5(x,y) &= 46\exp\left(-2(x^2+(y-1/2)^2)\right) \end{aligned}$$

The motivation for this choice is as follows. Intensity f_1 is intuitively the best forecast since it has roughly the correct shape up to a small shift and f_3 is a version of f_1 with a too high scaling factor. Intensity f_2 is similar to f_0 but linear while f_4 and f_5 have completely different shape. Except for f_3, all intensities put roughly identical mass on $[0,1]^2$. This allows for an assessment of how the scoring function reacts to misspecifications in scale instead of shape.

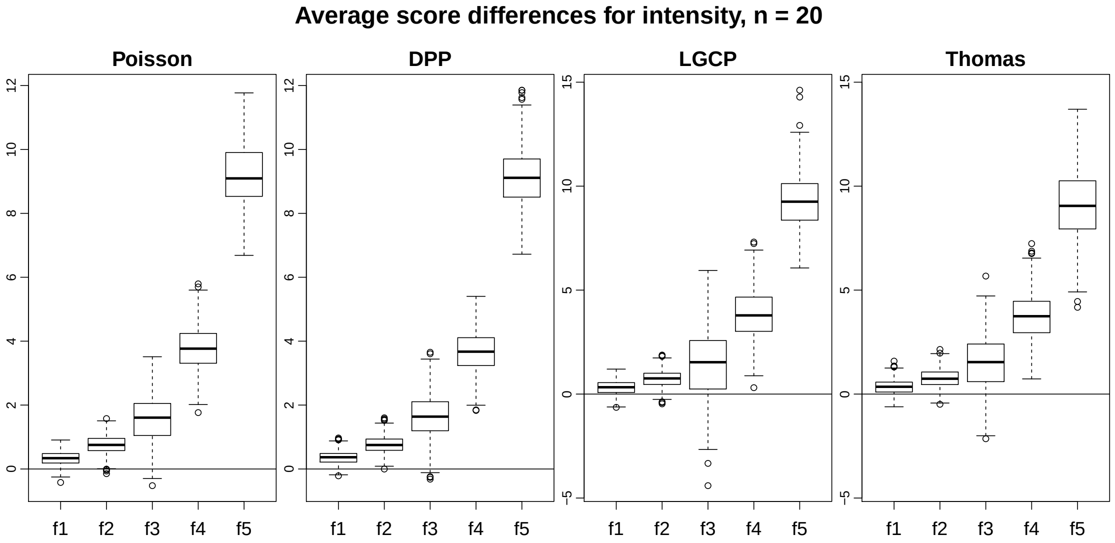

Figure 4.1: Boxplot of difference in average scores $\bar{s}_j - \bar{s}_0$ for $j = 1,\dots,5$ and scoring function S_1 from Example 4.10. From left to right, Φ is a Poisson point process, a Gaussian determinantal point process, a log-Gaussian Cox process, or an inhomogeneous Thomas process

Forecast comparison We begin with four simulation experiments based on the scoring function given in Example 4.10, which we denote via S_1 in the following. The scaling factor $c > 0$ is chosen such that the log and squared terms of S_1 are of the same order of magnitude in these simulations, giving $c = 0.1$. See Subsection 4.2.3 for a discussion of the choice of c.

In our first two experiments, Φ is a Poisson point process or a Gaussian determinantal point process, both having intensity λ. The latter is a determinantal point process (DPP) with Gaussian covariance, such that its points exhibit moderate inhibition. More details

Table 4.2: Fraction of times the 'row forecast' was preferred over the 'column forecast' by a standard DM test with level $\alpha = 0.05$, where Φ is a log-Gaussian Cox process (left) or an inhomogeneous Thomas process (right)

	f_0	f_1	f_2	f_3	f_4	f_5		f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.3	0.62	0.28	0.98	1	f_0		0.25	0.49	0.36	0.95	1
f_1	0		0.33	0.26	0.82	1	f_1	0.01		0.24	0.34	0.79	1
f_2	0	0.01		0.17	0.72	1	f_2	0	0		0.2	0.68	1
f_3	0	0	0.01		0.19	0.95	f_3	0.01	0.01	0.02		0.26	0.95
f_4	0	0	0	0.01		0.99	f_4	0	0	0	0.01		0.97
f_5	0	0	0	0	0		f_5	0	0	0	0	0	

are given in Subsection 4.4.2 and Lavancier et al. (2015). The left part of Figure 4.1 shows the average score differences between the five different forecasts $f_1, \ldots, f_5$ and the perfect forecast f_0 for both experiments. The plots share the same general properties, namely f_1 is close to the optimal forecast, f_2 and f_3 are worse, and the average score differences of the misspecified forecasts f_4 and f_5 are far from zero. Table 4.1 holds the results of DM tests (see Section 1.5) for both experiments. The probabilities of preferring f_0 against f_j, $j = 1, \ldots, 5$ (first row) are overall in line with the average score differences in Figure 4.1. The only difference is that f_3 is less often deemed inferior to f_0 than f_2, although it is clearly inferior to f_2 in terms of expected scores. The reason for this is the higher variance of the average score differences for f_3 compared to f_2 (see Figure 4.1), leading to less conclusive DM test results.

In the third and fourth simulation experiment Φ is a log-Gaussian Cox process (LGCP) or an inhomogeneous Thomas process. For the simulation of the LGCP, which has the intensity function

$$\lambda_{\mathrm{LGCP}}(s) = \exp\left(\mu(s) + \frac{1}{2}C(s,s)\right) \tag{4.16}$$

for $s \in \mathbb{R}^2$, we chose the exponential covariance $C(s,t) = 1/4 \exp(-\|s-t\|^2)$ and μ such that $\lambda_{\mathrm{LGCP}} = \lambda$ holds. The Thomas process is a cluster process which arises from an inhomogeneous Poisson process as parent and a random number of cluster points which are drawn from a normal distribution centered at its parent point. As intensity of the parent process we choose $\lambda/2$ and the number of points per cluster follow a Poisson distribution with parameter 2. The location of each cluster point is determined by a normal distribution which is centered at the parent point and where the components are uncorrelated and have standard deviation 0.05. As a result of the clustering the intensity of the Thomas process is only approximately equal to λ.

