Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Strictly Proper Scoring Rules, Prediction, and Estimation

Tilmann Gneiting & Adrian E Raftery

To cite this article: Tilmann Gneiting & Adrian E Raftery (2007) Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102:477, 359-378, DOI: 10.1198/016214506000001437

To link to this article: https://doi.org/10.1198/016214506000001437

Strictly Proper Scoring Rules, Prediction, and Estimation

Tilmann GNEITING and Adrian E. RAFTERY

Scoring rules assess the quality of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes. A scoring rule is proper if the forecaster maximizes the expected score for an observation drawn from the distribution F if he or she issues the probabilistic forecast F, rather than $G \neq F$. It is strictly proper if the maximum is unique. In prediction problems, proper scoring rules encourage the forecaster to make careful assessments and to be honest. In estimation problems, strictly proper scoring rules provide attractive loss and utility functions that can be tailored to the problem at hand. This article reviews and develops the theory of proper scoring rules on general probability spaces, and proposes and discusses examples thereof. Proper scoring rules derive from convex functions and relate to information measures, entropy functions, and Bregman divergences. In the case of categorical variables, we prove a rigorous version of the Savage representation. Examples of scoring rules for probabilistic forecasts in the form of predictive densities include the logarithmic, spherical, pseudospherical, and quadratic scores. The continuous ranked probability score applies to probabilistic forecasts that take the form of predictive cumulative distribution functions. It generalizes the absolute error and forms a special case of a new and very general type of score, the energy score. Like many other scoring rules, the energy score admits a kernel representation in terms of negative definite functions, with links to inequalities of Hoeffding type, in both univariate and multivariate settings. Proper scoring rules for quantile and interval forecasts are also discussed. We relate proper scoring rules to Bayes factors and to cross-validation, and propose a novel form of cross-validation known as random-fold cross-validation. A case study on probabilistic weather forecasts in the North American Pacific Northwest illustrates the importance of propriety. We note optimum score approaches to point and quantile estimation, and propose the intuitively appealing interval score as a utility function in interval estimation that addresses width as well as coverage.

KEY WORDS: Bayes factor; Bregman divergence; Brier score; Coherent; Continuous ranked probability score; Cross-validation; Entropy; Kernel score; Loss function; Minimum contrast estimation; Negative definite function; Prediction interval; Predictive distribution; Quantile forecast; Scoring rule; Skill score; Strictly proper; Utility function.

1. INTRODUCTION

One major purpose of statistical analysis is to make forecasts for the future and provide suitable measures of the uncertainty associated with them. Consequently, forecasts should be probabilistic in nature, taking the form of probability distributions over future quantities or events (Dawid 1984). Indeed, over the past two decades, probabilistic forecasting has become routine in such applications as weather and climate prediction (Palmer 2002; Gneiting and Raftery 2005), computational finance (Duffie and Pan 1997), and macroeconomic forecasting (Garratt, Lee, Pesaran, and Shin 2003; Granger 2006). In the statistical literature, advances in Markov chain Monte Carlo methodology (see, e.g., Besag, Green, Higdon, and Mengersen 1995) have led to explosive growth in the use of predictive distributions, mostly in the form of Monte Carlo samples from posterior predictive distributions of quantities of interest. In earlier work (Gneiting, Raftery, Balabdaoui, and Westveld 2003; Gneiting, Balabdaoui, and Raftery 2006), we contended that the goal of probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration. Calibration refers to the statistical consistency between the distributional forecasts and the observations, and is a joint property of the forecasts and the events or values that materialize. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only.

Scoring rules provide summary measures for the evaluation of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes. In terms of elicitation, the role of scoring rules is to encourage the assessor to make careful assessments and to be honest (Garthwaite, Kadane, and O'Hagan 2005). In terms of evaluation, scoring rules measure the quality of the probabilistic forecasts, reward probability assessors for forecasting jobs, and rank competing forecast procedures. Meteorologists refer to this broad task as *forecast verification*, and much of the underlying methodology has been developed by atmospheric scientists (Jolliffe and Stephenson 2003). In a Bayesian context, scores are frequently referred to as utilities, emphasizing the Bayesian principle of maximizing the expected utility of a predictive distribution (Bernardo and Smith 1994). We take scoring rules to be positively oriented rewards that a forecaster wishes to maximize. Specifically, if the forecaster quotes the predictive distribution P and the event x materializes, then his or her reward is $S(P, x)$. The function $S(P, \cdot)$ takes values in the real line $\mathbb{R}$ or in the extended real line $\overline{\mathbb{R}} = [-\infty, \infty]$, and we write $S(P, Q)$ for the expected value of $S(P, \cdot)$ under Q. Suppose, then, that the forecaster's best judgment is the distributional forecast Q. The forecaster has no incentive to predict any $P \neq Q$ and is encouraged to quote his or her true belief, $P = Q$, if $S(Q, Q) \geq S(P, Q)$ with equality if and only if $P = Q$. A scoring rule with this property is said to be *strictly proper*. If $S(Q, Q) \geq S(P, Q)$ for all P and Q, then the scoring rule is said to be *proper*. Propriety is essential in scientific and operational

Tilmann Gneiting is Associate Professor of Statistics (E-mail: *tilmann@stat.washington.edu*) and Adrian E. Raftery is Blumstein-Jordan Professor of Statistics and Sociology (E-mail: *raftery@u.washington.edu*), Department of Statistics, University of Washington, Seattle, WA 98195. This work was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under grant N00014-01-10745 and by the National Science Foundation under award 0134264. Part of Tilmann Gneiting's work was performed on sabbatical leave at the Soil Physics Group, Universität Bayreuth, 95440 Bayreuth, Germany. The authors thank Mark Albright, Veronica J. Berrocal, William M. Briggs, Andreas Buja, Ignacio Cascos, Claudia Czado, A. Philip Dawid, Werner Ehm, Thomas Gerds, Eric P. Grimit, Susanne Gschlößl, Eliezer Gurarie, Mark S. Handcock, Leonhard Held, Peter J. Huber, Nicholas A. Johnson, Ian T. Jolliffe, Hans Kuensch, Christian Lantuéjoul, Clifford F. Mass, Debashis Mondal, David B. Stephenson, Werner Stuetzle, Gabor J. Székely, Olivier Talagrand, Jon A. Wellner, Lawrence J. Wilson, Robert L. Winkler, and two anonymous reviewers for providing comments, preprints, references, and data.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Review Article
DOI 10.1198/016214506000001437

forecast evaluation; and we present a case study that provides a striking example of the potential issues that result from the use of intuitively appealing but improper scoring rules.

In estimation problems, strictly proper scoring rules provide attractive loss and utility functions that can be tailored to a scientific problem. To fix the idea, suppose that we wish to fit a parametric model P_θ based on a sample $X_1, \ldots, X_n$. To estimate θ, we might measure the goodness-of-fit by the mean score

$$\mathcal{S}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(P_\theta, X_i),$$

where S is a strictly proper scoring rule. If θ_0 denotes the true parameter value, then asymptotic arguments indicate that $\arg\max_\theta \mathcal{S}_n(\theta) \to \theta_0$ as $n \to \infty$. This suggests a general approach to estimation: Choose a strictly proper scoring rule that is tailored to the problem at hand and use $\hat{\theta}_n = \arg\max_\theta \mathcal{S}_n(\theta)$ as the *optimum score estimator* based on the scoring rule. Pfanzagl (1969) and Birgé and Massart (1993) studied this approach under the heading of *minimum contrast estimation*. Maximum likelihood estimation forms a special case of optimum score estimation, and optimum score estimation forms a special case of M-estimation (Huber 1964), in that the function to be optimized derives from a strictly proper scoring rule.

This article reviews and develops the theory of proper scoring rules on general probability spaces, proposes and discusses examples thereof, and presents case studies. The remainder of the article is organized as follows. In Section 2 we state a fundamental characterization theorem, review the links between proper scoring rules, information measures, entropy functions, and Bregman divergences, and introduce skill scores. In Section 3 we turn to scoring rules for categorical variables. We prove a rigorous version of the representation of Savage (1971) and relate to a more recent characterization of Schervish (1989) that applies to probability forecasts of a dichotomous event. Bremnes (2004, p. 346) noted that the literature on scoring rules for probabilistic forecasts of continuous variables is sparse. We address this issue in Section 4, where we discuss the spherical, pseudospherical, logarithmic, and quadratic scores. The *continuous ranked probability score*, which lately has attracted much attention, enjoys appealing properties and might serve as a standard score in evaluating probabilistic forecasts of real-valued variables. It forms a special case of a novel and very general type of scoring rule, the energy score. In Section 5 we introduce an even more general construction, giving rise to *kernel scores* based on negative definite functions and inequalities of Hoeffding type, with side results on expectation inequalities and positive definite functions. In Section 6 we study scoring rules for quantile and interval forecasts. We show that the class of proper scoring rules for quantile forecasts is larger than conjectured by Cervera and Muñoz (1996) and discuss the *interval score*, a scoring rule for prediction intervals that is proper and has intuitive appeal. In Section 7 we relate proper scoring rules to Bayes factors and to cross-validation, and propose a novel form of cross-validation known as random-fold cross-validation. In Section 8 we present a case study on the use of scoring rules in the evaluation of probabilistic weather forecasts. In Section 9 we turn to optimum score estimation. We discuss point, quantile, and interval estimation and propose using the interval score as a utility function that addresses width as well as coverage. We close the article with a discussion of avenues for future work in Section 10. Scoring rules show a superficial analogy to statistical depth functions, which we hint at in an Appendix.

2. CHARACTERIZATIONS OF PROPER SCORING RULES

In this section we introduce notation, provide characterizations of proper scoring rules, and relate them to convex functions, information measures, and Bregman divergences. The discussion here is more technical than that in the remainder of the article, and readers with more applied interests might skip ahead to Section 2.3, in which we discuss skill scores, without significant loss of continuity.

2.1 Proper Scoring Rules and Convex Functions

We consider probabilistic forecasts on a general sample space Ω. Let $\mathcal{A}$ be a σ-algebra of subsets of Ω, and let $\mathcal{P}$ be a convex class of probability measures on $(\Omega, \mathcal{A})$. A function defined on Ω and taking values in the extended real line, $\overline{\mathbb{R}} = [-\infty, \infty]$, is *$\mathcal{P}$-quasi-integrable* if it is measurable with respect to $\mathcal{A}$ and is quasi-integrable with respect to all $P \in \mathcal{P}$ (Bauer 2001, p. 64). A *probabilistic forecast* is any probability measure $P \in \mathcal{P}$. A *scoring rule* is any extended real-valued function $S: \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ such that $S(P, \cdot)$ is $\mathcal{P}$-quasi-integrable for all $P \in \mathcal{P}$. Thus if the forecast is P and ω materializes, the forecaster's reward is $S(P, \omega)$. We permit algebraic operations on the extended real line and deal with the respective integrals and expectations as described in section 2.1 of Mattner (1997) and section 3.1 of Grünwald and Dawid (2004). The scoring rules used in practice are mostly real-valued, but there are exceptions, such as the logarithmic rule (Good 1952), that allow for infinite scores.

We write

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

for the expected score under Q when the probabilistic forecast is P. The scoring rule S is *proper* relative to $\mathcal{P}$ if

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

It is *strictly proper* relative to $\mathcal{P}$ if (1) holds with equality if and only if $P = Q$, thereby encouraging honest quotes by the forecaster. If S is a proper scoring rule, $c > 0$ is a constant, and h is a $\mathcal{P}$-integrable function, then

$$S^*(P, \omega) = cS(P, \omega) + h(\omega) \tag{2}$$

is also a proper scoring rule. Similarly, if S is strictly proper, then S^* is strictly proper as well. Following Dawid (1998), we say that S and S^* are *equivalent*, and *strongly equivalent* if $c = 1$. The term *proper* was apparently coined by Winkler and Murphy (1968, p. 754), whereas the general idea dates back at least to Brier (1950) and Good (1952, p. 112). In a parametric context, and with respect to estimators, Lehmann and Casella (1998, p. 157) refer to the defining property in (1) as *risk unbiasedness*.

A function $G: \mathcal{P} \to \mathbb{R}$ is *convex* if

$$G((1-\lambda)P_0 + \lambda P_1) \leq (1-\lambda)G(P_0) + \lambda G(P_1) \quad \text{for all } \lambda \in (0, 1),\ P_0, P_1 \in \mathcal{P}. \tag{3}$$

It is *strictly convex* if (3) holds with equality if and only if $P_0 = P_1$. A function $G^*(P, \cdot) : \Omega \to \overline{\mathbb{R}}$ is a *subtangent* of G at the point $P \in \mathcal{P}$ if it is integrable with respect to P, quasi-integrable with respect to all $Q \in \mathcal{P}$, and

$$G(Q) \geq G(P) + \int G^*(P, \omega)\, \mathrm{d}(Q - P)(\omega) \tag{4}$$

for all $Q \in \mathcal{P}$. The following characterization theorem is more general and considerably simpler than previous results of McCarthy (1956) and Hendrickson and Buehler (1971).

Definition 1. A scoring rule $S : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ is *regular* relative to the class $\mathcal{P}$ if $S(P, Q)$ is real-valued for all $P, Q \in \mathcal{P}$, except possibly that $S(P, Q) = -\infty$ if $P \neq Q$.

Theorem 1. A regular scoring rule $S : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ is proper relative to the class $\mathcal{P}$ if and only if there exists a convex, real-valued function G on $\mathcal{P}$ such that

$$S(P, \omega) = G(P) - \int G^*(P, \omega)\, \mathrm{d}P(\omega) + G^*(P, \omega) \tag{5}$$

for $P \in \mathcal{P}$ and $\omega \in \Omega$, where $G^*(P, \cdot) : \Omega \to \overline{\mathbb{R}}$ is a subtangent of G at the point $P \in \mathcal{P}$. The statement holds with proper replaced by strictly proper, and convex replaced by strictly convex.

Proof. If the scoring rule S is of the stated form, then the subtangent inequality (4) implies the defining inequality (1), that is, propriety. Conversely, suppose that S is a regular proper scoring rule. Define $G : \mathcal{P} \to \mathbb{R}$ by $G(P) = S(P, P) = \sup_{Q \in \mathcal{P}} S(Q, P)$, which is the pointwise supremum over a class of convex functions and thus is convex on $\mathcal{P}$. Furthermore, the subtangent inequality (4) holds with $G^*(P, \omega) = S(P, \omega)$. This implies the representation (5) and proves the claim for propriety. By an argument of Hendrickson and Buehler (1971), strict inequality in (1) is equivalent to no subtangent of G at P being a subtangent of G at Q, for $P, Q \in \mathcal{P}$ and $P \neq Q$, which is equivalent to G being strictly convex on $\mathcal{P}$.

Expressed slightly differently, a regular scoring rule S is proper relative to the class $\mathcal{P}$ if and only if the expected score function $G(P) = S(P, P)$ is convex and $S(P, \omega)$ is a subtangent of G at the point P, for all $P \in \mathcal{P}$.

