Annals of the Institute of Statistical Mathematics (2020) 72:659–673
https://doi.org/10.1007/s10463-019-00705-7

Properization: constructing proper scoring rules via Bayes acts

Jonas R. Brehmer[1] · Tilmann Gneiting[2,3]

Received: 17 August 2018 / Revised: 14 December 2018 / Published online: 22 February 2019
© The Institute of Statistical Mathematics, Tokyo 2019

Abstract
Scoring rules serve to quantify predictive performance. A scoring rule is proper if truth telling is an optimal strategy in expectation. Subject to customary regularity conditions, every scoring rule can be made proper, by applying a special case of the Bayes act construction studied by Grünwald and Dawid (Ann Stat 32:1367–1433, 2004) and Dawid (Ann Inst Stat Math 59:77–93, 2007), to which we refer as properization. We discuss examples from the recent literature and apply the construction to create new types, and reinterpret existing forms, of proper scoring rules and consistent scoring functions. In an abstract setting, we formulate sufficient conditions under which Bayes acts exist and scoring rules can be made proper.

Keywords Bayes act · Consistent scoring function · Forecast evaluation · Misclassification error · Proper scoring rule

1 Introduction

Let $\mathscr{B}$ be a σ-algebra of subsets of a general sample space Ω. Let $\mathscr{P}$ be a convex class of probability measures on $(\Omega, \mathscr{B})$. A *scoring rule* is any extended real-valued function S on $\mathscr{P} \times \Omega$ such that

✉ Jonas R. Brehmer
jbrehmer@mail.uni-mannheim.de

Tilmann Gneiting
tilmann.gneiting@h-its.org

[1] Institute for Mathematics, University of Mannheim, A5, 6, 68131 Mannheim, Germany

[2] Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Englerstraße 2, 76131
Karlsruhe, Germany

[3] Computational Statistics Group, Heidelberg Institute for Theoretical Studies,
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

$$S(P, Q) = \int S(P, \omega) \, \mathrm{d}Q(\omega)$$

is well defined for $P, Q \in \mathscr{P}$. The scoring rule S is *proper* relative to $\mathscr{P}$ if

$$S(Q, Q) \leq S(P, Q) \quad \text{for all} \quad P, Q \in \mathscr{P}. \tag{1}$$

In words, we take scoring rules to be negatively oriented penalties that a forecaster wishes to minimize. If she believes that a future quantity or event has distribution Q, and the penalty for quoting the predictive distribution P when ω realizes is $S(P, \omega)$, then (1) implies that quoting $P = Q$ is an optimal strategy in expectation. The scoring rule is *strictly proper* if (1) holds with equality only if $P = Q$. For recent reviews of the theory and application of proper scoring rules, see Dawid (2007), Gneiting and Raftery (2007), Dawid and Musio (2014) and Gneiting and Katzfuss (2014).

The intent of this note is to draw attention to the simple fact that, subject to customary regularity conditions, any scoring rule can be *properized*, in the sense that it can be modified in a straightforward way to yield a proper scoring rule, so that truth telling becomes an optimal strategy. Implicitly, this construction has recently been used by various authors in various types of applications; see, e.g., Diks et al. (2011), Christensen et al. (2014) and Holzmann and Klar (2017).

Theorem 1 (properization) *Let S be a scoring rule. Suppose that for every $P \in \mathscr{P}$ there is a probability distribution $P^* \in \mathscr{P}$ such that*

$$S(P^*, P) \leq S(Q, P) \quad \textit{for all} \quad Q \in \mathscr{P}. \tag{2}$$

Then the function

$$S^* : \mathscr{P} \times \Omega \to \bar{\mathbb{R}}, \quad (P, \omega) \mapsto S^*(P, \omega) = S(P^*, \omega), \tag{3}$$

is a proper scoring rule.

Here and in what follows, we denote the real line by $\mathbb{R}$ and the extended real line by $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$. Any probability measure P^* with the property (2) is commonly called *Bayes act*; for the existence of Bayes acts, see Sect. 3. In case there are multiple minimizers of the expected score $Q \mapsto S(Q, P)$, the function S^* is well-defined by using a mapping $P \mapsto P^*$ that chooses a P^* out of the set of minimizers. If S is proper and $P^* = P$, then $S^* = S$, so the proper scoring rules are fixed points under the properization operator.

Importantly, Theorem 1 is a special case of a general and powerful construction studied in detail by Grünwald and Dawid (2004) and Dawid (2007). Specifically, given some action space $\mathscr{A}$ and a loss function $L : \mathscr{A} \times \Omega \to \bar{\mathbb{R}}$, suppose that for each $P \in \mathscr{P}$ there is a Bayes act $a_P \in \mathscr{A}$, such that

$$\int L(a_P, \omega) \, \mathrm{d}P(\omega) \leq \int L(a, \omega) \, \mathrm{d}P(\omega) \quad \text{for all} \quad a \in \mathscr{A}.$$

Then the function

$$S^* : \mathscr{P} \times \Omega \to \bar{\mathbb{R}}, \quad (P, \omega) \mapsto S^*(P, \omega) = L(a_P, \omega),$$

is a proper scoring rule. Note the natural connection to decision- and utility-based scoring approaches, where the quality of a forecast is judged by the monetary utility of the induced acts and decisions (Granger and Pesaran 2000; Granger and Machina 2006; Ehm et al. 2016).