The results of the third and fourth experiment are given in the right part of Figure 4.1. The overall behavior of average score differences is the same as in the previous two experiments (same figure), but the variance increases. As shown in Table 4.2, the results of DM tests are similar to the previous two experiments, as well, although the probabilities of preferring f_0 (first row) decrease overall. An intuitive reason for this is that clustering, which is a feature of the LGCP and the Thomas process, complicates the distinction between different intensity forecasts. In contrast, the inhibition of the Gaussian DPP seems to facilitate the comparison, see Table 4.1. An increase in sample size to $n = 50$ can compensate for clustering and leads to more definitive preferring probabilities for the LGCP and Thomas process (not shown).

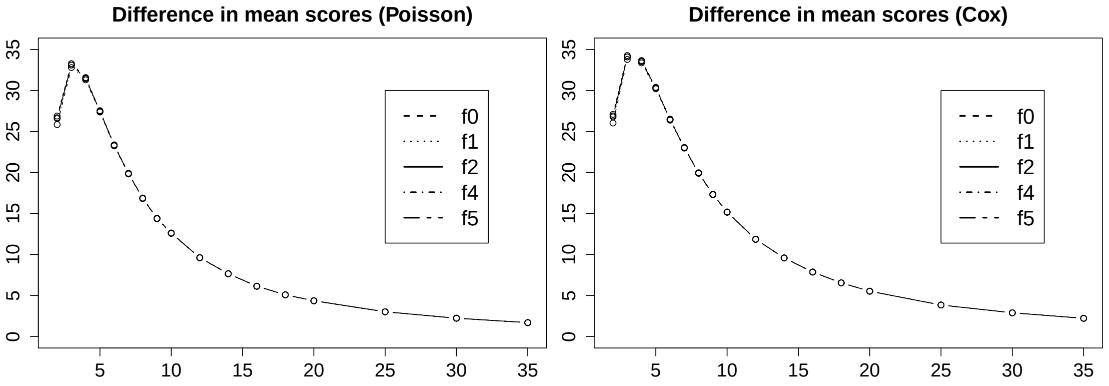

Figure 4.2: Average differences in realized scores $S_{\text{bin}}^{\mathcal{T}_n} - S_2 + C_n$ for different values of n, where C_n is a correction term as given in Proposition 4.20. The process Φ is a Poisson point process (left) or a log-Gaussian Cox process (right). Note that f_3 obtains the same values as f_1 since the difference $S_{\text{bin}}^{\mathcal{T}_n} - S_2$ is independent of the scaling of the reports

Relation to binning and earthquake likelihood model testing We now investigate how the forecast comparison changes when using scoring functions motivated by earthquake likelihood model testing (Subsection 4.3.3) instead of S_1 from Example 4.10. We start with the scoring function (4.15), denote it via S_2, and repeat the simulation experiments for the four different choices of Φ as above. These experiments lead to overall similar boxplots of average score differences as in Figure 4.1 up to an increase in variance and they are thus omitted. However, inspecting the results of DM tests given in Table 4.3 reveals some noteworthy differences. If we compare the values for f_2 and f_3 in Table 4.3 to the left-hand side of Table 4.1 and 4.2, we see that f_2 is more often preferred to f_3 under S_1 than vice versa. For S_2 the roles are switched and f_3 is now preferred to f_2 more often. A possible reason for this is that f_3 is a wrongly scaled version of the 'almost perfect' forecast f_1 and S_2 is less sensitive to scaling than S_1. It puts more emphasis on the shape of the intensity at the cost of less sensitivity to the number of points. As in the previous experiments, the clustering of the LGCP leads to less conclusive decisions between the forecasts.

A further sequence of experiments considers the speed of convergence in Proposition 4.20, i.e. how well the binned scoring function $S_{\text{bin}}^{\mathcal{T}_n}$ given in (4.13), together with a forecast-independent correction term, approximates S_2. We select a dissecting system of partitions $(\mathcal{T}_n)_{n\in\mathbb{N}}$ of $[0,1]^2$ which arises from partitioning both axes, i.e. each bin $B_{ij}^{(n)} \in \mathcal{T}_n$ is given by $[(i-1)/n, i/n] \times [(j-1)/n, j/n]$ for $i, j \in \{1, \dots, n\}$. The number of bins is thus $k_n = n^2$ and we choose $n \in \{1, 2, \dots, 35\}$ for the simulations. As forecasts we rely on the functions $f_0, \dots, f_5$ introduced above which we transform into bin reports $f_{l,ij}^{(n)}$ by computing the integral of f_l over the bin $B_{ij}^{(n)}$ for all bins. These reports are then compared to the number of points per bin via $S_{\text{bin}}^{\mathcal{T}_n}$. Figure 4.2 illustrates convergence of the mean difference to zero as n grows. The process Φ is a Poisson point process or

Table 4.3: Fraction of times the ‘row forecast’ was preferred over the ‘column forecast’ by a standard DM test with level $\alpha = 0.05$ based on the scoring function S_2 (see (4.15)), where Φ is a Poisson point process (left) or a log-Gaussian Cox process (right)

	f_0	f_1	f_2	f_3	f_4	f_5		f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.5	0.83	0.75	1	1	f_0		0.44	0.7	0.47	1	1
f_1	0		0.42	0.61	0.99	1	f_1	0		0.38	0.26	0.98	1
f_2	0	0		0.01	0.96	1	f_2	0	0		0.04	0.91	1
f_3	0	0	0.12		0.99	1	f_3	0	0.01	0.12		0.89	1
f_4	0	0	0	0		1	f_4	0	0	0	0		1
f_5	0	0	0	0	0		f_5	0	0	0	0	0	

a LGCP as specified in the previous experiments and $m = 500$ repetitions are used. It suggests that the speed of convergence does not depend on the reported intensity. The results do not change significantly when using the Gaussian DPP or the Thomas process, thus the corresponding plots are omitted.