2.2 Information Measures, Bregman Divergences, and Decision Theory

Suppose that the scoring rule S is proper relative to the class $\mathcal{P}$. Following Grünwald and Dawid (2004) and Buja, Stuetzle, and Shen (2005), we call the expected score function

$$G(P) = \sup_{Q \in \mathcal{P}} S(Q, P), \qquad P \in \mathcal{P}, \tag{6}$$

the *information measure* or *generalized entropy function* associated with the scoring rule S. This is the maximally achievable utility; the term *entropy function* is used as well. If S is regular and proper, then we call

$$d(P, Q) = S(Q, Q) - S(P, Q), \qquad P, Q \in \mathcal{P}, \tag{7}$$

the associated *divergence function*. Note the order of the arguments, which differs from previous practice in that the true distribution, Q, is preceded by an alternative probabilistic forecast, P. The divergence function is nonnegative, and if S is strictly proper, then $d(P, Q)$ is strictly positive, unless $P = Q$. If the sample space is finite and the entropy function is sufficiently smooth, then the divergence function becomes the *Bregman divergence* (Bregman 1967), associated with the convex function G. Bregman divergences play major roles in optimization and have recently attracted the attention of the machine learning community (Collins, Schapire, and Singer 2002). The term *Bregman distance* is also used, even though $d(P, Q)$ is not necessarily the same as $d(Q, P)$.

An interesting problem is to find conditions under which a divergence function d is a *score divergence*, in the sense that it admits the representation (7) for a proper scoring rule S, and to describe principled ways of finding such a scoring rule. The landmark work by Savage (1971) provides a necessary condition on a symmetric divergence function d to be a score divergence: If P and Q are concentrated on the same two mutually exclusive events and identified with the respective probabilities, $p, q \in [0, 1]$, then $d(P, Q)$ reduces to a linear function of $(p - q)^2$. Dawid (1998) noted that if d is a score convergence, then $d(P, Q) - d(P', Q)$ is an affine function of Q for all $P, P' \in \mathcal{P}$, and proved a partial converse.

Friedman (1983) and Nau (1985) studied a looser type of relationship between proper scoring rules and distance measures on classes of probability distributions. They restricted attention to metrics (i.e., distance measures that are symmetric and satisfy the triangle inequality) and called a scoring rule S *effective* with respect to a metric d if

$$S(P_1, Q) \geq S(P_2, Q) \quad \Longleftrightarrow \quad d(P_1, Q) \leq d(P_2, Q).$$

Nau (1985) called a metric *co-effective* if there is a proper scoring rule that is effective with respect to it. His proposition 1 implies that the l_1, l_∞, and Hellinger distances on spaces of absolutely continuous probability measures are not co-effective.

Sections 3–5 provide numerous examples of proper scoring rules on general sample spaces, along with the associated entropy and divergence functions. For example, the logarithmic score is linked to Shannon entropy and Kullback–Leibler divergence. Dawid (1998, 2006), Grünwald and Dawid (2004), and Buja et al. (2005) have given further examples of proper scoring rules, entropy, and divergence functions and have elaborated on the connections to the Bregman divergence.

Proper scoring rules occur naturally in statistical decision problems (Dawid 1998). Given an outcome space and an action space, let $U(\omega, a)$ be the utility for outcome ω and action a, and let $\mathcal{P}$ be a convex class of probability measures on the outcome space. Let a_P denote the Bayes act for $P \in \mathcal{P}$. Then the scoring rule

$$S(P, \omega) = U(\omega, a_P)$$

is proper relative to the class $\mathcal{P}$. Indeed,

$$\begin{aligned} S(Q, Q) &= \int U(\omega, a_Q)\, \mathrm{d}Q(\omega) \\ &\geq \int U(\omega, a_P)\, \mathrm{d}Q(\omega) = S(P, Q), \end{aligned}$$

by the fact that the optimal Bayesian decision maximizes expected utility. Dawid (2006) has given details and discussed the generality of the construction.

2.3 Skill Scores

In practice, scores are aggregated, and competing forecast procedures are ranked by the average score,

$$\mathcal{S}_n = \frac{1}{n}\sum_{i=1}^{n} S(P_i, x_i),$$

over a fixed set of forecast situations. We give examples of this in case studies in Sections 6 and 8. Recommendations for choosing a scoring rule have been given by Winkler (1994, 1996), by Buja et al. (2005), and throughout this article.

Scores for competing forecast procedures are directly comparable if they refer to exactly the same set of forecast situations. If scores for distinct sets of situations are compared, then considerable care must be exercised to separate the confounding effects of intrinsic predictability and predictive performance. For instance, there is substantial spatial and temporal variability in the predictability of weather and climate elements (Langland et al. 1999; Campbell and Diebold 2005). Thus a score that is superior for a given location or season might be inferior for another, or vice versa. To address this issue, atmospheric scientists have put forth *skill scores* of the form

$$\mathcal{S}_n^{\text{skill}} = \frac{\mathcal{S}_n^{\text{fcst}} - \mathcal{S}_n^{\text{ref}}}{\mathcal{S}_n^{\text{opt}} - \mathcal{S}_n^{\text{ref}}}, \tag{8}$$

where $\mathcal{S}_n^{\text{fcst}}$ is the forecaster's score, $\mathcal{S}_n^{\text{opt}}$ refers to a hypothetical ideal or optimal forecast, and $\mathcal{S}_n^{\text{ref}}$ is the score for a reference strategy (Murphy 1973; Potts 2003, p. 27; Briggs and Ruppert 2005; Wilks 2006, p. 259). Skill scores are standardized in that (8) takes the value 1 for an optimal forecast, which is typically understood as a point measure in the event or value that materializes, and the value 0 for the reference forecast. Negative values of a skill score indicate forecasts that are of lesser quality than the reference. The reference forecast is typically a *climatological* forecast, that is, an estimate of the marginal distribution of the predictand. For example, a climatological probabilistic forecast for maximum temperature on Independence Day in Seattle, Washington might be a smoothed version of the local historic record of July 4 maximum temperatures. Climatological forecasts are independent of the forecast horizon; they are calibrated by construction, but often lack sharpness.

Unfortunately, skill scores of the form (8) are generally improper, even if the underlying scoring rule S is proper. Murphy (1973) studied hedging strategies in the case of the Brier skill score for probability forecasts of a dichotomous event. He showed that the Brier skill score is asymptotically proper, in the sense that the benefits of hedging become negligible as the number of independent forecasts grows. Similar arguments may apply to skill scores based on other proper scoring rules. Mason's (2004) claim of the propriety of the Brier skill score rests on unjustified approximations and generally is incorrect.

3. SCORING RULES FOR CATEGORICAL VARIABLES

We now review the representations of Savage (1971) and Schervish (1989) that characterize scoring rules for probabilistic forecasts of categorical and binary variables, and give examples of proper scoring rules.

3.1 Savage Representation

We consider probabilistic forecasts of a categorical variable. Thus, the sample space $\Omega = \{1, \ldots, m\}$ consists of a finite number m of mutually exclusive events, and a probabilistic forecast is a probability vector $(p_1, \ldots, p_m)$. Using the notation of Section 2, we consider the convex class $\mathcal{P} = \mathcal{P}_m$, where

$$\mathcal{P}_m = \{\mathbf{p} = (p_1, \ldots, p_m) : p_1, \ldots, p_m \geq 0, p_1 + \cdots + p_m = 1\}.$$

A scoring rule S can then be identified with a collection of m functions,

$$S(\cdot, i) : \mathcal{P}_m \to \overline{\mathbb{R}}, \qquad i = 1, \ldots, m.$$

In other words, if the forecaster quotes the probability vector $\mathbf{p}$ and the event i materializes, then his or her reward is $S(\mathbf{p}, i)$. Theorem 2 is a special case of Theorem 1 and provides a rigorous version of the Savage (1971) representation of proper scoring rules on finite sample spaces. Our contributions lie in the notion of regularity, the rigorous treatment, and the introduction of appropriate tools for convex analysis (Rockafellar 1970, sects. 23–25). Specifically, let $G : \mathcal{P}_m \to \mathbb{R}$ be a convex function. A vector $\mathbf{G}'(\mathbf{p}) = (G_1'(\mathbf{p}), \ldots, G_m'(\mathbf{p}))$ is a *subgradient* of G at the point $\mathbf{p} \in \mathcal{P}_m$ if

$$G(\mathbf{q}) \geq G(\mathbf{p}) + \langle \mathbf{G}'(\mathbf{p}), \mathbf{q} - \mathbf{p} \rangle \tag{9}$$

for all $\mathbf{q} \in \mathcal{P}_m$, where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product. If G is differentiable at an interior point $\mathbf{p} \in \mathcal{P}_m$, then $\mathbf{G}'(\mathbf{p})$ is unique and equals the gradient of G at $\mathbf{p}$. We assume that the components of $\mathbf{G}'(\mathbf{p})$ are real-valued, except that we permit $G_i'(\mathbf{p}) = -\infty$ if $p_i = 0$.

Definition 2. A scoring rule S for categorical forecasts is *regular* if $S(\cdot, i)$ is real-valued for $i = 1, \ldots, m$, except possibly that $S(\mathbf{p}, i) = -\infty$ if $p_i = 0$.

Regular scoring rules assign finite scores, except that a forecast might receive a score of $-\infty$ if an event claimed to be impossible is realized. The logarithmic scoring rule (Good 1952) provides a prominent example of this.

Theorem 2 (McCarthy, Savage). A regular scoring rule S for categorical forecasts is proper if and only if

$$S(\mathbf{p}, i) = G(\mathbf{p}) - \langle \mathbf{G}'(\mathbf{p}), \mathbf{p} \rangle + G_i'(\mathbf{p}) \quad \text{for } i = 1, \ldots, m, \tag{10}$$

where $G : \mathcal{P}_m \to \mathbb{R}$ is a convex function and $\mathbf{G}'(\mathbf{p})$ is a subgradient of G at the point $\mathbf{p}$, for all $\mathbf{p} \in \mathcal{P}_m$. The statement holds with proper replaced by strictly proper, and convex replaced by strictly convex.

Phrased slightly differently, a regular scoring rule S is proper if and only if the expected score function $G(\mathbf{p}) = S(\mathbf{p}, \mathbf{p})$ is convex on $\mathcal{P}_m$, and the vector with components $S(\mathbf{p}, i)$ for $i = 1, \ldots, m$ is a subgradient of G at the point $\mathbf{p}$, for all $\mathbf{p} \in \mathcal{P}_m$. In view of these results, every bounded convex function G on $\mathcal{P}_m$ generates a regular proper scoring rule. This function G becomes the expected score function, information measure, or entropy function (6) associated with the score. The divergence function (7) is the respective Bregman distance.

We now give a number of examples. The scoring rules in Examples 1–3 are strictly proper. The score in Example 4 is proper but not strictly proper.

Example 1 (Quadratic or Brier score). If $G(\mathbf{p}) = \sum_{j=1}^m p_j^2 - 1$, then (10) yields the quadratic score or Brier score,

$$S(\mathbf{p}, i) = -\sum_{j=1}^m (\delta_{ij} - p_j)^2 = 2p_i - \sum_{j=1}^m p_j^2 - 1,$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. The associated Bregman divergence is the squared Euclidean distance, $d(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^m (p_j - q_j)^2$. This well-known scoring rule was proposed by Brier (1950). Selten (1998) gave an axiomatic characterization.

Example 2 (Spherical score). Let $\alpha > 1$ and consider the generalized entropy function $G(\mathbf{p}) = (\sum_{j=1}^m p_j^\alpha)^{1/\alpha}$. This corresponds to the pseudospherical score

$$S(\mathbf{p}, i) = \frac{p_i^{\alpha-1}}{(\sum_{j=1}^m p_j^\alpha)^{(\alpha-1)/\alpha}},$$

which reduces to the traditional spherical score when $\alpha = 2$. The associated Bregman divergence is

$$d(\mathbf{p}, \mathbf{q}) = \left(\sum_{j=1}^m q_j^\alpha\right)^{1/\alpha} - \sum_{j=1}^m p_j q_j^{\alpha-1} \Big/ \left(\sum_{j=1}^m q_j^\alpha\right)^{(\alpha-1)/\alpha}.$$

Example 3 (Logarithmic score). Negative Shannon entropy, $G(\mathbf{p}) = \sum_{j=1}^m p_j \log p_j$, corresponds to the logarithmic score, $S(\mathbf{p}, i) = \log p_i$. The associated Bregman distance is the Kullback–Leibler divergence, $d(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^m q_j \log(q_j/p_j)$. [Note the order of the arguments in the definition (7) of the divergence function.] This scoring rule dates back at least to Good (1952). Information-theoretic perspectives and interpretations in terms of gambling returns have been given by Roulston and Smith (2002) and Daley and Vere-Jones (2004). Despite its popularity, the logarithmic score has been criticized for its unboundedness, with Selten (1998, p. 51) arguing that it entails value judgments that are unacceptable. Feuerverger and Rahman (1992) noted a connection to Neyman–Pearson theory and an ensuing optimality property of the logarithmic score.

Example 4 (Zero–one score). The zero–one scoring rule rewards a probabilistic forecast if the mode of the predictive distribution materializes. In case of multiple modes, the reward is reduced proportionally, that is,

$$S(\mathbf{p}, i) = \begin{cases} 1/\#M(\mathbf{p}) & \text{if } i \text{ belongs to } M(\mathbf{p}) \\ 0 & \text{otherwise,} \end{cases}$$

where $M(\mathbf{p}) = \{i : p_i = \max_{j=1,\ldots,m} p_j\}$ denotes the set of modes of $\mathbf{p}$. This is also known as the *misclassification loss*, and the meteorological literature uses the term *success rate* to denote case-averaged zero–one scores (see, e.g., Toth, Zhu, and Marchok 2001). The associated expected score or generalized entropy function (6) is $G(\mathbf{p}) = \max_{j=1,\ldots,m} p_j$, and the divergence function (7) becomes

$$d(\mathbf{p}, \mathbf{q}) = \max_{j=1,\ldots,m} q_j - \frac{\sum_{j \in M(\mathbf{p})} q_j}{\#M(\mathbf{p})}.$$

This does not define a Bregman divergence, because the entropy function is neither differentiable nor strictly convex.

The scoring rules in the foregoing examples are *symmetric*, in the sense that

$$S((p_1, \ldots, p_m), i) = S\big((p_{\pi_1}, \ldots, p_{\pi_m}), \pi_i\big) \tag{11}$$

for all $\mathbf{p} \in \mathcal{P}_m$, for all permutations π on m elements and for all events $i = 1, \ldots, m$. Winkler (1994, 1996) argued that symmetric rules do not always appropriately reward forecasting skill and called for asymmetric ones, particularly in situations in which skills scores traditionally have been used. Asymmetric proper scoring rules can be generated by applying Theorem 2 to convex functions G that are not invariant under coordinate permutation.

3.2 Schervish Representation

The classical case of a probability forecast for a dichotomous event suggests further discussion. We follow Dawid (1986) in considering the sample space $\Omega = \{1, 0\}$. A probabilistic forecast is a quoted probability $p \in [0, 1]$ for the event to occur. A scoring rule S can be identified with a pair of functions $S(\cdot, 1) : [0, 1] \to \overline{\mathbb{R}}$ and $S(\cdot, 0) : [0, 1] \to \overline{\mathbb{R}}$. Thus, $S(p, 1)$ is the forecaster's reward if he or she quotes p and the event materializes, and $S(p, 0)$ is the reward if he or she quotes p and the event does not materialize. Note the subtle change from the previous section, where we used the convex class $\mathcal{P}_2 = \{(p_1, p_2) \in \mathbb{R}^2 : p_1 \in [0, 1], p_2 = 1 - p_1\}$ in place of the unit interval, $\mathcal{P} = [0, 1]$, to represent probability measures on binary sample spaces.