In general, a link function $\alpha : \mathscr{P} \to \mathscr{A}$, where $\alpha(P)$ is not necessarily a Bayes act for P, can be used to construct scoring rules from loss functions. Moreover, the reverse is possible and applied in a recent strand of the machine learning literature, where for finite Ω and bijective α a loss function on $\mathscr{A} \times \Omega$ is constructed by composing a scoring rule with α^{-1}. For details on this type of *composite loss*, see Reid and Williamson (2010), van Erven et al. (2012, Section 6) and Williamson et al. (2016).

In the remainder of the paper, we focus on the above special case in which the action domain $\mathscr{A}$ is the class $\mathscr{P}$. In Sect. 2, we identify scattered results in the literature as prominent special cases of properization (Examples 1–4), and we use Theorem 1 to construct new proper scoring rules from improper ones (Examples 5–7). Section 3 gives sufficient conditions for the existence of Bayes acts, and Sect. 4 contains a brief discussion. All proofs and technical details are moved to the "Appendix."

2 Examples

This section starts with an example in which we review the ubiquitous misclassification error from the perspective of properization. We go on to demonstrate how Theorem 1 has been used implicitly to construct proper scoring rules in econometric, meteorological, and statistical strands of literature. The notion of properization simplifies and shortens the respective proofs of propriety, makes them much more transparent, and puts the scattered examples into a unifying and principled joint framework. Further examples show other facets of properization: The scoring rules constructed in Example 5 are original, and the discussion in Example 6 illustrates a connection to the practical problem of the treatment of observational uncertainty in forecast evaluation. Finally, Example 7 includes an instance of a situation in which properization fails.

Example 1 Consider probability forecasts of a binary event, where $\Omega = \{0, 1\}$ and $\mathscr{P}$ is the class of the Bernoulli measures. We identify any $P \in \mathscr{P}$ with the probability $p = P(\{1\}) \in [0, 1]$ and consider the scoring rules

$$S_1(P, \omega) := 1 - p\omega - (1 - p)(1 - \omega) \quad \text{and} \quad S_2(P, \omega) := |p - \omega|.$$

The scoring rule S_1 corresponds to the mean probability rate (MPR) in machine learning (Ferri et al. 2009, p. 30). The equivalent form in S_2 was first considered by Dawid (1986). It agrees with the special case $c_1 = c_2$ in Section 4.2 of Parry (2016) and also corresponds to the mean absolute error (MAE) as discussed by Ferri et al. (2009,

p. 30).[1] These scoring rules are improper with Bayes act

$$p^* = \mathbb{1}\left(p \geq \tfrac{1}{2}\right) \in \{0, 1\},$$

and with properized score given by the zero-one rule

$$S^*(P, \omega) = \begin{cases} 0, & p^* = \omega, \\ 1, & \text{otherwise}. \end{cases}$$

A case-averaged zero-one score is typically referred to as *misclassification rate* or *misclassification error*; undoubtedly, this is the most popular and most frequently used performance measure in binary classification. While the scoring rule S^* is proper, it fails to be strictly proper (Gneiting and Raftery 2007, Example 4; Parry 2016, Section 4.3). Consequently, misclassification error has serious limitations as a performance measure, as persuasively argued by Harrell (2015, p. 258), among others. Nevertheless, the scoring rule S^* is proper, contrary to recent claims of impropriety in the blogosphere.[2]

For the remainder of the section, let $\Omega = \mathbb{R}$ and let $\mathscr{B}$ be the Borel σ-algebra. We let $\mathscr{L}$ be the class of Borel measures P with a Lebesgue density, p. Furthermore, we write $\mathscr{P}_k$ for the measures with finite kth moment and $\mathscr{P}_k^+$ for the subclasses when Dirac measures are excluded. Whenever it simplifies notation, we identify P with its cumulative distribution function $x \mapsto P((-\infty, x])$.

Example 2 Let S_0 be a proper scoring rule on some subclass $\mathscr{P}$ of $\mathscr{L}$ and let w be a nonnegative weight function such that $0 < \int w(z)\, p(z)\, \mathrm{d}z < \infty$ for $p \in \mathscr{P}$. Let

$$S : \mathscr{P} \times \mathbb{R} \to \mathbb{R}, \quad (P, y) \mapsto S(P, y) = w(y)\, S_0(P, y);$$

this score is improper unless the weight function is constant. Indeed, by Theorem 1 of Gneiting and Ranjan (2011), the Bayes act P^* under S has density

$$p^*(y) = \frac{w(y)\, p(y)}{\int w(z)\, p(z)\, \mathrm{d}z}.$$

From this, we see that the key statement in Theorem 1 of Holzmann and Klar (2017) constitutes a special case of Theorem 1. In the further special case in which S_0 is the logarithmic score, the properized score (3) recovers the conditional likelihood score of Diks et al. (2011) up to equivalence, as noted in Example 1 of Holzmann and Klar (2017). For analogous results for consistent scoring functions, see Theorem 5 of Gneiting (2011) and Example 2 of Holzmann and Klar (2017).

[1] As noted by Parry (2016), the improper score S_1 shares its (concave) expected score function $P \mapsto S_1(P, P)$ with the proper Brier score. This illustrates the importance of the second condition in Theorem 1 of Gneiting and Raftery (2007): For a scoring rule S, the (strict) concavity of the expected score function $G(P) := S(P, P)$ is equivalent to the (strict) propriety of S only if, furthermore, $-S(P, \cdot)$ is a subtangent of $-G$ at P.