An alternative way to study convergence consists of plotting the results of DM tests based on $S_{\text{bin}}^{\mathcal{T}_n}$ for increasing n. The corresponding fractions exhibit fast convergence to the values in Table 4.3, however, since the related plots contain 30 different curves, we omit them here.

4.4.2 Product density

In this subsection we focus on Scenario A again, however, we now consider second order properties and keep intensities fixed. We simulate a stationary and isotropic point processes Φ on the window $[0,1]^2$ with three different second order structures corresponding to inhibition, clustering, and no interaction. We draw $n = 20$ i.i.d. samples φ_i from Φ and compare the average scores for different forecasts, in the same way as in the previous subsection. The scoring function S is defined in Example 4.12 and the scaling factor $c > 0$ is chosen such that the log and squared terms are of the same order of magnitude, in this case $c = 10^{-5}$. We repeat the simulations $m = 500$ times to assess the variation in average scores.

Five different product density forecasts are compared below, given by

$$\begin{aligned}
f_1(r) &= \exp\left(2\mu + \sigma^2\left(1 + \exp(-400r^2)\right)\right) \\
f_2(r) &= \exp\left(2\mu + \sigma^2\left(1 + \exp(-20r)\right)\right) \\
f_3(r) &= \lambda^2 \\
f_4(r) &= \lambda^2\left(1 - \exp(-2r/\gamma)\right) \\
f_5(r) &= \lambda^2\left(1 - \exp(-2(r/\gamma)^2)\right),
\end{aligned}$$

where we choose $\mu = \log(\lambda) - \sigma^2/2$, $\sigma^2 = \log(2)$, $\gamma = 0.06$ and $\lambda = 40$. Figure 4.3 gives a graphical comparison of the different functions. The first two forecasts represent clustering, since they arise as product densities of log-Gaussian Cox processes (LGCPs). A stationary and isotropic LGCP is determined by a stationary and isotropic Gaussian process with mean μ and covariance $C(x,y) = C_0(\|x-y\|)$ for some $C_0 : [0,\infty) \to \mathbb{R}$.

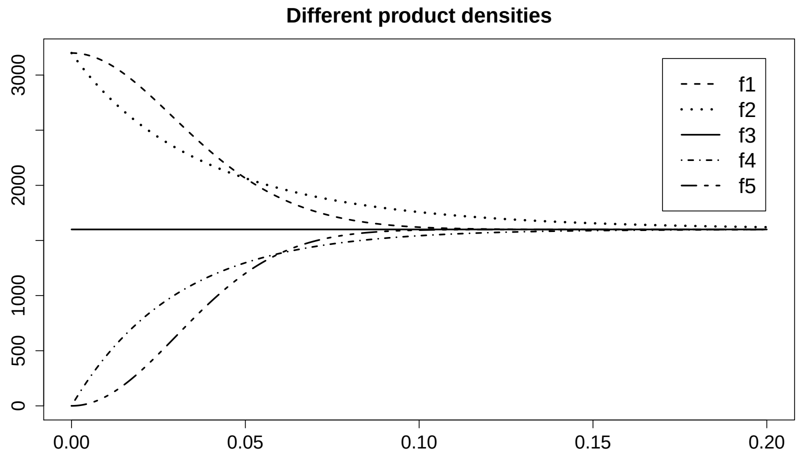

Figure 4.3: Plot of the five different product densities used as forecasts in Subsection 4.4.2

Its second order product density $\varrho^{(2)}$ is of the form $\varrho^{(2)}(x_1, x_2) = \varrho_0^{(2)}(\|x_1 - x_2\|)$ with

$$\varrho_0^{(2)}(r) = \exp(2\mu + C_0(0) + C_0(r)),$$

see e.g. Illian et al. (2008). The forecasts f_1 and f_2 are the product densities of a LGCP with Gaussian and exponential covariance function, respectively. The variance and scale are chosen as $\sigma^2 = \log(2)$ and $s = 0.05$ in both cases.

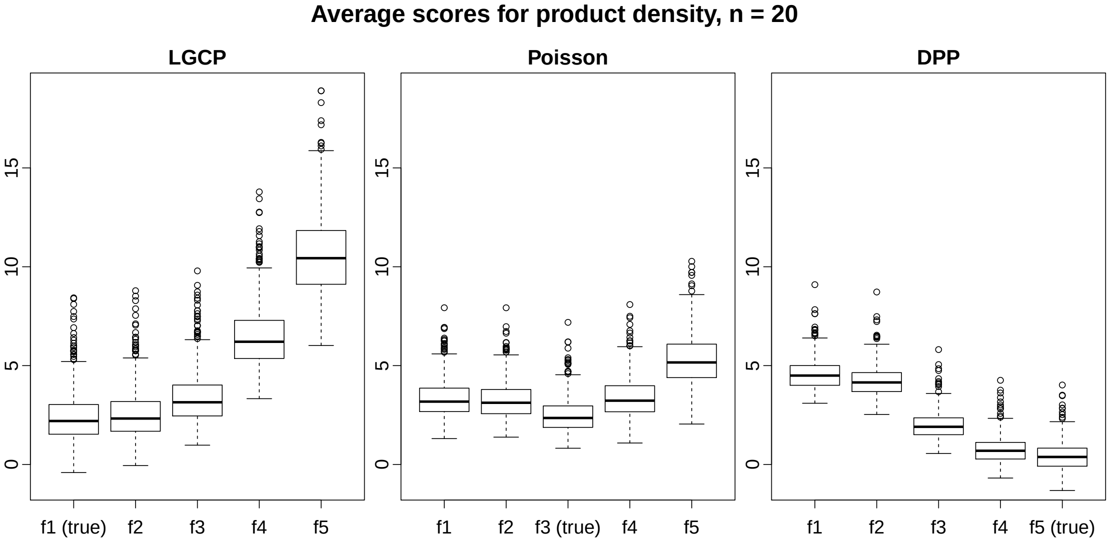

Figure 4.4: Boxplots of average scores $\bar{s}_j$ for different product density forecasts, where Φ is a log-Gaussian Cox process (left), a homogeneous Poisson process (center), or a Gaussian determinantal point process (right)