A scoring rule for binary variables is *regular* if $S(\cdot, 1)$ and $S(\cdot, 0)$ are real-valued, except possibly that $S(0, 1) = -\infty$ or $S(1, 0) = -\infty$. A variant of Theorem 2 shows that every regular proper scoring rule is of the form

$$\begin{aligned} S(p, 1) &= G(p) + (1 - p)G'(p), \\ S(p, 0) &= G(p) - pG'(p), \end{aligned} \tag{12}$$

where $G : [0, 1] \to \mathbb{R}$ is a convex function and $G'(p)$ is a subgradient of G at the point $p \in [0, 1]$, in the sense that

$$G(q) \geq G(p) + G'(p)(q - p)$$

for all $q \in [0, 1]$. The statement holds with proper replaced by strictly proper, and convex replaced by strictly convex. The subgradient $G'(p)$ is real-valued, except that we permit $G'(0) = -\infty$ and $G'(1) = \infty$. The function G is the expected score function $G(p) = pS(p, 1) + (1 - p)S(p, 0)$, and if G is differentiable at an interior point $p \in (0, 1)$, then $G'(p)$ is unique and equals the derivative of G at p. Related but slightly less general results were given by Shuford, Albert, and Massengil (1966). Figure 1 provides a geometric interpretation.

The Savage representation (12) implies various interesting properties of regular proper scoring rules. For instance, we conclude from theorem 24.2 of Rockafellar (1970) that

$$S(p, 1) = \lim_{q \to 1} G(q) - \int_p^1 (G'(q) - G'(p))\, \mathrm{d}q \tag{13}$$

for $p \in (0, 1)$, and because $G'(p)$ is increasing, $S(p, 1)$ is increasing as well. Similarly, $S(p, 0)$ is decreasing, as would be intuitively expected. The statements hold with proper, increasing, and decreasing replaced by strictly proper, strictly increasing, and strictly decreasing. Alternative proofs of these and other results have been given by Schervish (1989, the app.).

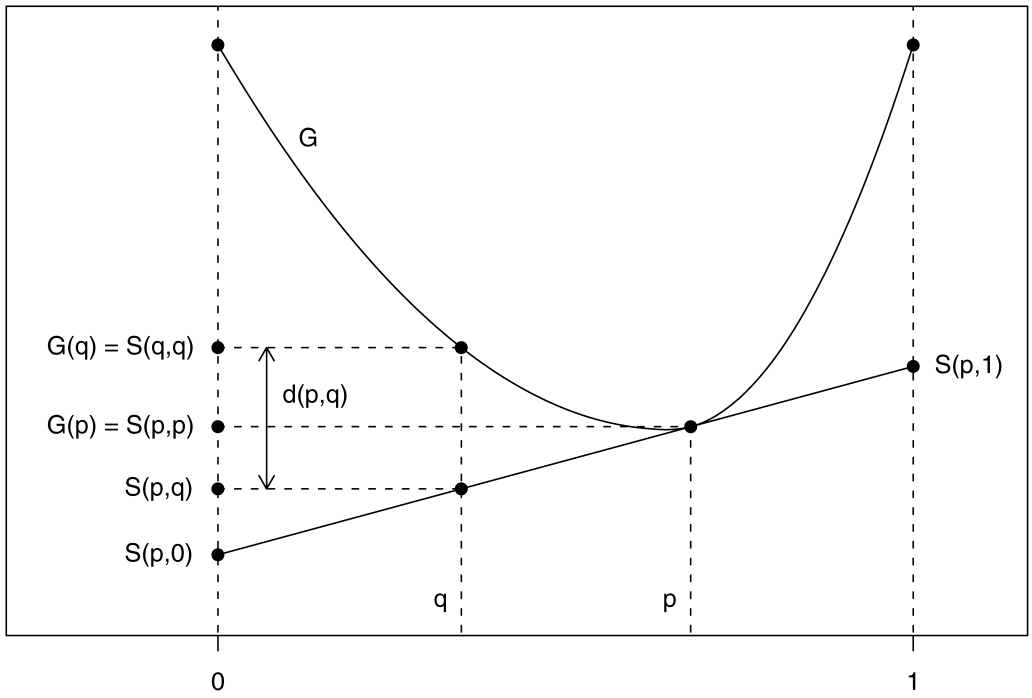

Figure 1. Schematic Illustration of the Relationships Between a Smooth Generalized Entropy Function G (solid convex curve) and the Associated Scoring Functions and Bregman Divergence. For any probability forecast $p \in [0, 1]$, the expected score $S(p, q) = qS(p, 1) + (1 - q)S(p, 0)$ equals the ordinate of the tangent to G at p [the solid line with slope $G'(p)$], when evaluated at $q \in [0, 1]$. In particular, the scores $S(p, 0) = G(p) - pG'(p)$ and $S(p, 1) = G(p) + (1 - p)G'(p)$ can be read off the tangent when evaluated at $q = 0$ and $q = 1$. The Bregman divergence $d(p, q) = S(q, q) - S(p, q)$ equals the difference between G and its tangent at p when evaluated at q. (For a similar interpretation see fig. 8 in Buja et al. 2005.)

Schervish (1989, p. 1861) suggested that his theorem 4.2 generalizes the Savage representation. Given Savage's (1971, p. 793) assessment of his representation (9.15) as "figurative," the claim can well be justified. However, in its rigorous form [eq. (12)], the Savage representation is perfectly general.

Hereinafter, we let $\mathbb{1}\{\cdot\}$ denote an indicator function that takes value 1 if the event in brackets is true and 0 otherwise.

Theorem 3 (Schervish). Suppose that S is a regular scoring rule. Then S is proper and such that $S(0, 1) = \lim_{p\to 0} S(p, 1)$, and $S(0, 0) = \lim_{p\to 0} S(p, 0)$, and both $S(p, 1)$ and $S(p, 0)$ are left continuous if and only if there exists a nonnegative measure ν on $(0, 1)$ such that

$$\begin{aligned} S(p, 1) &= S(1, 1) - \int (1 - c)\mathbb{1}\{p \le c\}\nu(\mathrm{d}c), \\ S(p, 0) &= S(0, 0) - \int c\mathbb{1}\{p > c\}\nu(\mathrm{d}c), \end{aligned} \tag{14}$$

for all $p \in [0, 1]$. The scoring rule is strictly proper if and only if ν assigns positive measure to every open interval.

Sketch of Proof. Suppose that S satisfies the assumptions of the theorem. To prove that $S(p, 1)$ is of the form (14), consider the representation (13), identify the increasing function $G'(p)$ with the left-continuous distribution function of a nonnegative measure ν on $(0, 1)$, and apply the partial integration formula. The proof of the representation for $S(p, 0)$ is analogous. For the proof of the converse, reverse the foregoing steps. The statement for strict propriety follows from well-known properties of convex functions.

A two-decision problem can be characterized by a cost–loss ratio $c \in (0, 1)$ that reflects the relative costs of the two possible types of inferior decision. The measure $\nu(\mathrm{d}c)$ in Schervish's representation (14) assigns relevance to distinct cost–loss ratios. This result also can be interpreted as a Choquet representation, in that every left-continuous bounded scoring rule is equivalent to a mixture of cost-weighted asymmetric zero–one scores,

$$S_c(p, 1) = (1 - c)\mathbb{1}\{p > c\}, \qquad S_c(p, 0) = c\mathbb{1}\{p \le c\}, \tag{15}$$

with a nonnegative mixing measure $\nu(\mathrm{d}c)$. Theorem 3 allows for unbounded scores, requiring a slightly more elaborate statement. Full equivalence to the Savage representation (12) can be achieved if the regularity conditions are relaxed (Schervish 1989; Buja et al. 2005).

Table 1 shows the mixing measure $\nu(\mathrm{d}c)$ for the quadratic or Brier score, the spherical score, the logarithmic score, and the asymmetric zero–one score. If the expected score function, G, is smooth, then $\nu(\mathrm{d}c)$ has Lebesgue density $G''(c)$ (Buja et al. 2005). For instance, the logarithmic score derives from Shannon entropy, $G(p) = p \log p + (1 - p) \log(1 - p)$, and corresponds to the infinite measure with Lebesgue density $(c(1 - c))^{-1}$.

Buja et al. (2005) introduced the beta family, a continuous two-parameter family of proper scoring rules that includes both symmetric and asymmetric members and derives from mixing measures of beta type.

Example 5 (Beta family). Let $\alpha, \beta > -1$ and consider the two-parameter family

$$S(p, 1) = -\int_p^1 c^{\alpha-1}(1 - c)^{\beta}\,\mathrm{d}c,$$

Table 1. Proper Scoring Rules for Probability Forecasts of a Dichotomous Event and the Respective Mixing Measure or Lebesgue Density in the Schervish Representation (14)

Scoring rule	*S(p, 1)*	*S(p, 0)*	*ν(dc)*
Brier	$-(1-p)^2$	$-p^2$	Uniform
Spherical	$p(1-2p+2p^2)^{-1/2}$	$(1-p)(1-2p+2p^2)^{-1/2}$	$(1-2c+2c^2)^{-3/2}$
Logarithmic	$\log p$	$\log(1-p)$	$(c(1-c))^{-1}$
Zero–one	$(1-c)\mathbb{1}\{p>c\}$	$c\,\mathbb{1}\{p\le c\}$	Point measure in c

$$S(p,0) = -\int_0^p c^{\alpha}(1-c)^{\beta-1}\,\mathrm{d}c,$$

which is of the form (14) for a mixing measure $\nu(\mathrm{d}c)$ with Lebesgue density $c^{\alpha-1}(1-c)^{\beta-1}$. This family includes the logarithmic score ($\alpha=\beta=0$), and versions of the Brier score ($\alpha=\beta=1$), and the zero–one score (15) with $c=\frac{1}{2}$ ($\alpha=\beta\to\infty$) as special or limiting cases. Asymmetric members arise when $\alpha\neq\beta$, with the scoring rule $S(p,1)=p-1$ and $S(p,0)=p+\log(1-p)$ being one such example ($\alpha=0,\beta=1$).

Winkler (1994) proposed a method for constructing asymmetric scoring rules from symmetric scoring rules. Specifically, if S is a symmetric proper scoring rule and $c\in(0,1)$, then

$$\begin{aligned} S^*(p,1) &= \frac{S(p,1)-S(c,1)}{T(c,p)},\\ S^*(p,0) &= \frac{S(p,0)-S(c,0)}{T(c,p)}, \end{aligned} \tag{16}$$

where $T(c,p)=S(0,0)-S(c,0)$ if $p\le c$ and $T(c,p)=S(1,1)-S(c,1)$ if $p>c$ is also a proper scoring rule, standardized in the sense that the expected score function attains a minimum value of 0 at $p=c$ and a maximum value of 1 at $p=0$ and $p=1$.

Example 6 (Winkler's score). Tetlock (2005) explored what constitutes good judgment in predicting future political and economic events, and looked at why experts are often wrong in their forecasts. In evaluating experts' predictions, he adjusted for the difficulty of the forecast task by using the special case of (16) that derives from the Brier score, that is,

$$\begin{aligned} S^*(p,1) &= \frac{(1-c)^2-(1-p)^2}{c^2\mathbb{1}\{p\le c\}+(1-c)^2\mathbb{1}\{p>c\}},\\ S^*(p,0) &= \frac{c^2-p^2}{c^2\mathbb{1}\{p\le c\}+(1-c)^2\mathbb{1}\{p>c\}}, \end{aligned} \tag{17}$$

with the value of $c\in(0,1)$ adapted to reflect a baseline probability. This was suggested by Winkler (1994, 1996) as an alternative to using skill scores.

Figure 2 shows the expected score or generalized entropy function, $G(p)$, and the scoring functions, $S(p,1)$ and $S(p,0)$, for the quadratic or Brier score and the logarithmic score (Table 1), the asymmetric zero–one score (15) with $c=.6$, and Winkler's standardized score (17) with $c=.2$.

4. SCORING RULES FOR CONTINUOUS VARIABLES

Bremnes (2004, p. 346) noted that the literature on scoring rules for probabilistic forecasts of continuous variables is sparse. We address this issue in the following.

4.1 Scoring Rules for Density Forecasts

Let μ be a σ-finite measure on the measurable space $(\Omega,\mathcal{A})$. For $\alpha>1$, let $\mathcal{L}_\alpha$ denote the class of probability measures on $(\Omega,\mathcal{A})$ that are absolutely continuous with respect to μ and have μ-density p such that

$$\|p\|_\alpha = \left(\int p(\omega)^\alpha \mu(\mathrm{d}\omega)\right)^{1/\alpha}$$

is finite. We identify a probabilistic forecast $P\in\mathcal{L}_\alpha$ with its μ-density, p, and call p a *predictive density* or *density forecast*. Predictive densities are defined only up to a set of μ-measure zero. Whenever appropriate, we follow Bernardo (1979, p. 689) and use the unique version defined by $p(\omega)=\lim_{\rho\to0}P(S_\rho(\omega))/\mu(S_\rho(\omega))$, where $S_\rho(\omega)$ is a sphere of radius ρ centered at ω.

We begin by discussing scoring rules that correspond to Examples 1, 2, and 3. The *quadratic score*,

$$\mathrm{QS}(p,\omega) = 2p(\omega) - \|p\|_2^2, \tag{18}$$

is strictly proper relative to the class $\mathcal{L}_2$. It has expected score or generalized entropy function $G(p)=\|p\|_2^2$, and the associated divergence function, $d(p,q)=\|p-q\|_2^2$, is symmetric. Good (1971) proposed the *pseudospherical score*,

$$\mathrm{PseudoS}(p,\omega) = p(\omega)^{\alpha-1}/\|p\|_\alpha^{\alpha-1},$$

that reduces to the *spherical score* when $\alpha=2$. He described original and generalized versions of the score—a distinction that in a measure-theoretic framework is obsolete. The pseudospherical score is strictly proper relative to the class $\mathcal{L}_\alpha$. The strict convexity of the associated entropy function, $G(p)=\|p\|_\alpha$, and the nonnegativity of the divergence function are straightforward consequences of the Hölder and Minkowski inequalities.

The *logarithmic score*,

$$\mathrm{LogS}(p,\omega) = \log p(\omega), \tag{19}$$

emerges as a limiting case ($\alpha\to1$) of the pseudospherical score when suitably scaled. This scoring rule was proposed by Good (1952) and has been widely used since then, under various names, including the *predictive deviance* (Knorr-Held and Rainer 2001) and the *ignorance score* (Roulston and Smith 2002). The logarithmic score is strictly proper relative to the class $\mathcal{L}_1$ of the probability measures dominated by μ. The associated expected score function or information measure is negative Shannon entropy, and the divergence function becomes the classical Kullback–Leibler divergence.