[2] See, e.g., http://www.fharrell.com/post/class-damage/ and http://www.fharrell.com/post/classification/.

Example 3 For a probability measure $P \in \mathscr{P}_4$, let μ_P, σ_P^2, and γ_P denote its mean, variance, and centered third moment. Let

$$S(P, y) = \left(\sigma_P^2 - (y - \mu_P)^2\right)^2$$

be the "trial score" in equation (16) of Christensen et al. (2014). As Christensen et al. (2014) show in their "Appendix A," if $\sigma_P^2 > 0$, any Bayes act P^* under S has mean $\mu_P + \frac{1}{2}\frac{\gamma_P}{\sigma_P^2}$ and variance

$$\sigma_P^2 \left(1 + \frac{1}{4}\frac{\gamma_P^2}{\sigma_P^6}\right),$$

so properization yields the spread-error score,

$$S^*(P, y) = \left(\sigma_P^2 - (y - \mu_P)^2 + (y - \mu_P)\,\frac{\gamma_P}{\sigma_P^2}\right)^2,$$

which is proper relative to the class $\mathscr{P}_4^+$. Hence, the construction of the spread-error score in Christensen et al. (2014) constitutes another special case of Theorem 1.

Example 4 The predictive model choice criterion of Laud and Ibrahim (1995) and Gelfand and Ghosh (1998) uses the scoring rule $S(P, y) = (y - \mu_P)^2 + \sigma_P^2$, where μ_P and σ_P^2 denote the mean and the variance of a distribution $P \in \mathscr{P}_2$, respectively. As pointed out by Gneiting and Raftery (2007), this score fails to be proper. Specifically, any Bayes act P^* under S has mean μ_P and vanishing variance, so properization yields the ubiquitous squared error, $S^*(P, y) = (y - \mu_P)^2$.

The original scoring rules of Examples 3 and 4 can be interpreted as functions $L : \mathscr{A} \times \Omega \to \mathbb{R}$ in the Bayes act setting of Grünwald and Dawid (2004) and Dawid (2007), where the action space $\mathscr{A}$ is given by $\mathbb{R} \times [0, \infty)$, as we formalize in Sect. 3. Hence, the properization method can be interpreted as an application of Theorem 3 of Gneiting (2011) to consistent scoring functions for elicitable two-dimensional functionals, as discussed by Fissler and Ziegel (2016).

Detailed arguments and calculations for the subsequent examples are deferred to the "Appendix."

Example 5 For $\alpha > 0$ consider the scoring rule

$$S_\alpha(P, y) = \int |P(x) - \mathbb{1}(y \le x)|^\alpha \, \mathrm{d}x,$$

where P is identified with its cumulative distribution function (CDF). For $\alpha = 2$, this is the well-known proper continuous ranked probability score (CRPS), as reviewed in Section 4.2 of Gneiting and Raftery (2007). For $\alpha = 1$, the score S_α was proposed by Müller et al. (2005), and Zamo and Naveau (2018) show in their "Appendix A" that for discrete distributions every Dirac measure in a median of P is a Bayes act. The same holds true for general distributions and for all $\alpha \in (0, 1]$. If $\alpha > 1$, the Bayes act P^* under S_α is given by

$$P^*(x) = \left(1 + \left(\frac{1 - P(x)}{P(x)}\right)^{1/(\alpha-1)}\right)^{-1} \mathbb{1}\left(P(x) > 0\right), \tag{4}$$

and all in all we see that properization of S_α works for any $\alpha > 0$.

Moreover, in the case $\alpha > 1$ the mapping $P \mapsto P^*$ is even injective. Consequently, if the class $\mathscr{P}$ is such that $P^* \in \mathscr{P}$ and $S_\alpha(P^*, P)$ is finite for $P \in \mathscr{P}$, the properized score (3) is even strictly proper relative to $\mathscr{P}$. If $\alpha \in (1, 2]$, this can be ensured by restricting S_α to the class $\mathscr{P}_1$. For $\alpha > 2$, the class $\mathscr{P}_\mathrm{c}$ of the Borel measures with compact support is a suitable choice.

Example 6 Friederichs and Thorarinsdottir (2012, p. 58) propose a modification of the CRPS that aims to account for observational error in forecast evaluation. Specifically, they consider the scoring rule

$$S_\Phi(P, y) = \int |P(x) - \Phi(x - y)|^2 \, \mathrm{d}x,$$

where $\Phi \in \mathscr{P}_1^+$ represents additive observation error. This scoring rule fails to be proper, as for probability measures $P, Q \in \mathscr{P}_1$ we have

$$S_\Phi(P, Q) = \mathrm{CRPS}(P, Q * \Phi) - \mathrm{CRPS}(\Phi, \Phi), \tag{5}$$

where $*$ denotes the convolution operator. Due to the strict propriety of the CRPS relative to the class $\mathscr{P}_1$, the unique Bayes act under S_Φ is given by $P^* = P * \Phi$. Theorem 1 now gives the scoring rule $S(P, y) := S_\Phi(P^*, y)$, which is proper relative to $\mathscr{P}_1$.