The forecast f_3 corresponds to a homogeneous Poisson process. The remaining two forecasts arise as product densities of determinantal point processes (DPPs) and thus represent inhibition. In general, a DPP is a locally finite point process with product

Table 4.4: Fraction of times the ‘row forecast’ was preferred over the ‘column forecast’ by a standard DM test with level $\alpha = 0.05$. Φ is a log-Gaussian Cox process (left), a homogeneous Poisson process (center), or a Gaussian determinantal point process (right)

	f_1	f_2	f_3	f_4	f_5
f_1		0.23	0.73	1	1
f_2	0.01		0.64	1	1
f_3	0	0		1	1
f_4	0	0	0		1
f_5	0	0	0	0	

	f_1	f_2	f_3	f_4	f_5
f_1		0	0	0	0
f_2	0.97		0	0	0
f_3	1	1		0	0
f_4	1	1	1		0
f_5	1	1	1	0.7	

	f_1	f_2	f_3	f_4	f_5
f_1		0.01	0	0.05	0.66
f_2	0.21		0	0.06	0.73
f_3	0.91	0.85		0.85	1
f_4	0.07	0.05	0		1
f_5	0	0	0	0	

densities given by

$$\rho^{(n)}(x_1, \ldots, x_n) = \det\left(C(x_i, x_j)\right)_{i,j=1,\ldots,n} \tag{4.17}$$

for all $n \in \mathbb{N}$, where $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a covariance, see Hough et al. (2006) and Lavancier et al. (2015) for details. A DPP has intensity $x \mapsto C(x, x)$ and it is stationary and isotropic whenever its covariance is. In this case, we have $C(x, y) = C_0(\|x - y\|)$ for some $C_0 : [0, \infty) \to \mathbb{R}$ such that the second order product density can be specified via

$$\varrho_0^{(2)}(r) = C_0(0)^2 - C_0(r)^2.$$

The forecasts f_4 and f_5 are the product densities of a DPP with exponential and Gaussian covariance function, respectively. The variance and scale are chosen as $\lambda^2 = 40^2$ and $\gamma = 0.06$ in both cases. Our parameter choices ensure that the point process models corresponding to $f_1, \ldots, f_5$ all have intensity equal to λ.

In the first experiment the true Φ is a LGCP with a Gaussian covariance function such that it has product density f_1 and intensity λ. In the second experiment Φ is a homogeneous Poisson process with intensity λ, such that f_3 becomes the perfect forecast in this situation. Lastly, we let Φ be a DPP with Gaussian covariance function and parameters such that f_5 is true. We thus perform one experiment for each of the three phenomena clustering, no interaction, and inhibition.

The simulated average scores are displayed in Figure 4.4 for all three experiments. The optimal forecast consistently achieves the lowest average scores. In the case of clustering (left subfigure) the LGCP related forecasts f_1 and f_2 perform roughly similar, while the misspecified no interaction and inhibition forecasts f_3, f_4 and f_5 lead to considerably higher average scores. A similar, but mirrored behavior is apparent in the inhibition experiment (right subfigure): The forecast f_4, which gets the nature of point interactions right, attains low average scores, even though it is not optimal. The average scores of the Poisson forecast f_3 are always in between the ‘extremes’. The DM test probabilities of the three experiments are given in Table 4.4 and support these observations. Additionally, the DM results illustrate that the clustering forecasts f_1 and f_2 are preferred more often over the inhibition forecasts f_5 in the case of Poisson data (center table).

4.4.3 Conditional intensity (temporal)

In this subsection we turn to Scenario B and simulate a stationary Hawkes point process (see Example 4.14) on an interval $[0,T]$ to compare forecasts of the conditional intensity. We compute realized scores with the scoring function S from Example 4.15 and normalize them by the length of the interval T. We repeat the simulation $m = 500$ times to assess the variation in realized scores.

We compare five different conditional intensity reports of the form given in Example 4.14 and fix the background rate $\nu = 2$ such that the reports differ in the triggering functions only. We define five different triggering function forecasts given by

$$\begin{aligned} f_1(t) &= 2\exp(-4t) \\ f_2(t) &= 5/4\exp(-2t) \\ f_3(t) &= 2\exp(-9/2t^2) \\ f_4(t) &= 2\max\{4-6t,0\} \\ f_5(t) &= \mathbb{1}(t \in [0,0.8]) \end{aligned}$$

We select f_1 and f_3 as candidates for the truth and let f_2, f_4, and f_5 differ in shape. A graphical comparison of the functions is given in Figure 4.5.

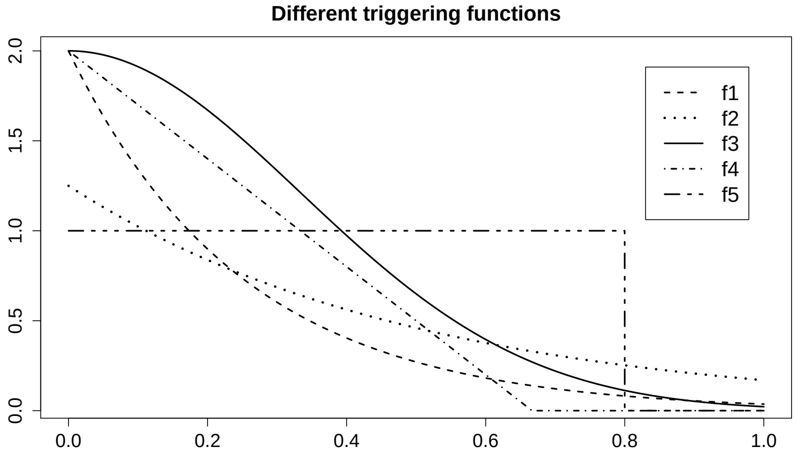

Figure 4.5: Plot of the five different triggering functions used as forecasts in Subsection 4.4.3

Forecast comparison In our first two experiments, the true process Φ is a Hawkes process with triggering functions f_1 or f_3. The left part of Figure 4.6 shows the corresponding boxplots for the score differences between the true forecast and the remaining four competitors on the interval $[0,T]$. In both experiments the score differences are overall positive, such that the true forecast can be identified. Increasing the interval

$[0, T]$ (not shown) does not change the overall appearance of the boxplots. However, the variance increases as we consider realized scores scaled by the interval length T instead of average scores.