Bernardo (1979, p. 689) argued that "when assessing the worthiness of a scientist's final conclusions, only the probability he attaches to a small interval containing the true value

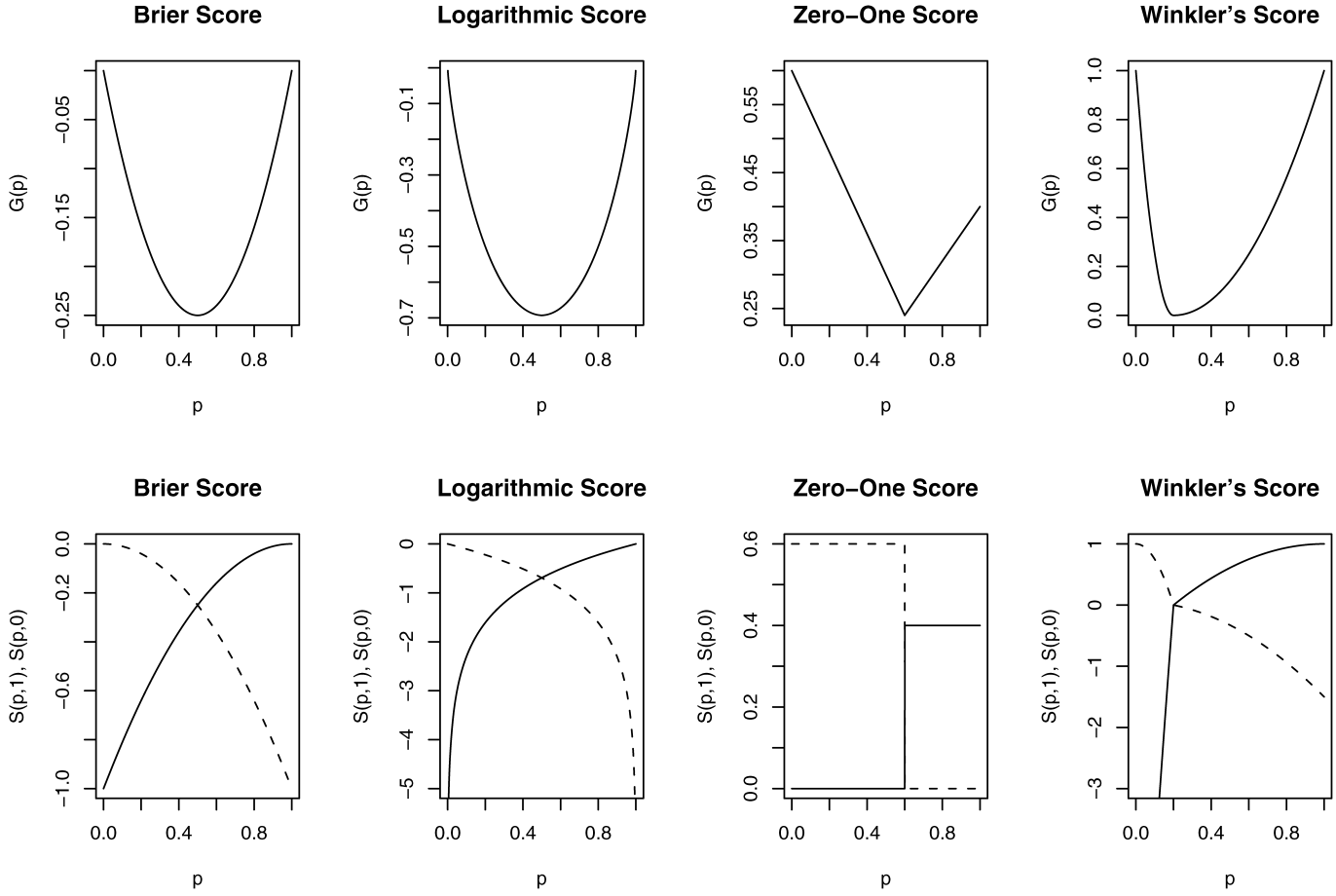

Figure 2. The Expected Score or Generalized Entropy Function G(p) (top row) and the Scoring Functions S(p, 1) (—) and S(p, 0) (- - -) (bottom row), for the Brier Score and Logarithmic Score (Table 1), the Asymmetric Zero–One Score (15) With c = .6 and Winkler's Standardized Score (17) With c = .2.

should be taken into account." This seems subject to debate, and atmospheric scientists have argued otherwise, putting forth scoring rules that are *sensitive to distance* (Epstein 1969; Staël von Holstein 1970). That said, Bernardo (1979) studied *local* scoring rules $S(p, \omega)$ that depend on the predictive density p only through its value at the event ω that materializes. Assuming regularity conditions, he showed that every proper local scoring rule is equivalent to the logarithmic score, in the sense of (2). Consequently, the *linear score*, $\mathrm{LinS}(p, \omega) = p(\omega)$, is not a proper scoring rule, despite its intuitive appeal. For instance, let φ and u denote the Lebesgue densities of a standard Gaussian distribution and the uniform distribution on $(-\epsilon, \epsilon)$. If $\epsilon < \sqrt{\log 2}$, then

$$\begin{aligned}\mathrm{LinS}(u, \varphi) &= \frac{1}{(2\pi)^{1/2}} \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} e^{-x^2/2}\, \mathrm{d}x \\ &> \frac{1}{2\pi^{1/2}} = \mathrm{LinS}(\varphi, \varphi),\end{aligned}$$

in violation of propriety. Essentially, the linear score encourages overprediction at the modes of an assessor's true predictive density (Winkler 1969). The probability score of Wilson, Burrows, and Lanzinger (1999) integrates the predictive density over a neighborhood of the observed, real-valued quantity. This resembles the linear score and is not a proper score either. Dawid (2006) constructed proper scoring rules from improper ones; an interesting question is whether this can be done for the probability score, similar to the way in which the proper quadratic score (18) derives from the linear score.

If Lebesgue densities on the real line are used to predict discrete observations, then the logarithmic score encourages the placement of artificially high density ordinates on the target values in question. This problem emerged in the Evaluating Predictive Uncertainty Challenge at a recent PASCAL Challenges Workshop (Kohonen and Suomela 2006; Quiñonero-Candela, Rasmussen, Sinz, Bousquet, and Schölkopf 2006). It disappears if scores expressed in terms of predictive cumulative distribution functions are used, or if the sample space is reduced to the target values in question.

4.2 Continuous Ranked Probability Score

The restriction to predictive densities is often impractical. For instance, probabilistic quantitative precipitation forecasts involve distributions with a point mass at zero (Krzysztofowicz and Sigrest 1999; Bremnes 2004), and predictive distributions are often expressed in terms of samples, possibly originating from Markov chain Monte Carlo. Thus it seems more compelling to define scoring rules directly in terms of predictive cumulative distribution functions. Furthermore, the aforementioned scores are not sensitive to distance, meaning that no credit is given for assigning high probabilities to values near but not identical to the one materializing.

To address this situation, let $\mathcal{P}$ consist of the Borel probability measures on $\mathbb{R}$. We identify a probabilistic forecast—a member of the class $\mathcal{P}$—with its cumulative distribution function F, and use standard notation for the elements of the sample space $\mathbb{R}$. The *continuous ranked probability score* (CRPS) is defined as

$$\mathrm{CRPS}(F, x) = -\int_{-\infty}^{\infty} (F(y) - \mathbb{1}\{y \geq x\})^2 \, \mathrm{d}y \tag{20}$$

and corresponds to the integral of the Brier scores for the associated binary probability forecasts at all real-valued thresholds (Matheson and Winkler 1976; Hersbach 2000).

Applications of the CRPS have been hampered by a lack of readily computable solutions to the integral in (20), and the use of numerical quadrature rules has been proposed instead (Staël von Holstein 1977; Unger 1985). However, the integral often can be evaluated in closed form. By lemma 2.2 of Baringhaus and Franz (2004) or identity (17) of Székely and Rizzo (2005),

$$\mathrm{CRPS}(F, x) = \frac{1}{2} E_F|X - X'| - E_F|X - x|, \tag{21}$$

where X and X' are independent copies of a random variable with distribution function F and finite first moment. If the predictive distribution is Gaussian with mean μ and variance σ^2, then it follows that

$$\mathrm{CRPS}(\mathcal{N}(\mu, \sigma^2), x) = \sigma\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma}\left(2\Phi\left(\frac{x-\mu}{\sigma}\right) - 1\right)\right],$$

where φ and Φ denote the probability density function and the cumulative distribution function of a standard Gaussian variable. If the predictive distribution takes the form of a sample of size n, then the right side of (20) can be evaluated in terms of the respective order statistics in a total of $\mathcal{O}(n \log n)$ operations (Hersbach 2000, sec. 4.b).

The CRPS is proper relative to the class $\mathcal{P}$ and strictly proper relative to the subclass $\mathcal{P}_1$ of the Borel probability measures that have finite first moment. The associated expected score function or information measure,

$$G(F) = -\int_{-\infty}^{\infty} F(y)(1 - F(y)) \, \mathrm{d}y = -\frac{1}{2} E_F|X - X'|,$$

coincides with the negative selectivity function (Matheron 1984), and the respective divergence function,

$$d(F, G) = \int_{-\infty}^{\infty} (F(y) - G(y))^2 \, \mathrm{d}y,$$

is symmetric and of the Cramér–von Mises type.

The CRPS lately has attracted renewed interest in the atmospheric sciences community (Hersbach 2000; Candille and Talagrand 2005; Gneiting, Raftery, Westveld, and Goldman 2005; Grimit, Gneiting, Berrocal, and Johnson 2006; Wilks 2006, pp. 302–303). It is typically used in negative orientation, say $\mathrm{CRPS}^*(F, x) = -\mathrm{CRPS}(F, x)$. The representation (21) then can be written as

$$\mathrm{CRPS}^*(F, x) = E_F|X - x| - \frac{1}{2} E_F|X - X'|,$$

which sheds new light on the score. In negative orientation, the CRPS can be reported in the same unit as the observations, and it generalizes the absolute error to which it reduces if F is a deterministic forecast—that is, a point measure. Thus the CRPS provides a direct way to compare deterministic and probabilistic forecasts.

4.3 Energy Score

We introduce a generalization of the CRPS that draws on Székely's (2003) statistical energy perspective. Let $\mathcal{P}_\beta$, $\beta \in (0, 2)$, denote the class of the Borel probability measures P on $\mathbb{R}^m$ that are such that $E_P\|\mathbf{X}\|^\beta$ is finite, where $\|\cdot\|$ denotes the Euclidean norm. We define the *energy score*,

$$\mathrm{ES}(P, \mathbf{x}) = \frac{1}{2} E_P\|\mathbf{X} - \mathbf{X}'\|^\beta - E_P\|\mathbf{X} - \mathbf{x}\|^\beta, \tag{22}$$

where $\mathbf{X}$ and $\mathbf{X}'$ are independent copies of a random vector with distribution $P \in \mathcal{P}_\beta$. This generalizes the CRPS, to which (22) reduces when $\beta = 1$ and $m = 1$, by allowing for an index $\beta \in (0, 2)$ and applying to distributional forecasts of a vector-valued quantity in $\mathbb{R}^m$. Theorem 1 of Székely (2003) shows that the energy score is strictly proper relative to the class $\mathcal{P}_\beta$. [For a different and more general argument, see Sec. 5.1.] In the limiting case $\beta = 2$, the energy score (22) reduces to the negative squared error,

$$\mathrm{ES}(P, \mathbf{x}) = -\|\boldsymbol{\mu}_p - \mathbf{x}\|^2, \tag{23}$$

where $\boldsymbol{\mu}_P$ denotes the mean vector of P. This scoring rule is regular and proper, but not strictly proper, relative to the class $\mathcal{P}_2$.

The energy score with index $\beta \in (0, 2)$ applies to all Borel probability measures on $\mathbb{R}^m$, by defining

$$\mathrm{ES}(P, \mathbf{x}) = -\frac{\beta 2^{\beta-2}\Gamma(\frac{m}{2} + \frac{\beta}{2})}{\pi^{m/2}\Gamma(1 - \frac{\beta}{2})} \int_{\mathbb{R}^m} \frac{|\phi_P(\mathbf{y}) - e^{i\langle \mathbf{x}, \mathbf{y}\rangle}|^2}{\|\mathbf{y}\|^{m+\beta}} \, \mathrm{d}\mathbf{y}, \tag{24}$$

where ϕ_P denotes the characteristic function of P. If P belongs to $\mathcal{P}_\beta$, then theorem 1 of Székely (2003) implies the equality of the right sides in (22) and (24). Essentially, the score computes a weighted distance between the characteristic function of P and the characteristic function of the point measure at the value that materializes.

4.4 Scoring Rules That Depend on First and Second Moments Only

An interesting question is that for proper scoring rules that apply to the Borel probability measures on $\mathbb{R}^m$ and depend on the predictive distribution, P, only through its mean vector, $\boldsymbol{\mu}_P$, and dispersion or covariance matrix, $\boldsymbol{\Sigma}_P$. Dawid (1998) and Dawid and Sebastiani (1999) studied proper scoring rules of this type. A particularly appealing example is the scoring rule

$$S(P, \mathbf{x}) = -\log\det\boldsymbol{\Sigma}_P - (\mathbf{x} - \boldsymbol{\mu}_P)'\boldsymbol{\Sigma}_P^{-1}(\mathbf{x} - \boldsymbol{\mu}_P), \tag{25}$$

which is linked to the generalized entropy function

$$G(P) = -\log\det\boldsymbol{\Sigma}_P - m,$$

and to the divergence function

$$d(P, Q) = \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) - \log\det(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) + (\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)'\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q) - m.$$

[Note the order of the arguments in the definition (7) of the divergence function.] This scoring rule is proper but not strictly proper relative to the class $\mathcal{P}_2$ of the Borel probability measures P for which $E_P\|\mathbf{X}\|^2$ is finite. It is strictly proper relative to any convex class of probability measures characterized by the first two moments, such as the Gaussian measures, for which (25) is equivalent to the logarithmic score (19). For other examples of scoring rules that depend on $\boldsymbol{\mu}_P$ and $\boldsymbol{\Sigma}_P$ only, see (23) and the right column of table 1 of Dawid and Sebastiani (1999).

The predictive model choice criterion of Laud and Ibrahim (1995) and Gelfand and Ghosh (1998) has lately attracted the attention of the statistical community. Suppose that we fit a predictive model to observed, real-valued data $x_1, \ldots, x_n$. The predictive model choice criterion (PMCC) assesses the model fit through the quantity

$$\text{PMCC} = \sum_{i=1}^{n}(x_i - \mu_i)^2 + \sum_{i=1}^{n}\sigma_i^2,$$

where μ_i and σ_i^2 denote the expected value and the variance of a replicate variable X_i, given the model and the observations. Within the framework of scoring rules, the PMCC corresponds to the positively oriented score

$$S(P, x) = -(x - \mu_P)^2 - \sigma_P^2, \tag{26}$$

where P has mean μ_P and variance σ_P^2. The scoring rule (26) depends on the predictive distribution through its first two moments only, but it is improper; if the forecaster's true belief is P and if he or she wishes to maximize the expected score, then he or she will quote the point measure at μ_P—that is, a deterministic forecast—rather than the predictive distribution P. This suggests that the predictive model choice criterion should be replaced by a criterion based on the scoring rule (25), which reduces to

$$S(P, x) = -\left(\frac{x - \mu_P}{\sigma_P}\right)^2 - \log\sigma_P^2 \tag{27}$$

in the case in which $m = 1$ and the observations are real-valued.

5. KERNEL SCORES, NEGATIVE AND POSITIVE DEFINITE FUNCTIONS, AND INEQUALITIES OF HOEFFDING TYPE

In this section we use negative definite functions to construct proper scoring rules and present expectation inequalities that are of independent interest.

5.1 Kernel Scores

Let Ω be a nonempty set. A real-valued function g on $\Omega \times \Omega$ is said to be a *negative definite kernel* if it is symmetric in its arguments and $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j g(x_i, x_j) \le 0$ for all positive integers n, all $a_1, \ldots, a_n \in \mathbb{R}$ that sum to 0, and all $x_1, \ldots, x_n \in \Omega$. Numerous examples of negative definite kernels have been given by Berg, Christensen, and Ressel (1984) and the references cited therein.

We now give the key result of this section, which generalizes a kernel construction of Eaton (1982, p. 335). The term *kernel score* was coined by Dawid (2006).

Theorem 4. Let Ω be a Hausdorff space and let g be a nonnegative, continuous negative definite kernel on $\Omega \times \Omega$. For a Borel probability measure P on Ω, let X and X' be independent random variables with distribution P. Then the scoring rule

$$S(P, x) = \frac{1}{2}E_P g(X, X') - E_P g(X, x) \tag{28}$$

is proper relative to the class of the Borel probability measures P on Ω for which the expectation $E_P g(X, X')$ is finite.