In order to account for noisy observational data in forecast evaluation, Eq. (5) suggests using the scoring rule $S(P, y) := \mathrm{CRPS}(P^*, y)$ if the noise is independent, additive, and has distribution Φ. This corresponds to predicting hypothetical true values, to which noise is added before they are compared to observations. The drawbacks of this approach and alternative techniques are discussed by Ferro (2017). The associated issues in forecast evaluation remain challenges to the scientific community at large; see, e.g., Ebert et al. (2013) and Ferro (2017).

Example 7 Let S be a scoring rule, and let $\Phi \in \mathscr{L}$ be a distribution with Lebesgue density φ. Suppose $\mathscr{P}$ is a class of distributions such that $P * \Phi \in \mathscr{P}$ for $P \in \mathscr{P}$. For $P \in \mathscr{P}$, define

$$S^\varphi(P, y) := \int \varphi(x - y) \, S(P, x) \, \mathrm{d}x,$$

which is again a scoring rule. If S is proper, a Bayes act under S^φ is given by $P^* = P*\Phi$, since $S^\varphi(P, Q) = S(P, Q*\Phi)$ for $Q \in \mathscr{P}$, and if S is strictly proper, the Bayes act is unique. Properization now gives the proper scoring rule $S(P, y) := S^{\varphi}(P^*, y)$. An interesting special case emerges when substituting the CRPS for S. This leads to

$$\mathrm{CRPS}^{\varphi}(P, y) = S_{\Phi}(P, y) + \mathrm{CRPS}(\Phi, \Phi), \tag{6}$$

where S_{Φ} is the scoring rule in the previous example. For another special case, let $c > 0$ and $P \in \mathscr{L}$, to yield

$$\mathrm{PS}_c(P, y) := -\int_{y-c}^{y+c} p(x)\,\mathrm{d}x,$$

which recovers the *probability score* of Wilson et al. (1999). We have that $\mathrm{PS}_c = 2c\,\mathrm{LinS}^{\varphi_c}$, where $\mathrm{LinS}(P, y) := -p(y)$ is the improper linear score and φ_c is a uniform density on $[-c, c]$. Properization is not feasible relative to sufficiently rich classes $\mathscr{P}$, as Bayes acts fail to exist under both the linear score and the probability score. For details, see the “Appendix.”

3 Existence of Bayes acts

In Example 7, we presented a scoring rule that cannot be properized, due to the non-existence of Bayes acts. This section addresses the question under which conditions on S and $\mathscr{P}$ a minimum of the expected score function exists. To illustrate the ideas, we start with a further example.

Example 8 Using the notation of Example 3, consider the normalized squared error (Gelman et al. 2014, p. 998),

$$S(P, y) = \frac{(y - \mu_P)^2}{\sigma_P^2},$$

as a scoring rule on the classes $\mathscr{P}_{2,m}^{+}$ of the Borel measures with variance at most m, and $\mathscr{P}_2^{+} = \cup_{m>0}\mathscr{P}_{2,m}^{+}$, respectively. Relative to $\mathscr{P}_{2,m}^{+}$ any Bayes act P^* under S has mean μ_P and variance m, so properization yields (non-normalized) squared error up to equivalence. Relative to $\mathscr{P}_2^{+}$ however, there is no Bayes act, since increasing the variance will always lead to a smaller expected score.

We now turn to a general perspective and discuss sufficient conditions for the existence of Bayes acts. At first, consider a finite probability space $\Omega = \{\omega_1, \ldots, \omega_k\}$. In this situation, geometrical arguments yield sufficient conditions. In particular, a Bayes act under S exists if the *risk set*

$$\mathscr{S} := \{(x_1, \ldots, x_k) \mid \exists\, P \in \mathscr{P} : x_j = S(P, \omega_j),\ j = 1, \ldots, k\} \subset \mathbb{R}^k$$

is closed from below and bounded from below; see Theorem 1 in Chapter 2.5 of Ferguson (1967). Extending this result to a general sample space Ω is non-trivial since in this case $\mathscr{S}$ can be a subset of an infinite-dimensional vector space. In the following, we employ well-known concepts of functional analysis in order to discuss extensions. All proofs are deferred to the “Appendix.”

Let $\mathscr{P}$ be a set of probability measures on a general probability space Ω and let $\mathscr{A}$ be a topological space. We return to the setting of Sect. 1 and consider functions of the form

$$S(P,\omega) = s(\alpha(P),\omega)$$

with mappings $\alpha : \mathscr{P} \to \mathscr{A}$ and $s : \mathscr{A} \times \Omega \to \bar{\mathbb{R}}$, such that

$$s(a, P) = \int s(a,\omega)\,\mathrm{d}P(\omega)$$

is well defined for all $a \in \mathscr{A}$ and $P \in \mathscr{P}$. This makes the results easier to apply in situations where the scoring rule depends on P only via some finite number of parameters. Concerning the latter point, note that the normalized squared error of Example 8 can be written as a composition of the mappings $\alpha(P) := (\mu_P, \sigma_P^2)$ and $s(x_1, x_2, y) := (y - x_1)^2/x_2$, with s being defined on $\mathscr{A} \times \Omega = \mathbb{R} \times (0,\infty) \times \mathbb{R}$. Consequently, the expected normalized squared error attains its minimum if the expected score of s attains its minimum. Note that such a decomposition of the scoring rule is possible for Examples 3 and 4 as well, as alluded to in the comments that succeed these examples.

We impose the following integrability condition.