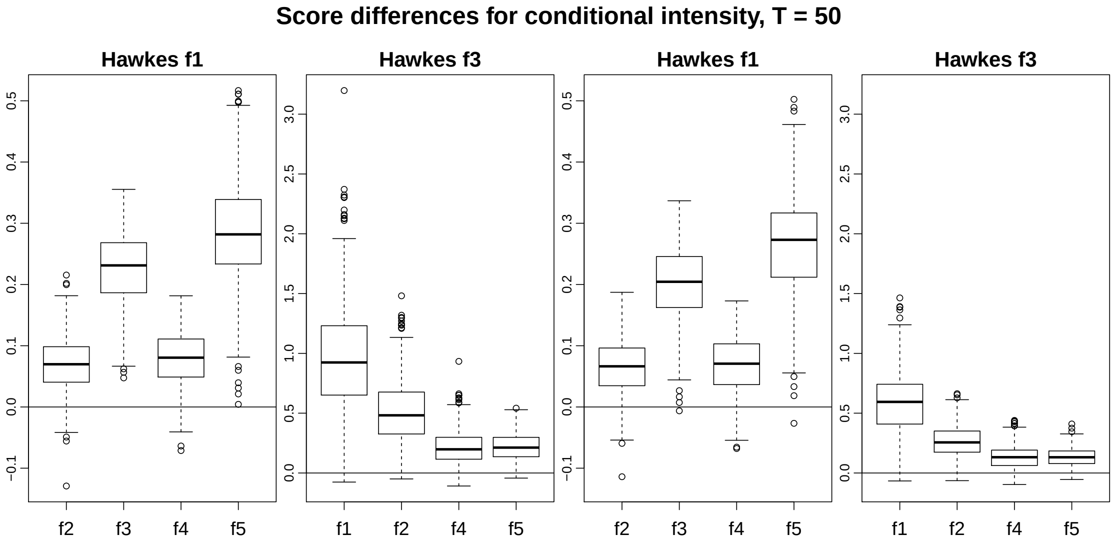

Figure 4.6: Boxplot of score differences between the true forecasts and the four remaining competitors, where Φ is a Hawkes process on the interval $[0, 50]$ with triggering function f_1 or f_3. The two plots on the left show the scoring function S from Example 4.15 while the two plots on the right show $S_{\text{int}}^{\mathcal{T}_n}$ (Equation (4.12)) with $n = 1000$

Relation to information gain approach We now use the idea of information gains discussed in Subsection 4.3.1 as an alternative method to compare forecasts for the conditional intensity. To this end, let $(\mathcal{T}_n)_{n\in\mathbb{N}}$ be a family of partitions of $(0, T]$ which consist of intervals $(a_i^{(n)}, b_i^{(n)}] = ((i-1)T/n, iT/n]$ for $i = 1, \ldots, n$. We again rely on the triggering functions $f_1, \ldots, f_5$ introduced above and transform them into collections of conditional probabilities $p_{l,i}^{(n)}$ of points materializing in each interval via the formula in Proposition 4.17. These reports are then compared to the realized data via the scoring function $S_{\text{int}}^{\mathcal{T}_n}$ defined in (4.12). The right part of Figure 4.6 displays boxplots for $m = 500$ realized scores, where the interval $[0, 50]$ is partitioned into $n = 1000$ intervals. As in the first two experiments, f_1 and f_3 are the true triggering functions. The behavior of the realized scores closely resembles the left part of Figure 4.6, suggesting that the forecast ranking of $S_{\text{int}}^{\mathcal{T}_n}$ is a good approximation to the one of S for $n = 1000$.

Proposition 4.17 ensures that the scores computed from $S_{\text{int}}^{\mathcal{T}_n}$ converge to the realized scores under S if a suitable correction term is added. Figure 4.7 illustrates the convergence of the mean of the scores for growing n based on the simulation of $m = 500$ samples and $T = 50$. It highlights that the speed of convergence depends on the underlying process. However, the absolute difference between $S_{\text{int}}^{\mathcal{T}_n}$ and S is less important than the overall forecast rankings of the scoring functions. As illustrated by Figure 4.6 these rankings are already in good agreement for $n = 1000$, corresponding to an interval length of 0.05.

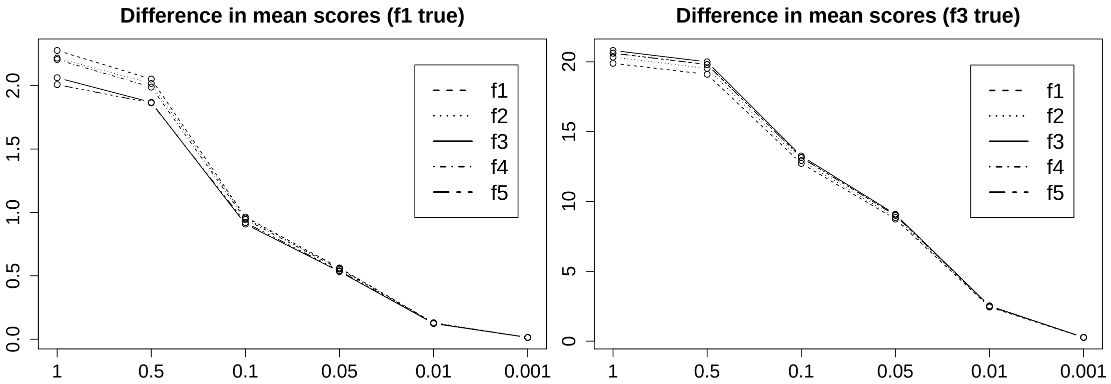

Figure 4.7: Average differences in realized scores $S_{\text{int}}^{\mathcal{T}_n} - S + C_n$ for different values of T/n, where C_n is a correction term as given in Proposition 4.17 and the scoring functions are defined in Equation (4.12) and Example 4.15, respectively. The process Φ is a Hawkes process with triggering function f_1 (left) or f_3 (right) on the interval $[0, 50]$

4.5 Discussion

Assessing accuracy via consistent scoring functions leads to principled tools for the choice of competing forecasts, which are both, theoretically underpinned and regularly used in many areas of applied statistics. As worked out in Section 4.2, consistent scoring functions transfer to the point process setting in a straightforward manner and are available for a variety of popular point process characteristics. The corresponding model evaluation methods outlined in Section 4.1 can improve forecast evaluation for point processes in applications and moreover encompass several existing techniques for model comparison (Section 4.3).