Proof. Let P and Q be Borel probability measures on Ω, and suppose that X, X' and Y, Y' are independent random variates with distribution P and Q. We need to show that

$$-\frac{1}{2}E_Q g(Y, Y') \ge \frac{1}{2}E_P g(X, X') - E_{P,Q} g(X, Y). \tag{29}$$

If the expectation $E_{P,Q} g(X, Y)$ is infinite, then the inequality is trivially satisfied; if it is finite, then theorem 2.1 of Berg et al. (1984, p. 235) implies (29).

Next we give examples of scoring rules that admit a kernel representation. In each case, we equip the sample space with the standard topology. Note that evaluating the kernel scores is straightforward if P is discrete and has only a moderate number of atoms.

Example 7 (Quadratic or Brier score). Let $\Omega = \{1, 0\}$ and suppose that $g(0, 0) = g(1, 1) = 0$ and $g(0, 1) = g(1, 0) = 1$. Then (28) recovers the quadratic or Brier score.

Example 8 (CRPS). If $\Omega = \mathbb{R}$ and $g(x, x') = |x - x'|$ for $x, x' \in \mathbb{R}$ in Theorem 4, we obtain the CRPS (21).

Example 9 (Energy score). If $\Omega = \mathbb{R}^m$, $\beta \in (0, 2)$, and $g(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^\beta$ for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, where $\|\cdot\|$ denotes the Euclidean norm, then (28) recovers the energy score (22).

Example 10 (CRPS for circular variables). We let $\Omega = \mathbb{S}$ denote the circle and write $\alpha(\theta, \theta')$ for the angular distance between two points $\theta, \theta' \in \mathbb{S}$. Let P be a Borel probability measure on $\mathbb{S}$, and let Θ and Θ' be independent random variates with distribution P. By theorem 1 of Gneiting (1998), angular distance is a negative definite kernel. Thus,

$$S(P, \theta) = \frac{1}{2}E_P\alpha(\Theta, \Theta') - E_P\alpha(\Theta, \theta) \tag{30}$$

defines a proper scoring rule relative to the class of the Borel probability measures on the circle. Grimit et al. (2006) introduced (30) as an analog of the CRPS (21) that applies to directional variables, and used Fourier analytic tools to prove the propriety of the score.

We turn to a far-reaching generalization of the energy score. For $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and $\alpha \in (0, \infty]$, define the vector norm $\|\mathbf{x}\|_\alpha = (\sum_{i=1}^{m} |x_i|^\alpha)^{1/\alpha}$ if $\alpha \in (0, \infty)$ and $\|\mathbf{x}\|_\alpha = \max_{1 \le i \le m} |x_i|$ if $\alpha = \infty$. Schoenberg's theorem (Berg et al. 1984, p. 74) and a strand of literature culminating in the work of Koldobskiĭ (1992) and Zastavnyi (1993) imply that if $\alpha \in (0, \infty]$ and $\beta > 0$, then the kernel

$$g(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_\alpha^\beta, \qquad \mathbf{x}, \mathbf{x}' \in \mathbb{R}^m,$$

is negative definite if and only if the following holds.

Assumption 1. Suppose that (a) $m = 1$, $\alpha \in (0, \infty]$, and $\beta \in (0, 2]$; (b) $m \geq 2$, $\alpha \in (0, 2]$, and $\beta \in (0, \alpha]$; or (c) $m = 2$, $\alpha \in (2, \infty]$, and $\beta \in (0, 1]$.

Example 11 (Non-Euclidean energy score). Under Assumption 1, the scoring rule

$$S(P, \mathbf{x}) = \frac{1}{2} E_P \|\mathbf{X} - \mathbf{X}'\|_\alpha^\beta - E_P \|\mathbf{X} - \mathbf{x}\|_\alpha^\beta$$

is proper relative to the class of the Borel probability measures P on $\mathbb{R}^m$ for which the expectation $E_P \|\mathbf{X} - \mathbf{X}'\|_\alpha^\beta$ is finite. If $m = 1$ or $\alpha = 2$, then we recover the energy score; if $m \geq 2$ and $\alpha \neq 2$, then we obtain non-Euclidean analogs. Mattner (1997, sec. 5.2) showed that if $\alpha \geq 1$, then $E_{P,Q} \|\mathbf{X} - \mathbf{Y}\|_\alpha^\beta$ is finite if and only if $E_P \|\mathbf{X}\|_\alpha^\beta$ and $E_Q \|\mathbf{Y}\|_\alpha^\beta$ are finite. In particular, if $\alpha \geq 1$, then $E_P \|\mathbf{X} - \mathbf{X}'\|_\alpha^\beta$ is finite if and only if $E_P \|\mathbf{X}\|_\alpha^\beta$ is finite.

The following result sharpens Theorem 4 in the crucial case of Euclidean sample spaces and spherically symmetric negative definite functions. Recall that a function η on $(0, \infty)$ is said to be *completely monotone* if it has derivatives $\eta^{(k)}$ of all orders and $(-1)^k \eta^{(k)}(t) \geq 0$ for all nonnegative integers k and all $t > 0$.

Theorem 5. Let ψ be a continuous function on $[0, \infty)$ with $-\psi'$ completely monotone and not constant. For a Borel probability measure P on $\mathbb{R}^m$, let $\mathbf{X}$ and $\mathbf{X}'$ be independent random vectors with distribution P. Then the scoring rule

$$S(P, \mathbf{x}) = \frac{1}{2} E_P \psi(\|\mathbf{X} - \mathbf{X}'\|_2^2) - E_P \psi(\|\mathbf{X} - \mathbf{x}\|_2^2)$$

is strictly proper relative to the class of the Borel probability measures P on $\mathbb{R}^m$ for which $E_P \psi(\|\mathbf{X} - \mathbf{X}'\|_2^2)$ is finite.

The proof of this result is immediate from theorem 2.2 of Mattner (1997). In particular, if $\psi(t) = t^{\beta/2}$ for $\beta \in (0, 2)$, then Theorem 5 ensures the strict propriety of the energy score relative to the class of the Borel probability measures P on $\mathbb{R}^m$ for which $E_P \|\mathbf{X}\|_2^\beta$ is finite.

5.2 Inequalities of Hoeffding Type and Positive Definite Kernels

A number of side results seem to be of independent interest, even though they are easy consequences of previous work. Briefly, if the expectations $E_P g(X, X')$ and $E_P g(Y, Y')$ are finite, then (29) can be written as a Hoeffding-type inequality,

$$2E_{P,Q} g(X, Y) - E_P g(X, X') - E_Q g(Y, Y') \geq 0. \tag{31}$$

Theorem 1 of Székely and Rizzo (2005) provides a nearly identical result and a converse: If g is not negative definite, then there are counterexamples to (31), and the respective scoring rule is improper. Furthermore, if Ω is a group and the negative definite function g satisfies $g(x, x') = g(-x, -x')$ for $x, x' \in \Omega$, then a special case of (31) can be stated as

$$E_P g(X, -X') \geq E_P g(X, X'). \tag{32}$$

In particular, if $\Omega = \mathbb{R}^m$ and Assumption 1 holds, then inequalities (31) and (32) apply and reduce to

$$2E\|\mathbf{X} - \mathbf{Y}\|_\alpha^\beta - E\|\mathbf{X} - \mathbf{X}'\|_\alpha^\beta - E\|\mathbf{Y} - \mathbf{Y}'\|_\alpha^\beta \geq 0 \tag{33}$$

and

$$E\|\mathbf{X} - \mathbf{X}'\|_\alpha^\beta \leq E\|\mathbf{X} + \mathbf{X}'\|_\alpha^\beta, \tag{34}$$

thereby generalizing results of Buja, Logan, Reeds, and Shepp (1994), Székely (2003), and Baringhaus and Franz (2004).

In the foregoing case in which Ω is a group and g satisfies $g(x, x') = g(-x, -x')$ for $x, x' \in \Omega$, the argument leading to theorem 2.3 of Buja et al. (1994) and theorem 4 of Ma (2003) implies that

$$h(x, x') = g(x, -x') - g(x, x'), \qquad x, x' \in \Omega, \tag{35}$$

is a *positive definite kernel*, in the sense that h is symmetric in its arguments and $\sum_{i=1}^n \sum_{j=1}^n a_i a_j h(x_i, x_j) \geq 0$ for all positive integers n, all $a_1, \ldots, a_n \in \mathbb{R}$, and all $x_1, \ldots, x_n \in \Omega$. Specifically, under Assumption 1,

$$h(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} + \mathbf{x}'\|_\alpha^\beta - \|\mathbf{x} - \mathbf{x}'\|_\alpha^\beta, \qquad \mathbf{x}, \mathbf{x}' \in \mathbb{R}^m, \tag{36}$$

is a positive definite kernel, a result that extends and completes the aforementioned theorem of Buja et al. (1994).

5.3 Constructions With Complex-Valued Kernels

With suitable modifications, the foregoing results allow for complex-valued kernels. A complex-valued function h on $\Omega \times \Omega$ is said to be a *positive definite kernel* if it is Hermitian, that is, $h(x, x') = \overline{h(x', x)}$ for $x, x' \in \Omega$, and $\sum_{i=1}^n \sum_{j=1}^n c_i \bar{c}_j h(x_i, x_j) \geq 0$ for all positive integers n, all $c_1, \ldots, c_n \in \mathbb{C}$, and all $x_1, \ldots, x_n \in \Omega$. The general idea (Dawid 1998, 2006) is that if h is continuous and positive definite, then

$$S(P, x) = E_P h(X, x) + E_P h(x, X) - E_P h(X, X') \tag{37}$$

defines a proper scoring rule. If h is positive definite, then $g = -h$ is negative definite; thus, if h is real-valued and sufficiently regular, then the scoring rules (37) and (28) are equivalent.

In the next example, we discuss scoring rules for Borel probability measures and observations on Euclidean spaces. However, the representation (37) allows for the construction of proper scoring rules in more general settings, such as probabilistic forecasts of structured data, including strings, sequences, graphs, and sets, based on positive definite kernels defined on such structures (Hofmann, Schölkopf, and Smola 2005).

Example 12. Let $\Omega = \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^m$, and consider the positive definite kernel $h(\mathbf{x}, \mathbf{x}') = e^{i\langle \mathbf{x} - \mathbf{x}', \mathbf{y} \rangle} - 1$, where $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$. Then (37) reduces to

$$S(P, \mathbf{x}) = -\left|\phi_P(\mathbf{y}) - e^{i\langle \mathbf{x}, \mathbf{y} \rangle}\right|^2, \tag{38}$$

that is, the negative squared distance between the characteristic function of the predictive distribution, ϕ_P, and the characteristic function of the point measures in the value that materializes, evaluated at $\mathbf{y} \in \mathbb{R}^m$. If we integrate with respect to a nonnegative measure $\mu(\mathrm{d}\mathbf{y})$, then the scoring rule (38) generalizes to

$$S(P, \mathbf{x}) = -\int_{\mathbb{R}^m} \left|\phi_P(\mathbf{y}) - e^{i\langle \mathbf{x}, \mathbf{y} \rangle}\right|^2 \mu(\mathrm{d}\mathbf{y}). \tag{39}$$

If the measure μ is finite and assigns positive mass to all intervals, then this scoring rule is strictly proper relative to the class of the Borel probability measures on $\mathbb{R}^m$. Eaton, Giovagnoli, and Sebastiani (1996) used the associated divergence function

to define metrics for probability measures. If μ is the infinite measure with Lebesgue density $\|\mathbf{y}\|^{-m-\beta}$, where $\beta \in (0, 2)$, then the scoring rule (39) is equivalent to the Euclidean energy score (24).

6. SCORING RULES FOR QUANTILE AND INTERVAL FORECASTS

Occasionally, full predictive distributions are difficult to specify, and the forecaster might quote predictive quantiles, such as value at risk in financial applications (Duffie and Pan 1997) or prediction intervals (Christoffersen 1998) only.

6.1 Proper Scoring Rules for Quantiles

We consider probabilistic forecasts of a continuous quantity that take the form of predictive quantiles. Specifically, suppose that the quantiles at the levels $\alpha_1, \ldots, \alpha_k \in (0, 1)$ are sought. If the forecaster quotes quantiles $r_1, \ldots, r_k$ and x materializes, then he or she will be rewarded by the score $S(r_1, \ldots, r_k; x)$. We define

$$S(r_1, \ldots, r_k; P) = \int S(r_1, \ldots, r_k; x)\, \mathrm{d}P(x)$$

as the expected score under the probability measure P when the forecaster quotes the quantiles $r_1, \ldots, r_k$. To avoid technical complications, we suppose that P belongs to the convex class $\mathcal{P}$ of Borel probability measures on $\mathbb{R}$ that have finite moments of all orders and whose distribution function is strictly increasing on $\mathbb{R}$. For $P \in \mathcal{P}$, let $q_1, \ldots, q_k$ denote the true P-quantiles at levels $\alpha_1, \ldots, \alpha_k$. Following Cervera and Muñoz (1996), we say that a scoring rule S is *proper* if

$$S(q_1, \ldots, q_k; P) \geq S(r_1, \ldots, r_k; P)$$

for all real numbers $r_1, \ldots, r_k$ and for all probability measures $P \in \mathcal{P}$. If S is proper, then the forecaster who wishes to maximize the expected score is encouraged to be honest and to volunteer his or her true beliefs.

To avoid technical overhead, we tacitly assume $\mathcal{P}$-integrability whenever appropriate. Essentially, we require that the functions $s(x)$ and $h(x)$ in (40) and (42) be $\mathcal{P}$-measurable and grow at most polynomially in x. Theorem 6 addresses the prediction of a single quantile; Corollary 1 turns to the general case.

Theorem 6. If s is nondecreasing and h is arbitrary, then the scoring rule

$$S(r; x) = \alpha s(r) + (s(x) - s(r))\mathbb{1}\{x \leq r\} + h(x) \tag{40}$$

is proper for predicting the quantile at level $\alpha \in (0, 1)$.

Proof. Let q be the unique α-quantile of the probability measure $P \in \mathcal{P}$. We identify P with the associated distribution function so that $P(q) = \alpha$. If $r < q$, then

$$\begin{aligned} &S(q; P) - S(r; P) \\ &\quad = \int_{(r,q)} s(x)\, \mathrm{d}P(x) + s(r)P(r) - \alpha s(r) \\ &\quad \geq s(r)(P(q) - P(r)) + s(r)P(r) - \alpha s(r) \\ &\quad = 0, \end{aligned}$$

as desired. If $r > q$, then an analogous argument applies.

If $s(x) = x$ and $h(x) = -\alpha x$, then we obtain the scoring rule

$$S(r; x) = (x - r)(\mathbb{1}\{x \leq r\} - \alpha), \tag{41}$$

which has been proposed by Koenker and Machado (1999), Taylor (1999), Giacomini and Komunjer (2005), Theis (2005, p. 232), and Friederichs and Hense (2006) for measuring in-sample goodness of fit and out-of-sample forecast performance in meteorological and financial applications. In negative orientation, the econometric literature refers to the scoring rule (41) as the *tick* or *check* loss function.

Corollary 1. If s_i is nondecreasing for $i = 1, \ldots, k$ and h is arbitrary, then the scoring rule

$$\begin{aligned} &S(r_1, \ldots, r_k; x) \\ &\qquad = \sum_{i=1}^{k} \big[\alpha_i s_i(r_i) + (s_i(x) - s_i(r_i))\mathbb{1}\{x \leq r_i\}\big] + h(x) \end{aligned} \tag{42}$$

is proper for predicting the quantiles at levels $\alpha_1, \ldots, \alpha_k \in (0, 1)$.