Definition 1 The mapping $s : \mathscr{A} \times \Omega \to \bar{\mathbb{R}}$ is *uniformly bounded from below* if there exists a function $g : \Omega \to \mathbb{R}$ which is integrable with respect to any $P \in \mathscr{P}$ and such that $s(a,\cdot) \geq g(\cdot)$ holds for all $a \in \mathscr{A}$.

Our first result is similar to Theorem 2 in Chapter 2.9 of Ferguson (1967), which proves the existence of minimax decision rules.

Theorem 2 *Suppose s is lower semicontinuous in its first component and uniformly bounded from below. If $\mathscr{A}$ is compact, then the function $a \mapsto s(a, P)$ attains its minimum for any $P \in \mathscr{P}$.*

This theorem can be used to prove the existence of a Bayes act for a given scoring rule. However, its applicability to Example 8 is limited. To see this, recall the above decomposition and note that restricting S to $\mathscr{P}_{2,m}^{+}$ corresponds to restricting s to $\mathbb{R} \times (0, m]$. The latter set is not a compact space and neither is its closure. We can further restrict the domain of S to Borel measures with means in some bounded interval and variances that are bounded away from both zero and infinity. This corresponds to restricting s to $[-t, t] \times [1/m, m]$ for some $t > 0$ and $m > 0$, and Theorem 2 now applies. As a consequence of this observation, we aim to dispense with the compactness assumption.

To do so, we need additional concepts from functional analysis. Let $\mathscr{X}$ be a real normed vector space. Recall that a function $h : \mathscr{X} \to \mathbb{R}$ is called *coercive* if for any sequence $(x_n)_{n\in\mathbb{N}} \subset \mathscr{X}$ the implication

$$\lim_{n\to\infty} \|x_n\| = \infty \quad \Rightarrow \quad \lim_{n\to\infty} h(x_n) = \infty$$

holds true, see, e.g., Definition III.5.7 in Werner (2018). By *weak topology* on $\mathscr{X}$, we mean the weakest topology such that all real-valued linear mappings on $\mathscr{X}$ are continuous; see, e.g., Chapters 2.13 and 6.5 in Aliprantis and Border (2006). The space $\mathscr{X}$ is called a *reflexive Banach space* if it is complete and the canonical embedding of $\mathscr{X}$ into its bidual space is surjective; see, e.g., Chapter III.3 in Werner (2018) or Chapter 6.3 in Aliprantis and Border (2006). Combining these concepts, we obtain a complement to Theorem 2.

Theorem 3 *Let $\mathscr{A}$ be a weakly closed subset of a reflexive Banach space. Moreover, suppose s is weakly lower semicontinuous in its first component and uniformly bounded from below. If the function $a \mapsto s(a, P)$ is coercive, then it attains its minimum.*

This result yields the existence of Bayes acts as long as the integrated scoring rule is coercive for any $P \in \mathscr{P}$, where $\mathscr{P}$ is a reflexive Banach space. We can connect to Example 8 as follows: The function $s(\cdot, \cdot, y)$ from the above decomposition of S is defined on $\mathbb{R} \times (0, \infty)$, which is a subset of the reflexive Banach space $\mathbb{R}^2$. Moreover, s is bounded from below by zero and continuous in its first component. As mentioned above, restricting the class $\mathscr{P}_2^+$ to $\mathscr{P}_{2,m}^+$ corresponds to restricting the domain of s to $\mathbb{R} \times (0, m]$ and in this situation, integrating s with respect to y gives a coercive function. Consequently, Theorem 3 can be used to show that S can be properized if restricted to $\mathscr{P}_{2,m}^+$.

We conclude this section by stressing that Theorem 3 represents only one of several possible ways to modify Theorem 2. Its limitations are illustrated in the following example. Details are again deferred to the “Appendix.”

Example 9 For $x, y \in \mathbb{R}$, the symmetric absolute percentage error (sAPE) is defined as

$$s(x, y) := \begin{cases} 0, & x = y, \\ \frac{|x-y|}{|x|+|y|}, & x \neq y. \end{cases}$$

It features prominently in forecast contests, such as the recent M4 competition (M4 Team 2018; Makridakis et al. 2018), where the sAPE is used to rank participants and award prizes. For any probability measure P and $x \in \mathbb{R}$, we have $s(x, P) \in [0, 1]$, and the mapping $x \mapsto s(x, P)$ is continuous in $\mathbb{R}\backslash\{0\}$. Moreover,

$$\lim_{x\to-\infty} s(x, P) = \lim_{x\to 0-} s(x, P) = \lim_{x\to 0+} s(x, P) = \lim_{x\to\infty} s(x, P) = 1 \qquad (7)$$

and $s(0, P) = P(\mathbb{R}\backslash\{0\})$. The behavior of the expected score thus implies that Bayes acts exist. In particular, any scoring rule obtained from a composition with the sAPE can be properized. However, the mapping $x \mapsto s(x, P)$ is defined on a non-compact set and fails to be coercive, so Theorems 2 and 3 do not apply.

4 Discussion

In this article, we have introduced the concept of properization, which is rooted in the Bayes act construction of Grünwald and Dawid (2004) and Dawid (2007), and we have drawn attention to its widespread implicit use in the transdisciplinary literature on proper scoring rules, where our unified approach yields simplified, shorter, and considerably more instructive and transparent proofs than extant methods. Moreover, using new examples, we have demonstrated the power of the properization approach in the creation of new proper scoring rules from existing improper ones. We anticipate further, important uses of the general Bayes act construction in a wide range of applied settings, where scoring rules are to be tailored to forecast users' needs (Ebert et al. 2018).