The comparison of point process functionals, as emphasized in this chapter, can be contrasted to the approach of Heinrich et al. (2019), who use an estimator in form of a function g and then compare the resulting distributions in an 'estimator space' via proper scoring rules. In practice such distributions will usually not be explicitly available and only accessible via simulations from the reported point process models. On the one hand, Heinrich et al. (2019) argue that their approach has better discrimination ability, as the whole point process distribution is taken into account. On the other hand, approximating the resulting scores via simulations leads to high computational costs, which might be prohibitive in routine evaluations. Moreover, when using characteristics such as the intensity, anybody can issue a report, without having a fully specified point process model in mind. Whichever approach is more suitable will likely depend on the problem at hand.

Absolute and relative performance Apart from the methods of Section 4.3 the majority of point process model evaluation tools focus on absolute performance and goodness-of-fit, instead of a comparison of two or more models. The most prominent example are point process residuals, which form an empirical process arising from fitting a conditional intensity to data. Classical residual methods were developed for temporal processes and extended to spatial processes, see Schoenberg (2003) and Baddeley et al.

(2005). Apart from a quantitative assessment, they can be valuable tools for graphical checks of under- or over-prediction in certain regions, see e.g. Bray et al. (2014) and Clements et al. (2011) for applications of different residual methods to earthquake forecasting models.

In a similar spirit, Thorarinsdottir (2013) transfers the probability integral transform (PIT) to point process forecasts. The PIT is a widely used tool to assess calibration, which roughly means consistency between forecasts and observations, see Dawid (1984) and Gneiting et al. (2007). Thorarinsdottir (2013) propose to choose a binning and then compare the number of points in each bin to the reported distribution of this number. Based on this, the PIT can be used to assess calibration of point process forecasts.

In contrast to such measures of absolute performance, the central use of scoring functions is the comparison of (at least) two competing models, even if both are misspecified. Although both, absolute and relative evaluation, are important in choosing suitable models, a selection among the available models has to be done eventually, and measures of absolute performance are not designed, and hence often badly positioned, for such a choice. Moreover, as pointed out by Nolde and Ziegel (2017), using absolute performance measures may lead to wrong incentives in designing candidate models. Hence, scoring functions are not competing with absolute performance measures, but are a useful and necessary addition to such methods.

Refined use of scoring functions Several results are available to tailor scoring functions to certain practical applications or address further issues in forecast evaluation. Most notably, *weighted scoring rules* provide a way to emphasize forecast performance on regions of interest and even ignore it on others. Lerch et al. (2017) illustrate that a simple weighing of the realized scores distorts the evaluation and Holzmann and Klar (2017) provide a general construction principle which ensures propriety. While previous work focuses mainly on the tails of the distribution, weighted scoring functions for point processes might focus on some spatial area or on the distribution of certain marks.

In case of noisy observational data, *error corrected scoring rules* provide a way to incorporate data inaccuracies into the forecast comparison framework. Assuming a certain error distribution, e.g. additive Gaussian noise, they allow for a consistent comparison of the underlying distribution of interest, separate from the error, see Ferro (2017) for details.

Thirdly, there are situations where the statistical property of interest cannot be computed explicitly from a model and is only accessible via simulations. In this case it is intuitive to use simulated realizations in order to approximate the realized score of the forecasts. This idea is a central element of the model comparison approach of Heinrich et al. (2019). *Fair versions* of proper scoring rules aim to reduce a possible bias which might occur in such approximations, see e.g. Ferro (2014) for an overview. Krüger et al. (2020) address this issue when the model simulation is done via Markov chain Monte Carlo methods.

Outlook Based on the results of this chapter, consistent scoring functions for point processes, and the methods relying on them, present themselves as useful tools in evaluating point process-based forecasts and choosing suitable models. Since we focus on working out the theoretical foundations, there are several avenues for future work. A first important task consists of investigations concerning the choice of scoring function, including, but not limited to, simulation studies in the spirit of Section 4.4 or case studies using real data. Moreover, implementing the here proposed methods for concrete real world applications poses more difficult, but highly relevant further problems, in particular if refinements such as weighing or approximations are used. Finally, some technical results which enable DM tests in certain (spatio-) temporal settings, as discussed in Section 4.1, remain to be worked out. We thus believe that model and forecast evaluation for point processes will remain an area of active research.

Bibliography

Aitchison, J. and Dunsmore, I. R. (1968). Linear-loss interval estimation of location and scale parameters. *Biometrika*, 55:141–148.

Anscombe, F. J. (1952). Large-sample theory of sequential estimation. *Proceedings of the Cambridge Philosophical Society*, 48:600–607.

Askanazi, R., Diebold, F. X., Schorfheide, F., and Shin, M. (2018). On the comparison of interval forecasts. *Journal of Time Series Analysis*, 39:953–965.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R.* Chapman and Hall/CRC Press, London.

Baddeley, A. and Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42.

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 67:617–666.

Barndorff-Nielsen, O. (2014). *Information and exponential families in statistical theory.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Barthelmé, S., Trukenbrod, H., Engbert, R., and Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision*, 13:1–34.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance*, 15:725–733.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2020). Evaluating epidemic forecasts in an interval format. *arXiv e-prints*, 2005.12881. URL `https://arxiv.org/pdf/2005.12881.pdf`.