Cervera and Muñoz (1996, pp. 515 and 519) proved Corollary 1 in the special case in which each s_i is linear. They asked whether the resulting rules are the only proper ones for quantiles. Our results give a negative answer; that is, the class of proper scoring rules for quantiles is considerably larger than anticipated by Cervera and Muñoz. We do not know whether or not (40) and (42) provide the general form of proper scoring rules for quantiles.

6.2 Interval Score

Interval forecasts form a crucial special case of quantile prediction. We consider the classical case of the central $(1 - \alpha) \times 100\%$ prediction interval, with lower and upper endpoints that are the predictive quantiles at level $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$. We denote a scoring rule for the associated interval forecast by $S_\alpha(l, u; x)$, where l and u represent for the quoted $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles. Thus, if the forecaster quotes the $(1-\alpha) \times 100\%$ central prediction interval $[l, u]$ and x materializes, then his or her score will be $S_\alpha(l, u; x)$. Putting $\alpha_1 = \frac{\alpha}{2}$, $\alpha_2 = 1 - \frac{\alpha}{2}$, $s_1(x) = s_2(x) = 2\frac{x}{\alpha}$, and $h(x) = -2\frac{x}{\alpha}$ in (42), and reversing the sign of the scoring rule, yields the negatively oriented *interval score*,

$$\begin{aligned} &S_\alpha^{\mathrm{int}}(l, u; x) \\ &\qquad = (u - l) + \frac{2}{\alpha}(l - x)\mathbb{1}\{x < l\} + \frac{2}{\alpha}(x - u)\mathbb{1}\{x > u\}. \end{aligned} \tag{43}$$

This scoring rule has intuitive appeal and can be traced back to Dunsmore (1968), Winkler (1972), and Winkler and Murphy (1979). The forecaster is rewarded for narrow prediction intervals, and he or she incurs a penalty, the size of which depends on α, if the observation misses the interval. In the case $\alpha = \frac{1}{2}$, Hamill and Wilks (1995, p. 622) used a scoring rule that is equivalent to the interval score. They noted that "a strategy for gaming [...] was not obvious," thereby conjecturing propriety, which is confirmed by the foregoing. We anticipate novel applications, particularly for the evaluation of volatility forecasts in computational finance.

6.3 Case Study: Interval Forecasts for a Conditionally Heteroscedastic Process

This section illustrates the use of the interval score in a time series context. Kabaila (1999) called for rigorous ways of specifying prediction intervals for conditionally heteroscedastic processes and proposed a relevance criterion in terms of conditional coverage and width dependence. We contend that the notion of proper scoring rules provides an alternative and possibly simpler, more general, and more rigorous paradigm. The prediction intervals that we deem appropriate derive from the true conditional distribution, as implied by the data-generating mechanism, and optimize the expected value of all proper scoring rules.

To fix the idea, consider the stationary bilinear process $\{X_t : t \in \mathbb{Z}\}$ defined by

$$X_{t+1} = \frac{1}{2}X_t + \frac{1}{2}X_t\epsilon_t + \epsilon_t, \tag{44}$$

where the ϵ_t's are independent standard Gaussian random variates. Kabaila and He (2001) studied central one-step-ahead prediction intervals at the 95% level. The process is Markovian, and the conditional distribution of X_{t+1} given $X_t, X_{t-1}, \ldots$ is Gaussian with mean $\frac{1}{2}X_t$ and variance $(1 + \frac{1}{2}X_t)^2$, thereby suggesting the prediction interval

$$I = \left[\frac{1}{2}X_t - c\left|1 + \frac{1}{2}X_t\right|, \frac{1}{2}X_t + c\left|1 + \frac{1}{2}X_t\right|\right], \tag{45}$$

where $c = \Phi^{-1}(.975)$. This interval satisfies the relevance property of Kabaila (1999), and Kabaila and He (2001) adopted I as the standard prediction interval. We agree with this choice, but we prefer the aforementioned more direct justification; the prediction interval I is the standard interval because its lower and upper endpoints are the 2.5% and 97.5% percentiles of the true conditional distribution function. Kabaila and He considered two alternative prediction intervals,

$$J = [F^{-1}(.025), F^{-1}(.975)], \tag{46}$$

where F denotes the unconditional, stationary distribution function of X_t, and

$$K = \left[\frac{1}{2}X_t - \gamma\left(\left|1 + \frac{1}{2}X_t\right|\right), \frac{1}{2}X_t + \gamma\left(\left|1 + \frac{1}{2}X_t\right|\right)\right], \tag{47}$$

where $\gamma(y) = (2(\log 7.36 - \log y))^{1/2}y$ for $y \le 7.36$ and $\gamma(y) = 0$ otherwise. This choice minimizes the expected width of the prediction interval under the constraint of nominal coverage. However, the interval forecast K seems misguided, in that it collapses to a point forecast when the conditional predictive variance is highest.

We generated a sample path $\{X_t : t = 1, \ldots, 100{,}001\}$ from the bilinear process (44) and considered sequential one-step-ahead interval forecasts for X_{t+1}, where $t = 1, \ldots, 100{,}000$. Table 2 summarizes the results of this experiment. The interval forecasts I, J, and K all showed close to nominal coverage, with the prediction interval K being sharpest on average. Nevertheless, the classical prediction interval I performed best in terms of the interval score.

Table 2. Comparison of One-Step-Ahead 95% Interval Forecasts for the Stationary Bilinear Process (44)

Interval forecast		Empirical coverage	Average width	Average interval score
I	(45)	95.01%	4.00	4.77
J	(46)	95.08%	5.45	8.04
K	(47)	94.98%	3.79	5.32

NOTE: The table shows the empirical coverage, the average width, and the average value of the negatively oriented interval score (43) for the prediction intervals I, J, and K in 100,000 sequential forecasts in a sample path of length 100,001. See text for details.

6.4 Scoring Rules for Distributional Forecasts

Specifying a predictive cumulative distribution function is equivalent to specifying all predictive quantiles; thus we can build scoring rules for predictive distributions from scoring rules for quantiles. Matheson and Winkler (1976) and Cervera and Muñoz (1996) suggested ways of doing this. Specifically, if S_α denotes a proper scoring rule for the quantile at level α and ν is a Borel measure on $(0, 1)$, then the scoring rule

$$S(F, x) = \int_0^1 S_\alpha(F^{-1}(\alpha); x)\nu(\mathrm{d}\alpha) \tag{48}$$

is proper, subject to regularity and integrability constraints.

Similarly, we can build scoring rules for predictive distributions from scoring rules for binary probability forecasts. If S denotes a proper scoring rule for probability forecasts and ν is a Borel measure on $\mathbb{R}$, then the scoring rule

$$S(F, x) = \int_{-\infty}^{\infty} S(F(y), \mathbb{1}\{x \le y\})\nu(\mathrm{d}y) \tag{49}$$

is proper, subject to integrability constraints (Matheson and Winkler 1976; Gerds 2002). The CRPS (20) corresponds to the special case in (49) in which S is the quadratic or Brier score and ν is the Lebesgue measure. If S is the Brier score and ν is a sum of point measures, then the ranked probability score (Epstein 1969) emerges.

The construction carries over to multivariate settings. If $\mathcal{P}$ denotes the class of the Borel probability measures on $\mathbb{R}^m$, then we identify a probabilistic forecast $P \in \mathcal{P}$ with its cumulative distribution function F. A multivariate analog of the CRPS can be defined as

$$CRPS(F, \mathbf{x}) = -\int_{\mathbb{R}^m} (F(\mathbf{y}) - \mathbb{1}\{\mathbf{x} \le \mathbf{y}\})^2\nu(\mathrm{d}\mathbf{y}).$$

This is a weighted integral of the Brier scores at all m-variate thresholds. The Borel measure ν can be chosen to encourage the forecaster to concentrate his or her efforts on the important ones. If ν is a finite measure that dominates the Lebesgue measure, then this scoring rule is strictly proper relative to the class $\mathcal{P}$.

7. SCORING RULES, BAYES FACTORS, AND RANDOM–FOLD CROSS–VALIDATION

We now relate proper scoring rules to Bayes factors and to cross-validation and propose a novel form of cross-validation: random-fold cross-validation.

7.1 Logarithmic Score and Bayes Factors

Probabilistic forecasting rules are often generated by probabilistic models, and the standard Bayesian approach to comparing probabilistic models is by Bayes factors. Suppose that we have a sample $\mathbf{X} = (X_1, \ldots, X_n)$ of values to be forecast. Suppose also that we have two forecasting rules, based on probabilistic models H_1 and H_2. So far in this article we have concentrated on the situation where the forecasting rule is completely specified before any of the X_i's are observed; that is, there are no parameters to be estimated from the data being forecast. In that situation, the *Bayes factor* for H_1 against H_2 is

$$B = \frac{P(\mathbf{X}|H_1)}{P(\mathbf{X}|H_2)}, \tag{50}$$

where $P(\mathbf{X}|H_k) = \prod_{i=1}^{n} P(X_i|H_k)$ for $k = 1, 2$ (Jeffreys 1939; Kass and Raftery 1995).

Thus, if the logarithmic score is used, then the log Bayes factor is the difference of the scores for the two models,

$$\log B = \mathrm{LogS}(H_1, \mathbf{X}) - \mathrm{LogS}(H_2, \mathbf{X}). \tag{51}$$

This was pointed out by Good (1952), who called the log Bayes factor the *weight of evidence*. It establishes two connections: (1) the Bayes factor is equivalent to the logarithmic score in this no-parameter case, and (2) the Bayes factor applies more generally than merely to the comparison of parametric probabilistic models, but also to the comparison of probabilistic forecasting rules of any kind.

So far in this article we have taken probabilistic forecasts to be fully specified, but often they are specified only up to unknown parameters estimated from the data. Now suppose that the forecasting rules considered are specified only up to unknown parameters, θ_k for H_k, to be estimated from the data. Then the Bayes factor is still given by (50), but now $P(\mathbf{X}|H_k)$ is the *integrated likelihood*,

$$P(\mathbf{X}|H_k) = \int p(\mathbf{X}|\theta_k, H_k) p(\theta_k|H_k)\, \mathrm{d}\theta_k,$$

where $p(\mathbf{X}|\theta_k, H_k)$ is the (usual) likelihood under model H_k, and $p(\theta_k|H_k)$ is the prior distribution of the parameter θ_k.

Dawid (1984) showed that when the data come in a particular order, such as time order, the integrated likelihood can be reformulated in predictive terms,

$$P(\mathbf{X}|H_k) = \prod_{t=1}^{n} P(X_t|\mathbf{X}^{t-1}, H_k), \tag{52}$$

where $\mathbf{X}^{t-1} = \{X_1, \ldots, X_{t-1}\}$ if $t \geq 1$, X^0 is the empty set and $P(X_t|\mathbf{X}^{t-1}, H_k)$ is the predictive distribution of X_t given the past values under H_k, namely

$$P(X_t|\mathbf{X}^{t-1}, H_k) = \int p(X_t|\theta_k, H_k) P(\theta_k|\mathbf{X}^{t-1}, H_k)\, \mathrm{d}\theta_k,$$

with $P(\theta_k|\mathbf{X}^{t-1}, H_k)$ the posterior distribution of θ_k given the past observations $\mathbf{X}^{t-1}$.

We let $S_{k,B} = \log P(\mathbf{X}|H_k)$ denote the log-integrated likelihood, viewed now as a scoring rule. To view it as a scoring rule it helps to rewrite it as

$$S_{k,B} = \sum_{t=1}^{n} \log P(X_t|\mathbf{X}^{t-1}, H_k). \tag{53}$$

Dawid (1984) showed that $S_{k,B}$ is asymptotically equivalent to the plug-in maximum likelihood prequential score

$$S_{k,D} = \sum_{t=1}^{n} \log P(X_t|\mathbf{X}^{t-1}, \hat{\theta}_k^{t-1}), \tag{54}$$

where $\hat{\theta}_k^{t-1}$ is the maximum likelihood estimator (MLE) of θ_k based on the past observations, $\mathbf{X}^{t-1}$, in the sense that $S_{k,D}/S_{k,B} \to 1$ as $n \to \infty$. Initial terms for which $\hat{\theta}_k^{t-1}$ is possibly undefined can be ignored. Dawid also showed that $S_{k,B}$ is asymptotically equivalent to the Bayes information criterion (BIC) score,

$$S_{k,\mathrm{BIC}} = \sum_{t=1}^{n} \log P(X_t|\mathbf{X}^{t-1}, \hat{\theta}_k^{n}) - \frac{d_k}{2} \log n,$$

where $d_k = \dim(\theta_k)$, in the same sense, namely $S_{k,\mathrm{BIC}}/S_{k,B} \to 1$ as $n \to \infty$. This justifies using the BIC for comparing forecasting rules, extending the previous justification of Schwarz (1978), which related only to comparing models.

These results have two limitations, however. First, they assume that the data come in a particular order. Second, they use only the logarithmic score, not other scores that might be more appropriate for the task at hand. We now briefly consider how these limitations might be addressed.

7.2 Scoring Rules and Random-Fold Cross-Validation

Suppose now that the data are unordered. We can replace (53) by

$$S_{k,B}^{*} = \sum_{t=1}^{n} E_D\big[\log p\big(X_t|\mathbf{X}^{(D)}, H_k\big)\big], \tag{55}$$

where D is a random sample from $\{1, \ldots, t-1, t+1, \ldots, n\}$, the size of which is a random variable with a discrete uniform distribution on $\{0, 1, \ldots, n-1\}$. Dawid's results imply that this is asymptotically equivalent to the plug-in maximum likelihood version,

$$S_{k,D}^{*} = \sum_{t=1}^{n} E_D\big[\log p\big(X_t|\mathbf{X}^{(D)}, \hat{\theta}_k^{(D)}, H_k\big)\big], \tag{56}$$

where $\hat{\theta}_k^{(D)}$ is the MLE of θ_k based on $\mathbf{X}^{(D)}$. Terms for which the size of D is small and $\hat{\theta}_k^{(D)}$ is possibly undefined can be ignored.

The formulations (55) and (56) may be useful because they turn a score that was a sum of nonidentically distributed terms into one that is a sum of identically distributed exchangeable terms. This opens the possibility of evaluating $S_{k,B}^{*}$ or $S_{k,D}^{*}$ by Monte Carlo, which would be a form of cross-validation. In this cross-validation, the amount of data left out would be random rather than fixed, leading us to call it *random-fold cross-validation*. Smyth (2000) used the log-likelihood as the criterion function in cross-validation, as here, calling the resulting method cross-validated likelihood, but used a fixed hold-out sample size. This general approach can be traced back at least to Geisser and Eddy (1979). One issue in cross-validation generally is how much data to leave out; different choices lead to different versions of cross-validation, such as leave-one-out,

10-fold, and so on. Considering versions of cross-validation in the context of scoring rules may shed some light on this issue.

We have seen by (51) that when there are no parameters being estimated, the Bayes factor is equivalent to the difference in the logarithmic score. Thus we could replace the logarithmic score by another proper score, and the difference in scores could be viewed as a kind of predictive Bayes factor with a different type of score. In $S_{k,B}$, $S_{k,D}$, $S_{k,\mathrm{BIC}}$, $S^*_{k,B}$, and $S^*_{k,D}$, we could replace the terms in the sums (each of which has the form of a logarithmic score) by another proper scoring rule, such as the CRPS, and we conjecture that similar asymptotic equivalences would remain valid.