Since the central element in the construction of a properized score is a Bayes act, we have discussed conditions on the scoring rule S and the class $\mathscr{P}$ that guarantee its existence. Undoubtedly, there are alternative paths to existence results in the spirit of Theorems 2 and 3, and the derivation of sufficient conditions in alternative situations is an interesting open problem. The expected score in Example 9 hints at more general conditions since it is not coercive, but its simple asymptotic behavior nevertheless ensures the existence of global minima. Furthermore, we have not explored necessary conditions for the existence of Bayes acts in this work.

A useful generic heuristic appears to be that Bayes acts exist (and properization is feasible) if the scoring rule is selective, in the sense that there must not be a sequence $(P_n)_{n \in \mathbb{N}}$ in $\mathscr{P}$ such that $S(P_n, \omega)$ tends to the infimum of $S(\cdot, \omega)$ for all $\omega \in \Omega$, and the score is bounded from below in a suitable sense. In Example 7, the linear score and the probability score satisfy the former condition but not the latter. In Example 8, the normalized squared error is bounded from below but fails to be selective. The scoring rules in our other examples admit Bayes acts and satisfy both of these conditions. Generally, the derivation of necessary conditions and the refinement of sufficient conditions for the existence of Bayes acts remain challenges that we leave for future work.

Acknowledgements Tilmann Gneiting is grateful for funding by the Klaus Tschira Foundation and by the European Union Seventh Framework Programme under grant agreement 290976. Part of his research leading to these results has been done within subproject C7 "Statistical postprocessing and stochastic physics for ensemble predictions" of the Transregional Collaborative Research Center SFB / TRR 165 "Waves to Weather" (www.wavestoweather.de) funded by the German Research Foundation (DFG). Jonas Brehmer gratefully acknowledges support by DFG through Research Training Group RTG 1953. We thank Tobias Fissler, Rafael Frongillo, Alexander Jordan, and Matthew Parry for instructive discussions, and we are grateful to the editor and two anonymous referees for thoughtful comments and suggestions.

Appendix: Proofs

Here, we present detailed arguments for the technical claims in Examples 5, 6, 7, and 9 as well as the proofs of Theorems 2 and 3.

Details for Example 5

We fix some distribution P and start with the case $\alpha > 1$. An application of Fubini's theorem gives

$$S_\alpha(Q, P) = \int \int |Q(x) - \mathbb{1}(y \le x)|^\alpha \,\mathrm{d}P(y)\,\mathrm{d}x. \tag{8}$$

Given $x \in \mathbb{R}$, we seek the value $Q(x) \in [0, 1]$ that minimizes the inner integral in (8). If x is such that $P(x) \in \{0, 1\}$, the equality $\mathbb{1}(y \le x) = P(x)$ holds for P-almost all y, hence $Q(x) = P(x)$ is the unique minimizer. If x satisfies $P(x) \in (0, 1)$, define the function

$$g_{x,P}(q) := \int |q - \mathbb{1}(y \le x)|^\alpha \,\mathrm{d}P(y) = (1 - P(x))q^\alpha + P(x)(1 - q)^\alpha,$$

which is strictly convex in $q \in (0, 1)$ with derivative

$$g'_{x,P}(q) = \alpha(1 - P(x))q^{\alpha-1} - \alpha P(x)(1 - q)^{\alpha-1}$$

and a unique minimum at $q = q^*_{x,P} \in (0, 1)$. As a consequence, the minimizing value $Q(x)$ is given by

$$Q(x) = q^*_{x,P} = \left(1 + \left(\frac{1 - P(x)}{P(x)}\right)^{1/(\alpha-1)}\right)^{-1}.$$

The function Q defined by the minimizers $Q(x), x \in \mathbb{R}$ is a minimizer of $S_\alpha(\cdot, P)$ and if $S_\alpha(Q, P)$ is finite, it is unique Lebesgue almost surely. Since $\alpha > 1$, the function Q has the properties of a distribution function, and hence, P^* defined by (4) is a Bayes act for P. Moreover, Eq. (4) shows that the relation between P and P^* is one-to-one.

It remains to be checked under which conditions the properization of S_α is not only proper but strictly proper. The representation (4) along with two Taylor expansions implies that P^* behaves like $P^{1/(\alpha-1)}$ in the tails. This has two consequences. At first, the above arguments show that for $S_\alpha(P^*, P)$ to be finite $x \mapsto g_{x,P}(P^*(x))$ has to be integrable with respect to Lebesgue measure. Hence, the tail behavior of P^* and the inequality $\alpha/(\alpha - 1) > 1$ for $\alpha > 1$ show that $S_\alpha(P^*, P)$ is finite for $P \in \mathscr{P}_1$. Second, P^* has a lighter tail than P for $\alpha \in (1, 2)$ and a heavier tail for $\alpha > 2$. In the latter case, $P \in \mathscr{P}_1$ does not necessarily imply $P^* \in \mathscr{P}_1$. Hence, without additional assumptions, strict propriety of the properized score (3) can only be ensured relative to $\mathscr{P}_\mathrm{c}$ for $\alpha > 2$ and relative to the class $\mathscr{P}_1$ for $\alpha \in (1, 2]$.