Bray, A. and Schoenberg, F. P. (2013). Assessment of point process models for earthquake forecasting. *Statistical Science*, 28:510–520.

Bray, A., Wong, K., Barr, C. D., and Schoenberg, F. P. (2014). Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts. *The Annals of Applied Statistics*, 8:2247–2267.

Brehmer, J. R. and Gneiting, T. (2020). Scoring interval forecasts: Equal-tailed, shortest, and modal interval. *arXiv e-prints*, 2007.05709. URL `https://arxiv.org/pdf/2007.05709.pdf`.

Brehmer, J. R. and Strokorb, K. (2019). Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, 13:4015–4034.

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

Cadena, M. and Kratz, M. (2016). New results for tails of probability distributions according to their asymptotic decay. *Statistics & Probability Letters*, 109:178–183.

Casella, G., Hwang, J. T. G., and Robert, C. (1993). A paradox in decision-theoretic interval estimation. *Statistica Sinica*, 3:141–155.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, third edition.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39:841–862.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *Economic Journal*, 114:844–866.

Clements, M. P. and Harvey, D. I. (2010). Forecast encompassing tests and probability forecasts. *Journal of Applied Econometrics*, 25:1028–1062.

Clements, R. A., Schoenberg, F. P., and Schorlemmer, D. (2011). Residual analysis methods for space-time point processes with applications to earthquake forecast models in California. *The Annals of Applied Statistics*, 5:2549–2571.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65:1254–1261.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I*. Probability and its Applications (New York). Springer, New York, second edition.

Daley, D. J. and Vere-Jones, D. (2004). Scoring probability forecasts for point processes: The entropy score and information gain. *Journal of Applied Probability*, 41A:297–312.

Daley, D. J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes. Vol. II*. Probability and its Applications (New York). Springer, New York, second edition.

Dawid, A. P. (1984). Statistical theory. The prequential approach. *Journal of the Royal Statistical Society. Series A. General*, 147:278–292.

Dawid, A. P. (1986). Probability forecasting. In Kotz, S., Johnson, N. L., and Read, C. B., editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. John Wiley & Sons, Inc., New York.

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

Dawid, A. P., Lauritzen, S., and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40:593–608.

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72:169–183.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *The Annals of Statistics*, 27:65–81.

de Haan, L. and Ferreira, A. (2006). *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York.

Dearborn, K. and Frongillo, R. (2020). On the indirect elicitability of the mode and modal interval. *Annals of the Institute of Statistical Mathematics*, 72:1095–1108.

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

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

Dimitriadis, T. and Schnaitmann, J. (2020). Forecast encompassing tests for the expected shortfall. *arXiv e-prints*, 1908.04569. URL `https://arxiv.org/pdf/1908.04569.pdf`.

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

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 78:505–562.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk*, 18:31–60.

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140:1917–1923.

Ferro, C. A. T. (2017). Measuring forecast performance in the presence of observation error. *Quarterly Journal of the Royal Meteorological Society*, 143:2665–2676.

Ferro, C. A. T. and Stephenson, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26:699–713.

Field, E. H. (2007). Overview of the working group for the development of regional earthquake likelihood models (RELM). *Seismological Research Letters*, 78:7–16.

Fissler, T., Frongillo, R., Hlavinová, J., and Rudloff, B. (2020). Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *arXiv e-prints*, 1910.07912. URL `https://arxiv.org/pdf/1910.07912.pdf`.

Fissler, T. and Ziegel, J. F. (2015). Higher order elicitability and Osband's principle. *arXiv e-prints*, 1503.08123. URL `https://arxiv.org/pdf/1503.08123v3.pdf`.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44:1680–1707.

Fissler, T. and Ziegel, J. F. (2019a). Erratum: Higher order elicitability and Osband's principle. *arXiv e-prints*, 1901.08826. URL `https://arxiv.org/pdf/1901.08826.pdf`.

Fissler, T. and Ziegel, J. F. (2019b). Order-sensitivity and equivariance of scoring functions. *Electronic Journal of Statistics*, 13:1166–1211.

Flaxman, S., Chirico, M., Pereira, P., and Loeffler, C. (2019). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "Real-Time Crime Forecasting Challenge". *The Annals of Applied Statistics*, 13:2564–2585.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D., and Bertozzi, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, 111:564–584.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23:579–594.

Frongillo, R. and Kash, I. A. (2015). Vector-valued property elicitation. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 40:1–18.

Frongillo, R. and Kash, I. A. (2019). General truthfulness characterizations via convex analysis. *arXiv e-prints*, 1211.3043. URL `https://arxiv.org/pdf/1211.3043.pdf`.

Frongillo, R. and Kash, I. A. (2020). Elicitation complexity of statistical properties. *arXiv e-prints*, 1506.07212. URL `https://arxiv.org/pdf/1506.07212.pdf`.

Giacomini, R. and Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, 23:416–431.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica. Journal of the Econometric Society*, 74:1545–1578.

Gneiting, T. (2011a). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762.

Gneiting, T. (2011b). Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27:197–207.

Gneiting, T. (2017). When is the mode functional the Bayes classifier? *Stat*, 6:204–206.

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

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

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

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

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

Gut, A. (2012). Anscombe's theorem 60 years later. *Sequential Analysis. Design Methods & Applications*, 31:368–396.

Harte, D. (2015). Log-likelihood of earthquake models: Evaluation of models and forecasts. *Geophysical Journal International*, 201:711–723.

Harte, D. and Vere-Jones, D. (2005). The entropy score and its uses in earthquake forecasting. *Pure and Applied Geophysics*, 162:1229–1253.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90.

Heinrich, C. (2014). The mode functional is not elicitable. *Biometrika*, 101:245–251.

Heinrich, C., Schneider, M., Guttorp, P., and Thorarinsdottir, T. (2019). Validation of point process forecasts. Preprint, available at `https://www.nr.no/en/nrpublication?query=/file/1564572954/PointProcessValidation-Heinrich.pdf`.

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *Annals of Mathematical Statistics*, 42:1916–1921.