8. CASE STUDY: PROBABILISTIC FORECASTS OF SEA–LEVEL PRESSURE OVER THE NORTH AMERICAN PACIFIC NORTHWEST

Our goals in this case study are to illustrate the use and the properties of scoring rules and to demonstrate the importance of propriety.

8.1 Probabilistic Weather Forecasting Using Ensembles

Operational probabilistic weather forecasts are based on *ensemble prediction systems*. Ensemble systems typically generate a set of perturbations of the best estimate of the current state of the atmosphere, run each of them forward in time using a numerical weather prediction model, and use the resulting set of forecasts as a sample from the predictive distribution of future weather quantities (Palmer 2002; Gneiting and Raftery 2005).

Grimit and Mass (2002) described the University of Washington ensemble prediction system over the Pacific Northwest, which covers Oregon, Washington, British Columbia, and parts of the Pacific Ocean. This is a five-member ensemble comprising distinct runs of the MM5 numerical weather prediction model with initial conditions taken from distinct national and international weather centers. We consider 48-hour-ahead forecasts of sea-level pressure in January–June 2000, the same period as that on which the work of Grimit and Mass was based. The unit used is the millibar (mb). Our analysis builds on a verification data base of 16,015 records scattered over the North American Pacific Northwest and the aforementioned 6-month period. Each record consists of the five ensemble member forecasts and the associated verifying observation. The root mean squared error of the ensemble mean forecast was 3.30 mb, and the square root of the average variance of the five-member forecast ensemble was 2.13 mb, resulting in a ratio of $r_0 = 1.55$.

This underdispersive behavior—that is, observed errors that tend to be larger on average than suggested by the ensemble spread—is typical of ensemble systems and seems unavoidable, given that ensembles capture only some of the sources of uncertainty (Raftery, Gneiting, Balabdaoui, and Polakowski 2005). Thus, to obtain calibrated predictive distributions, it seems necessary to carry out some form of statistical postprocessing. One natural approach is to take the predictive distribution for sea-level pressure at any given site as Gaussian, centered at the ensemble mean forecast, and with predictive standard deviation equal to r times the standard deviation of the forecast ensemble. Density forecasts of this type were proposed by Déqué, Royer, and Stroe (1994) and Wilks (2002). Following Wilks, we refer to r as an *inflation factor*.

8.2 Evaluation of Density Forecasts

In the aforementioned approach, the predictive density is Gaussian, say $\varphi_{\mu,r\sigma}$; its mean, μ, is the ensemble mean forecast, and its standard deviation, $r\sigma$, is the product of the inflation factor, r, and the standard deviation of the five-member forecast ensemble, σ. We considered various scoring rules S and computed the average score,

$$s(r) = \frac{1}{16{,}015} \sum_{i=1}^{16{,}015} S\bigl(\varphi_{\mu_i, r\sigma_i}, x_i\bigr), \qquad r > 0, \tag{57}$$

as a function of the inflation factor r. The index i refers to the ith record in the verification database, and x_i denotes the value that materialized. Given the underdispersive character of the ensemble system, we expect $s(r)$ to be maximized at some $r > 1$, possibly near the observed ratio, $r_0 = 1.55$, of the root mean squared error of the ensemble mean forecast over the square root of the average ensemble variance.

We computed the mean score (57) for inflation factors $r \in (0, 5)$ and for the quadratic score (QS), spherical score (SphS), logarithmic score (LogS), CRPS, linear score (LinS), and probability score (PS), as defined in Section 4. Briefly, if p denotes the predictive density and x denotes the observed value, then

$$\mathrm{QS}(p, x) = 2p(x) - \int_{-\infty}^{\infty} p(y)^2 \, \mathrm{d}y,$$

$$\mathrm{SphS}(p, x) = p(x) \Big/ \left(\int_{-\infty}^{\infty} p(y)^2 \, \mathrm{d}y \right)^{1/2},$$

$$\mathrm{LogS}(p, x) = \log p(x),$$

$$\mathrm{CRPS}(p, x) = \frac{1}{2} E_p |X - X'| - E_p |X - x|,$$

$$\mathrm{LinS}(p, x) = p(x),$$

and

$$\mathrm{PS}(p, x) = \int_{x-1}^{x+1} p(y) \, \mathrm{d}y.$$

Figure 3 and Table 3 summarize the results of this experiment. The scores shown in the figure are linearly transformed, so that the graphs can be compared side by side, and the transformations are listed in the rightmost column of the table. In the case of the quadratic score, for instance, we plotted 40 times the value in (57) plus 6. Clearly, transformed and original scores are equivalent in the sense of (2). The quadratic score, spherical score, logarithmic score and CRPS were maximized at values of $r > 1$, thereby confirming the underdispersive character of

Table 3. Probabilistic Forecasts of Sea-Level Pressure Over the North American Pacific Northwest in January–July 2000

Score	*Argmax$_r$ $s(r)$ in eq. (57)*	*Linear transformation plotted in Figure 3*
Quadratic score (QS)	2.18	$40s + 6$
Spherical score (SphS)	1.84	$108s - 22$
Logarithmic score (LogS)	2.41	$s + 13$
CRPS	1.62	$10s + 8$
Linear score (LinS)	.05	$105s - 5$
Probability score (PS)	.02	$60s - 5$

NOTE: The predictive density is Gaussian, centered at the ensemble mean forecast, and with predictive standard deviation equal to r times the standard deviation of the forecast ensemble.

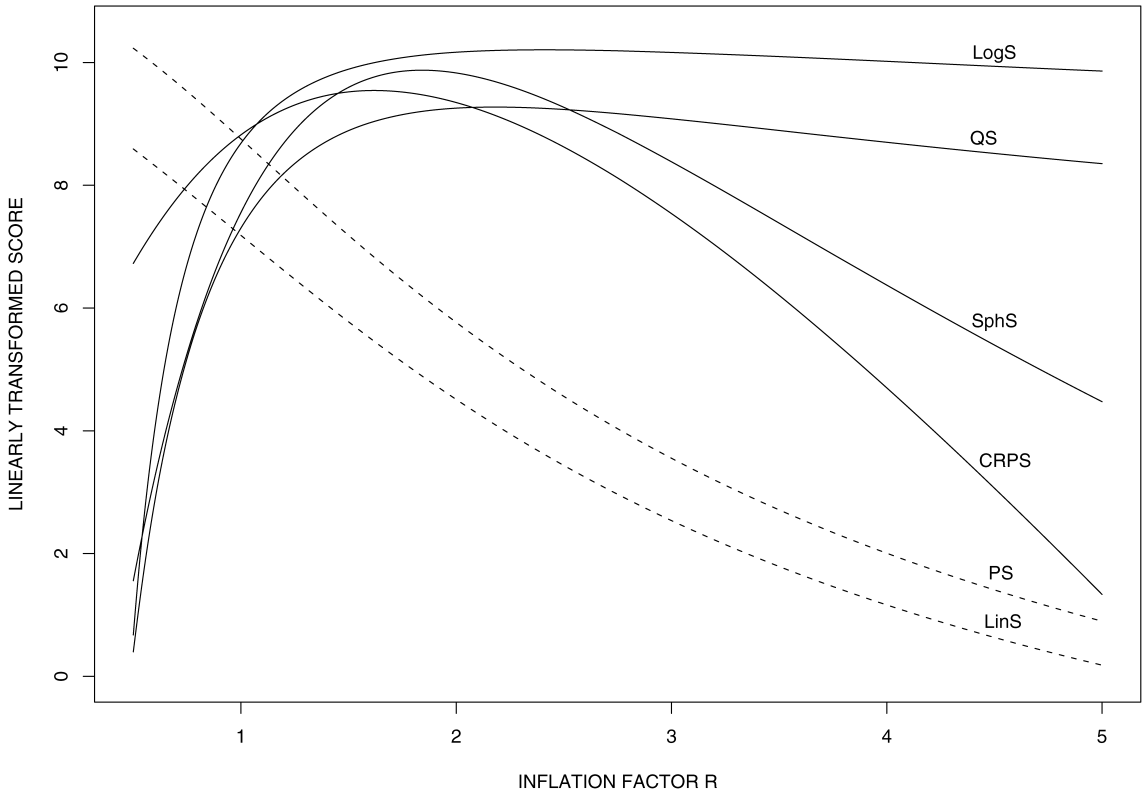

Figure 3. Probabilistic Forecasts of Sea-Level Pressure Over the North American Pacific Northwest in January–July 2000. The scores are shown as a function of the inflation factor r, where the predictive density is Gaussian, centered at the ensemble mean forecast, and with predictive standard deviation equal to r times the standard deviation of the forecast ensemble. The scores were subject to linear transformations as detailed in Table 3.

the ensemble. These scores are proper. The linear and probability scores were maximized at $r = .05$ and $r = .02$, thereby suggesting ignorable forecast uncertainty and essentially deterministic forecasts. The latter two scores have intuitive appeal, and the probability score has been used to assess forecast ensembles (Wilson et al. 1999). However, they are improper, and their use may result in misguided scientific inferences, as in this experiment. A similar comment applies to the predictive model choice criterion given in Section 4.4.

It is interesting to observe that the logarithmic score gave the highest maximizing value of r. The logarithmic score is strictly proper but involves a harsh penalty for low probability events and thus is highly sensitive to extreme cases. Our verification database includes a number of low-spread cases for which the ensemble variance implodes. The logarithmic score penalizes the resulting predictions unless the inflation factor r is large. Weigend and Shi (2000, p. 382) noted similar concerns and considered the use of trimmed means when computing the logarithmic score. In our experience, the CRPS is less sensitive to extreme cases or outliers and provides an attractive alternative.

8.3 Evaluation of Interval Forecasts

The aforementioned predictive densities also provide interval forecasts. We considered the central $(1-\alpha)\times 100\%$ prediction interval where $\alpha = .50$ and $\alpha = .10$. The associated lower and upper prediction bounds l_i and u_i are the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of a Gaussian distribution with mean μ_i and standard deviation $r\sigma_i$, as described earlier. We assessed the interval forecasts in their dependence on the inflation factor r in two ways: by computing the empirical coverage of the prediction intervals and by computing

$$s_\alpha(r) = \frac{1}{16{,}015}\sum_{i=1}^{16{,}015} S_\alpha^{\text{int}}(l_i, u_i; x_i), \qquad r > 0, \tag{58}$$

where S_α^{int} denotes the negatively oriented interval score (43). This scoring rule assesses both calibration and sharpness, by rewarding narrow prediction intervals and penalizing intervals missed by the observation. Figure 4(a) shows the empirical coverage of the interval forecasts. Clearly, the coverage increases with r. For $\alpha = .50$ and $\alpha = .10$, the nominal coverage was obtained at $r = 1.78$ and $r = 2.11$, which confirms the underdispersive character of the ensemble. Figure 4(b) shows the interval score (58) as a function of the inflation factor r. For $\alpha = .50$ and $\alpha = .10$, the score was optimized at $r = 1.56$ and $r = 1.72$.

9. OPTIMUM SCORE ESTIMATION

Strictly proper scoring rules also are of interest in estimation problems, where they provide attractive loss and utility functions that can be adapted to the problem at hand.

9.1 Point Estimation

We return to the generic estimation problem described in Section 1. Suppose that we wish to fit a parametric model P_θ based on a sample $X_1, \ldots, X_n$ of identically distributed observations. To estimate θ, we can measure the goodness of fit by

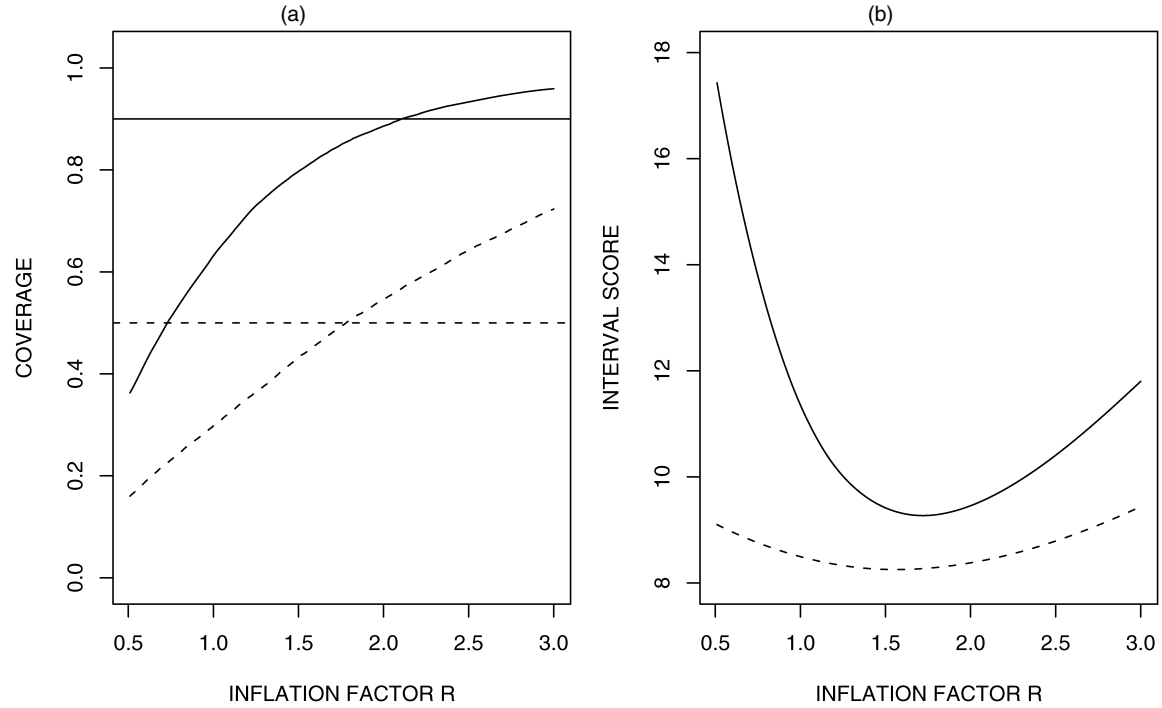

Figure 4. Interval Forecasts of Sea-Level Pressure Over the North American Pacific Northwest in January–July 2000. (a) Nominal and actual coverage and (b) the negatively oriented interval score (58), for the 50% central prediction interval ($\alpha = .50$, - - -) and the 90% central prediction interval ($\alpha = .10$, —; score scaled by a factor of .60). The predictive density is Gaussian, centered at the ensemble mean forecast, and with predictive standard deviation equal to r times the standard deviation of the forecast ensemble.

the mean score

$$\mathcal{S}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(P_\theta, X_i),$$

where S is a scoring rule that is strictly proper relative to a convex class of probability measures that contains the parametric model. If θ_0 denotes the true parameter value, then asymptotic arguments indicate that

$$\arg\max_\theta \mathcal{S}_n(\theta) \to \theta_0 \quad \text{as } n \to \infty. \tag{59}$$

This suggests a general approach to estimation: Choose a strictly proper scoring rule tailored to the problem at hand and take $\hat{\theta}_n = \arg\max_\theta \mathcal{S}_n(\theta)$ as the respective *optimum score estimator*. The first four values of the arg max in Table 3, for instance, refer to the optimum score estimates of the inflation factor r based on the logarithmic score, spherical score, quadratic score, and CRPS. Pfanzagl (1969) and Birgé and Massart (1993) studied optimum score estimators under the heading of *minimum contrast estimators*. This class includes many of the most popular estimators in various situations, such as MLEs, least squares and other estimators of regression models, and estimators for mixture models or deconvolution. Pfanzagl (1969) proved rigorous versions of the consistency result (59), and Birgé and Massart (1993) related rates of convergence to the entropy structure of the parameter space. Maximum likelihood estimation forms the special case of optimum score estimation based on the logarithmic score, and optimum score estimation forms a special case of M-estimation (Huber 1964), in that the function to be optimized derives from a strictly proper scoring rule. When estimating the location parameter in a Gaussian population with known variance, for example, the optimum score estimator based on the CRPS amounts to an M-estimator with a ψ-function of the form $\psi(x) = 2\Phi(\frac{x}{c}) - 1$, where c is a positive constant and Φ denotes the standard Gaussian cumulative. This provides a smooth version of the ψ-function for Huber's (1964) robust minimax estimator (see Huber 1981, p. 208). Asymptotic results for M-estimators, such as the consistency theorems of Huber (1967) and Perlman (1972), then apply to optimum scores estimators as well. Wald's (1949) classical proof of the consistency of MLEs relies heavily on the strict propriety of the logarithmic score, which is proved in his lemma 1.