We now turn to $\alpha \in (0, 1)$. In this case, the function $g_{x,P}$ is strictly concave, and its unique minimum is at $q = 0$ for $P(x) < \frac{1}{2}$ and at $q = 1$ for $P(x) > \frac{1}{2}$. If $P(x) = \frac{1}{2}$, then both 0 and 1 are minima. Arguing as above, every Bayes act P^* is a Dirac measure in a median of P.

Finally, $\alpha = 1$ implies that $g_{x,P}$ is linear, thus, as for $\alpha \in (0, 1)$, every Dirac measure in a median of P is a Bayes act. The only difference to the case $\alpha \in (0, 1)$ is that if there is more than one median, there are Bayes acts other than Dirac measures, since $g_{x,P}$ is constant for all x satisfying $P(x) = \frac{1}{2}$.

Details for Example 6

Let P, Q and Φ be distribution functions. By the definition of the convolution operator

$$\int \mathbb{1}\,(y \le x)\ \mathrm{d}(Q * \Phi)(y) = \int \Phi(x - y)\,\mathrm{d}Q(y)$$

holds for $x \in \mathbb{R}$. Using this identity and Fubini's theorem leads to

$$\begin{aligned} S_\Phi(P, Q) &= \int\int \Big(P(x)^2 - 2P(x)\Phi(x - y) + \Phi(x - y)^2\Big)\ \mathrm{d}Q(y)\,\mathrm{d}x \\ &= \int\int \Big(P(x)^2 - 2P(x)\mathbb{1}\,(y \le x) + \mathbb{1}\,(y \le x)\Big)\ \mathrm{d}(Q * \Phi)(y)\,\mathrm{d}x \\ &\quad + \int\int \Phi(x - y)(\Phi(x - y) - 1)\,\mathrm{d}Q(y)\,\mathrm{d}x \\ &= \int\int (P(x) - \mathbb{1}\,(y \le x))^2\,\mathrm{d}x\,\mathrm{d}(Q * \Phi)(y) - \int \Phi(x)(1 - \Phi(x))\,\mathrm{d}x, \end{aligned}$$

which verifies equality in (5). Moreover, the strict propriety of the CRPS relative to the class $\mathscr{P}_1$ gives $S_\Phi(P, Q) < \infty$ for $P, Q, \Phi \in \mathscr{P}_1$, thereby demonstrating that the Bayes act is unique in this situation.

Details for Example 7

For distributions $P, Q \in \mathscr{P}$ and $c > 0$, the Fubini–Tonelli theorem and the definition of the convolution operator give

$$\begin{aligned} S^\varphi(P, Q) &= -\int\int \varphi(x - y)S(P, x)\,\mathrm{d}Q(y)\,\mathrm{d}x \\ &= \int\int \varphi(x - y)\,\mathrm{d}Q(y)\,S(P, x)\,\mathrm{d}x = S(P, Q * \Phi), \end{aligned}$$

so the stated (unique) Bayes act under S^φ follows from the (strict) propriety of S. Proceeding as in the details for Example 6, we verify identity (6).

For $P \in \mathscr{L}$, the same calculations as above show that the probability score satisfies

$$\mathrm{PS}_c(P, Q) = 2c \int \frac{Q(x + c) - Q(x - c)}{2c}\ \mathrm{LinS}(P, x)\,\mathrm{d}x,$$

where $\mathrm{LinS}(P, y) = -p(y)$ is the linear score. Consequently, to demonstrate that Theorem 1 is neither applicable to PS_c nor to LinS, it suffices to show that there is a distribution Q such that $P \mapsto \mathrm{LinS}(P, Q)$ does not have a minimizer. We use an argument that generalizes the construction in Section 4.1 of Gneiting and Raftery (2007) who show that LinS is improper. Let q be a density, symmetric around zero and strictly increasing on $(-\infty, 0)$. Let $\epsilon > 0$ and define the interval $I_k := ((2k-1)\epsilon, (2k+1)\epsilon]$ for $k \in \mathbb{Z}$. Suppose p is a density with positive mass on some interval I_k for $k \neq 0$. Due to the properties of q, the score $\mathrm{LinS}(P, Q)$ can be reduced by substituting the density defined by

$$\tilde{p}(x) := p(x) - \mathbb{1}\,(x \in I_k)\; p(x) + \mathbb{1}\,(x + 2k\epsilon \in I_k)\; p(x + 2k\epsilon)$$

for p, i.e., by shifting the entire probability mass from I_k to the modal interval I_0. Repeating this argument for any $\epsilon > 0$ shows that no density p can be a minimizer of the expected score $\mathrm{LinS}(P, Q)$. Note that the assumptions on q are stronger than necessary in order to facilitate the argument. They can be relaxed at the cost of a more elaborate proof.

Details for Example 9

For any probability distribution P and $x \in \mathbb{R}$, we obtain

$$s(x, P) = \int \frac{|x-y|}{|x| + |y|}\mathbb{1}\,(x \neq y)\; \mathrm{d}P(y),$$

which immediately gives $s(0, P) = P(\mathbb{R}\backslash\{0\})$. This representation together with the dominated convergence theorem imply the continuity of $x \mapsto s(x, P)$ in $\mathbb{R}\backslash\{0\}$ as well as the limits given in (7).