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

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32:896–913.

Hough, J. B., Krishnapur, M., Peres, Y., and Virág, B. (2006). Determinantal processes and independence. *Probability Surveys*, 3:206–229.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition.

Hwang, E. and Shin, D. W. (2012). Random central limit theorems for linear processes with weakly dependent innovations. *Journal of the Korean Statistical Society*, 41:313–322.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. Statistics in Practice. John Wiley & Sons, Ltd., Chichester.

Kagan, Y. Y. and Jackson, D. D. (1995). New seismic gap hypothesis: Five years after. *Journal of Geophysical Research: Solid Earth*, 100:3943–3959.

Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, 32:788–803.

Krüger, F., Lerch, S., Thorarinsdottir, T. L., and Gneiting, T. (2020). Predictive inference based on Markov chain Monte Carlo output. *arXiv e-prints*, 1608.06802. URL `https://arxiv.org/pdf/1608.06802.pdf`.

Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, EC '08, pages 129–138. ACM.

Lambert, N. S. and Shoham, Y. (2009). Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 109–118. ACM.

Lavancier, F., Møller, J., and Rubak, E. (2015). Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 77:853–877.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187.

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society. Series B. Methodological*, 59:475–499.

Lee, S. (1997). Random central limit theorem for the linear process generated by a strong mixing process. *Statistics & Probability Letters*, 35:189–196.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32:106–127.

M Open Forecasting Center (2020). The M5 competition: Competitor's Guide. Available at `https://mofc.unic.ac.cy/m5-competition/`.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36:54–74.

Marzocchi, W., Zechar, J. D., and Jordan, T. H. (2012). Bayesian forecast evaluation and ensemble earthquake forecasting. *Bulletin of the Seismological Society of America*, 102:2574–2584.

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

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42:654–655.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management.* Princeton Series in Finance. Princeton University Press, Princeton, NJ, revised edition.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8:1612–1639.

Mikosch, T. (2009). *Non-life insurance mathematics.* Universitext. Springer-Verlag, Berlin, second edition.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106:100–108.

Nau, R. F. (1985). Should scoring rules be 'effective'? *Management Science*, 31:527–535.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica. Journal of the Econometric Society*, 55:819–847.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11:1833–1874.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27:23–31.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50:379–402.

Ogata, Y. (2013). A prospect of earthquake prediction research. *Statistical Science*, 28:521–541.

Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society. Series B. Methodological*, 46:496–518.

Osband, K. (1985). *Providing Incentives for Better Cost Forecasting.* PhD thesis, University of California, Berkeley.

Peng, R. D., Schoenberg, F. P., and Woods, J. A. (2005). A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, 100:26–35.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33:299–318.

Rényi, A. (1957). On the asymptotic distribution of the sum of a random number of independent random variables. *Acta Mathematica. Academiae Scientiarum Hungaricae*, 8:193–199.

Resnick, S. I. (1987). *Extreme values, regular variation, and point processes*, volume 4 of *Applied Probability. A Series of the Applied Probability Trust.* Springer-Verlag, New York.

Rhoades, D., Schorlemmer, D., Gerstenberger, M., Christophersen, A., Zechar, J. D., and Imoto, M. (2011). Efficient testing of earthquake forecasting models. *Acta Geophysica*, 59:728–747.

Rockafellar, R. T. (1970). *Convex analysis.* Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J.

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

Schoenberg, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, 98:789–795.

Schoenberg, F. P., Hoffmann, M., and Harrigan, R. J. (2019). A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics*, 71:1271–1287.

Schorlemmer, D., Gerstenberger, M., Wiemer, S., and Jackson, D. (2007). Earthquake likelihood model testing. *Seismological Research Letters*, 78:17–29.

Schorlemmer, D. and Gerstenberger, M. C. (2007). RELM testing center. *Seismological Research Letters*, 78:30–36.

Schorlemmer, D., Werner, M. J., Marzocchi, W., Jordan, T. H., Ogata, Y., Jackson, D. D., Mak, S., Rhoades, D. A., Gerstenberger, M. C., Hirata, N., Liukis, M., Maechling, P. J., Strader, A., Taroni, M., Wiemer, S., Zechar, J. D., and Zhuang, J. (2018). The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismological Research Letters*, 89:1305–1313.

Shang, Y. (2012). A central limit theorem for randomly indexed m-dependent random variables. *Filomat*, 26:713–717.

Steinwart, I., Pasin, C., Williamson, R., and Zhang, S. (2014). Elicitation and identification of properties. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 35:1–45.

Stephenson, D. B., Casati, B., Ferro, C. A. T., and Wilson, C. A. (2008). The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorological Applications*, 15:41–50.

Stoyan, D. and Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*, 15:61–78.

Strader, A., Schneider, M., and Schorlemmer, D. (2017). Prospective and retrospective evaluation of five-year earthquake forecast models for California. *Geophysical Journal International*, 211:239–251.

Taillardat, M., Fougères, A.-L., Naveau, P., and De Fondeville, R. (2019). Extreme events evaluation using CRPS distributions. *arXiv e-prints*, 1905.04022. URL `https://arxiv.org/pdf/1905.04022.pdf`.

Taylor, S. W., Woolford, D. G., Dean, C. B., and Martell, D. L. (2013). Wildfire prediction to inform fire management: Statistical science challenges. *Statistical Science*, 28:586–615.

Thorarinsdottir, T. L. (2013). Calibration diagnostic for point process models via the probability integral transform. *Stat*, 2:150–158.

Vere-Jones, D. (1998). Probability and information gain for earthquake forecasting. *Computational Seismology and Geodynamics*, 30:248–263.

Weber, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16:419–441.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67:187–191.

Xu, H. and Schoenberg, F. P. (2011). Point process modeling of wildfire hazard in Los Angeles County, California. *The Annals of Applied Statistics*, 5:684–704.

Zechar, J. D., Gerstenberger, M. C., and Rhoades, D. A. (2010). Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts. *Bulletin of the Seismological Society of America*, 100:1184–1195.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97:369–380.