Proof of Theorem 2

Let $(a_n)_{n\in\mathbb{N}} \subset \mathscr{A}$ be a sequence with $a := \lim_{n\to\infty} a_n$. Since s is lower semicontinuous in its first component and uniformly bounded from below by g, Fatou's lemma gives

$$\liminf_{n\to\infty} \int s(a_n, \omega)\, \mathrm{d}P(\omega) \geq \int \liminf_{n\to\infty} s(a_n, \omega)\, \mathrm{d}P(\omega) \geq s(a, P)$$

for any $P \in \mathscr{P}$. Hence, $a \mapsto s(a, P)$ is a lower semicontinuous function for any $P \in \mathscr{P}$ and due to the assumed compactness of $\mathscr{A}$, the result now follows from Theorem 2.43 in Aliprantis and Border (2006). □

Proof of Theorem 3

The same arguments as in the proof of Theorem 2 show that $a \mapsto s(a, P)$ is a weakly lower semicontinuous function for any $P \in \mathscr{P}$. If $P \in \mathscr{P}$ is such that this function is also coercive, we conclude by proceeding as in the proof of Satz III.5.8 in Werner (2018): In case $\inf_{a\in\mathscr{A}} s(a, P) = \infty$, there is nothing to prove. Otherwise, if $(a_n)_{n\in\mathbb{N}} \subset \mathscr{A}$ is a sequence such that $\lim_{n\to\infty} s(a_n, P) = \inf_{a\in\mathscr{A}} s(a, P)$ holds, the coercivity of $a \mapsto s(a, P)$ implies that this sequence is bounded. Together with the assumption that $\mathscr{A}$ is a subset of a reflexive Banach space, we obtain a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ of $(a_n)_{n\in\mathbb{N}}$ which weakly converges to some element a^*; see, e.g., Theorem 6.25 in Aliprantis and Border (2006). Since $\mathscr{A}$ is weakly closed, it contains a^* and weak lower semicontinuity gives $s(a^*, P) \leq \lim_{k\to\infty} s(a_{n_k}, P) = \inf_{a\in\mathscr{A}} s(a, P)$, concluding the proof. □

References

Aliprantis, C. D., Border, K. C. (2006). *Infinite dimensional analysis* third ed. Berlin: Springer.

Christensen, H. M., Moroz, I. M., Palmer, T. N. (2014). Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, *141*, 538–549.

Dawid, A. P. (1986). Probability forecasting. In S. Kotz, N. L. Johnson, C. B. Read (Eds.), *Encyclopedia of statistical sciences*, Vol. 7, pp. 210–218. New York: Wiley.

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

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

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

Ebert, E., Brown, B., Göber, M., Haiden, T., Mittermaier, M., Nurmi, P., Wilson, L., Jackson, S., Johnston, P., Schuster, D. (2018). The WMO challenge to develop and demonstrate the best new user-oriented forecast verification metric. *Meteorologische Zeitschrift*, *27*, 435–440.

Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Nurmi, P., Gill, P., Göber, M., Joslyn, S., Brown, B., Fowler, T., Watkins, A. (2013). Progress and challenges in forecast verification. *Meteorological Applications*, *20*, 130–139.

Ehm, W., Gneiting, T., Jordan, A., 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.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach. Probability and mathematical statistics*, Vol. 1. New York: Academic Press.

Ferri, C., Hernández-Orallo, J., Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*, 27–38.

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

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

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

Gelfand, A. E., Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, *85*, 1–11.

Gelman, A., Hwang, J., Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*, 746–762.

Gneiting, T., Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

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

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

Granger, C. W., Machina, M. J. (2006). Forecasting and decision theory. In G. Elliott, C. Granger, A. Timmermann (Eds.), *Handbook of economic forecasting*, Vol. 1, pp. 81–98. Amsterdam: Elsevier.

Granger, C. W. J., Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*, 537–560.

Grünwald, P. D., Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, *32*, 1367–1433.

Harrell, F. E, Jr. (2015). *Regression modeling strategies. Springer series in statistics* 2nd ed. Cham: Springer.

Holzmann, H., Klar, B. (2017). Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics*, *11*, 2404–2431.

Laud, P. W., Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society Series B. Methodological*, *57*, 247–262.

M4 Team. (2018). *M4 competitor's guide: Prizes and rules*. Available online at https://www.m4.unic.ac.cy/wp-content/uploads/2018/03/M4-Competitors-Guide.pdf. Accessed 13 Dec 2018.

Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*, 802–808.

Müller, W. A., Appenzeller, C., Doblas-Reyes, F. J., Liniger, M. A. (2005). A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *Journal of Climate*, *18*, 1513–1523.

Parry, M. (2016). Linear scoring rules for probabilistic binary classification. *Electronic Journal of Statistics*, *10*, 1596–1607.

Reid, M. D., Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research*, *11*, 2387–2422.

van Erven, T., Reid, M. D., Williamson, R. C. (2012). Mixability is Bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, *13*, 1639–1663.

Werner, D. (2018). *Funktionalanalysis* 8th ed. Berlin: Springer.

Williamson, R. C., Vernet, E., Reid, M. D. (2016). Composite multiclass losses. *Journal of Machine Learning Research*, *17*, 1–52.

Wilson, L. J., Burrows, W. R., Lanzinger, A. (1999). A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review*, *127*, 956–970.

Zamo, M., Naveau, P. (2018). Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, *50*, 209–234.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.