The appeal of optimum score estimation lies in the potential adaption of the scoring rule to the problem at hand. Gneiting et al. (2005) estimated a predictive regression model using the optimum score estimator based on the CRPS—a choice motivated by the meteorological problem. They showed empirically that such an approach can yield better predictive results than approaches using maximum likelihood plug-in estimates. This agrees with the findings of Copas (1983) and Friedman (1989), who showed that the use of maximum likelihood and least squares plug-in estimates can be suboptimal in prediction problems. Buja et al. (2005) argued that strictly proper scoring rules are the natural loss functions or fitting criteria in binary class probability estimation, and proposed tailoring scoring rules in situations in which false positives and false negatives have different cost implications.

9.2 Quantile Estimation

Koenker and Bassett (1978) proposed quantile regression using an optimum score estimator based on the proper scoring rule (41).

9.3 Interval Estimation

We now turn to interval estimation. Casella, Hwang, and Robert (1993, p. 141) pointed out that "the question of measuring optimality (either frequentist or Bayesian) of a set estimator against a loss criterion combining size and coverage does not yet have a satisfactory answer."

Their work was motivated by an apparent paradox due to J. O. Berger, which concerns interval estimators of the location parameter θ in a Gaussian population with unknown scale. Under the loss function

$$L(I;\theta) = c\lambda(I) - \mathbb{1}\{\theta \in I\}, \tag{60}$$

where c is a positive constant and $\lambda(I)$ denotes the Lebesgue measure of the interval estimate I, the classical t-interval is dominated by a misguided interval estimate that shrinks to the sample mean in the cases of the highest uncertainty. Casella et al. (1993, p. 145) commented that "we have a case where a disconcerting rule dominates a time honored procedure. The only reasonable conclusion is that there is a problem with the loss function." We concur, and propose using proper scoring rules to assess interval estimators based on a loss criterion that combines width and coverage.

Specifically, we contend that a meaningful comparison of interval estimators requires either equal coverage or equal width. The loss function (60) applies to all set estimates, regardless of coverage and size, which seems unnecessarily ambitious. Instead, we focus attention on interval estimators with equal nominal coverage and use the negatively oriented interval score (43). This loss function can be written as

$$L_\alpha(I;\theta) = \lambda(I) + \frac{2}{\alpha}\inf_{\eta\in I}|\theta-\eta| \tag{61}$$

and applies to interval estimates with upper and lower exceedance probability $\frac{\alpha}{2} \times 100\%$. This approach can again be traced back to Dunsmore (1968) and Winkler (1972) and avoids paradoxes, as a consequence of the propriety of the interval score. Compared with (60), the loss function (61) provides a more flexible assessment of the coverage, by taking the distance between the interval estimate and the estimand into account.

10. AVENUES FOR FUTURE WORK

Our paper aimed to bring proper scoring rules to the attention of a broad statistical and general scientific audience. Proper scoring rules lie at the heart of much statistical theory and practice, and we have demonstrated ways in which they bear on prediction and estimation. We close with a succinct, necessarily incomplete, and subjective discussion of directions for future work.

Theoretically, the relationships between proper scoring rules and divergence functions are not fully understood. The Savage representation (10), Schervish's Choquet-type representation (14), and the underlying geometric arguments surely allow generalizations, and the characterization of proper scoring rules for quantiles remains open. Little is known about the propriety of skill scores, despite Murphy's (1973) pioneering work and their ubiquitous use by meteorologists. Briggs and Ruppert (2005) have argued that skill score departures from propriety do little harm. Although we tend to agree, there is a need for follow-up studies. Diebold and Mariano (1995), Hamill (1999), Briggs (2005), Briggs and Ruppert (2005), and Jolliffe (2006) have developed formal tests of forecast performance, skill, and value. This is a promising avenue for future work, particularly in concert with biomedical applications (Pepe 2003; Schumacher, Graf, and Gerds 2003). Proper scoring rules form key tools within the broader framework of diagnostic forecast evaluation (Murphy and Winkler 1992; Gneiting et al. 2006), and in addition to hydrometeorological and biomedical uses, we see a wealth of potential applications in computational finance.

Guidelines for the selection of scoring rules are in strong demand, both for the assessment of predictive performance and in optimum score approaches to estimation. The tailoring approach of Buja et al. (2005) applies to binary class probability estimation, and we wonder whether it can be generalized. Last but not least, we anticipate novel applications of proper scoring rules in model selection and model diagnosis problems, particularly in prequential (Dawid 1984) and cross-validatory frameworks, and including Bayesian posterior predictive distributions and Markov chain Monte Carlo output (Gschlößl and Czado 2005). More traditional approaches to model selection such as Bayes factors (Kass and Raftery 1995), the Akaike information criterion, the BIC, and the deviance information criterion (Spiegelhalter, Best, Carlin, and van der Linde 2002) are likelihood-based and relate to the logarithmic scoring rule, as discussed in Section 7. We would like to know more about their relationships to cross-validatory approaches based directly on proper scoring rules, including, but not limited to, the logarithmic rule.

APPENDIX: STATISTICAL DEPTH FUNCTIONS

Statistical depth functions (Zuo and Serfling 2000) provide useful tools in nonparametric inference for multivariate data. In Section 1 we hinted at a superficial analogy to scoring rules. Specifically, if P is a Borel probability measure on $\mathbb{R}^m$, then a *depth function* $D(P,\mathbf{x})$ gives a P-based center-outward ordering of points $\mathbf{x} \in \mathbb{R}^m$. Formally, this resembles a scoring rule $S(P,\mathbf{x})$ that assigns a P-based numerical value to an event $\mathbf{x} \in \mathbb{R}^m$. Liu (1990) and Zuo and Serfling (2000) have listed desirable properties of depth functions, including maximality at the center, monotonicity relative to the deepest point, affine invariance, and vanishing at infinity. The latter two properties are not necessarily defendable requirements for scoring rules; conversely, propriety is irrelevant for depth functions.

[Received December 2005. Revised September 2006.]

REFERENCES

Baringhaus, L., and Franz, C. (2004), "On a New Multivariate Two-Sample Test," *Journal of Multivariate Analysis*, 88, 190–206.

Bauer, H. (2001), *Measure and Integration Theory*, Berlin: Walter de Gruijter.

Berg, C., Christensen, J. P. R., and Ressel, P. (1984), *Harmonic Analysis on Semigroups*, New York: Springer-Verlag.

Bernardo, J. M. (1979), "Expected Information as Expected Utility," *The Annals of Statistics*, 7, 686–690.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computing and Stochastic Systems," *Statistical Science*, 10, 3–66.

Birgé, L., and Massart, P. (1993), "Rates of Convergence for Minimum Contrast Estimators," *Probability Theory and Related Fields*, 97, 113–150.

Bregman, L. M. (1967), "The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Bremnes, J. B. (2004), "Probabilistic Forecasts of Precipitation in Terms of Quantiles Using NWP Model Output," *Monthly Weather Review*, 132, 338–347.

Brier, G. W. (1950), "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3.

Briggs, W. (2005), "A General Method of Incorporating Forecast Cost and Loss in Value Scores," *Monthly Weather Review*, 133, 3393–3397.

Briggs, W., and Ruppert, D. (2005), "Assessing the Skill of Yes/No Predictions," *Biometrics*, 61, 799–807.

Buja, A., Logan, B. F., Reeds, J. A., and Shepp, L. A. (1994), "Inequalities and Positive-Definite Functions Arising From a Problem in Multidimensional Scaling," *The Annals of Statistics*, 22, 406–438.

Buja, A., Stuetzle, W., and Shen, Y. (2005), "Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications," manuscript, available at *www-stat.wharton.upenn.edu/~buja/*.

Campbell, S. D., and Diebold, F. X. (2005), "Weather Forecasting for Weather Derivatives," *Journal of the American Statistical Association*, 100, 6–16.

Candille, G., and Talagrand, O. (2005), "Evaluation of Probabilistic Prediction Systems for a Scalar Variable," *Quarterly Journal of the Royal Meteorological Society*, 131, 2131–2150.

Casella, G., Hwang, J. T. G., and Robert, C. (1993), "A Paradox in Decision-Theoretic Interval Estimation," *Statistica Sinica*, 3, 141–155.

Cervera, J. L., and Muñoz, J. (1996), "Proper Scoring Rules for Fractiles," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 513–519.

Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39, 841–862.

Collins, M., Schapire, R. E., and Singer, J. (2002), "Logistic Regression, AdaBoost and Bregman Distances," *Machine Learning*, 48, 253–285.

Copas, J. B. (1983), "Regression, Prediction and Shrinkage," *Journal of the Royal Statistical Society*, Ser. B, 45, 311–354.

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.

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

——— (1986), "Probability Forecasting," in *Encyclopedia of Statistical Sciences*, Vol. 7, eds. S. Kotz, N. L. Johnson, and C. B. Read, New York: Wiley, pp. 210–218.

——— (1998), "Coherent Measures of Discrepancy, Uncertainty and Dependence, With Applications to Bayesian Predictive Experimental Design," Research Report 139, University College London, Dept. of Statistical Science.

——— (2006), "The Geometry of Proper Scoring Rules," Research Report 268, University College London, Dept. of Statistical Science.

Dawid, A. P., and Sebastiani, P. (1999), "Coherent Dispersion Criteria for Optimal Experimental Design," *The Annals of Statistics*, 27, 65–81.

Déqué, M., Royer, J. T., and Stroe, R. (1994), "Formulation of Gaussian Probability Forecasts Based on Model Extended-Range Integrations," *Tellus*, Ser. A, 46, 52–65.

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

Duffie, D., and Pan, J. (1997), "An Overview of Value at Risk," *Journal of Derivatives*, 4, 7–49.

Dunsmore, I. R. (1968), "A Bayesian Approach to Calibration," *Journal of the Royal Statistical Society*, Ser. B, 30, 396–405.

Eaton, M. L. (1982), "A Method for Evaluating Improper Prior Distributions," in *Statistical Decision Theory and Related Topics III*, eds. S. S. Gupta and J. O. Berger, New York: Academic Press, pp. 329–352.

Eaton, M. L., Giovagnoli, A., and Sebastiani, P. (1996), "A Predictive Approach to the Bayesian Design Problem With Application to Normal Regression Models," *Biometrika*, 83, 111–125.

Epstein, E. S. (1969), "A Scoring System for Probability Forecasts of Ranked Categories," *Journal of Applied Meteorology*, 8, 985–987.

Feuerverger, A., and Rahman, S. (1992), "Some Aspects of Probability Forecasting," *Communications in Statistics—Theory and Methods*, 21, 1615–1632.

Friederichs, P., and Hense, A. (2006), "Statistical Down-Scaling of Extreme Precipitation Events Using Censored Quantile Regression," *Monthly Weather Review*, in press.

Friedman, D. (1983), "Effective Scoring Rules for Probabilistic Forecasts," *Management Science*, 29, 447–454.

Friedman, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, 165–175.

Garratt, A., Lee, K., Pesaran, M. H., and Shin, Y. (2003), "Forecast Uncertainties in Macroeconomic Modelling: An Application to the U.K. Economy," *Journal of the American Statistical Association*, 98, 829–838.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005), "Statistical Methods for Eliciting Probability Distributions," *Journal of the American Statistical Association*, 100, 680–700.

Geisser, S., and Eddy, W. F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160.

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11.

Gerds, T. (2002), "Nonparametric Efficient Estimation of Prediction Error for Incomplete Data Models," unpublished doctoral dissertation, Albert-Ludwigs-Universität Freiburg, Germany, Mathematische Fakultät.

Giacomini, R., and Komunjer, I. (2005), "Evaluation and Combination of Conditional Quantile Forecasts," *Journal of Business & Economic Statistics*, 23, 416–431.

Gneiting, T. (1998), "Simple Tests for the Validity of Correlation Function Models on the Circle," *Statistics & Probability Letters*, 39, 119–122.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2006), "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society*, Ser. B, in press.

Gneiting, T., and Raftery, A. E. (2005), "Weather Forecasting With Ensemble Methods," *Science*, 310, 248–249.

Gneiting, T., Raftery, A. E., Balabdaoui, F., and Westveld, A. (2003), "Verifying Probabilistic Forecasts: Calibration and Sharpness," presented at the Workshop on Ensemble Forecasting, Val-Morin, Québec.

Gneiting, T., Raftery, A. E., Westveld, A., and Goldman, T. (2005), "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation," *Monthly Weather Review*, 133, 1098–1118.

Good, I. J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society*, Ser. B, 14, 107–114.

——— (1971), Comment on "Measuring Information and Uncertainty," by R. J. Buehler, in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott, Toronto: Holt, Rinehart and Winston, pp. 337–339.

Granger, C. W. J. (2006), "Preface: Some Thoughts on the Future of Forecasting," *Oxford Bulletin of Economics and Statistics*, 67S, 707–711.

Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. (2006), "The Continuous Ranked Probability Score for Circular Variables and Its Application to Mesoscale Forecast Ensemble Verification," *Quarterly Journal of the Royal Meteorological Society*, in press.

Grimit, E. P., and Mass, C. F. (2002), "Initial Results of a Mesoscale Short-Range Ensemble System Over the Pacific Northwest," *Weather and Forecasting*, 17, 192–205.

Grünwald, P. D., and Dawid, A. P. (2004), "Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory," *The Annals of Statistics*, 32, 1367–1433.

Gschlößl, S., and Czado, C. (2005), "Spatial Modelling of Claim Frequency and Claim Size in Insurance," Discussion Paper 461, Ludwig-Maximilians-Universität, Munich, Germany, Sonderforschungsbereich 368.

Hamill, T. M. (1999), "Hypothesis Tests for Evaluating Numerical Precipitation Forecasts," *Weather and Forecasting*, 14, 155–167.

Hamill, T. M., and Wilks, D. S. (1995), "A Probabilistic Forecast Contest and the Difficulty in Assessing Short-Range Forecast Uncertainty," *Weather and Forecasting*, 10, 620–631.

Hendrickson, A. D., and Buehler, R. J. (1971), "Proper Scores for Probability Forecasters," *The Annals of Mathematical Statistics*, 42, 1916–1921.

Hersbach, H. (2000), "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems," *Weather and Forecasting*, 15, 559–570.

Hofmann, T., Schölkopf, B., and Smola, A. (2005), "A Review of RKHS Methods in Machine Learning," preprint.

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101.

——— (1967), "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability I*, eds. L. M. Le Cam and J. Neyman, Berkeley, CA: University of California Press, pp. 221–233.

——— (1981), *Robust Statistics*, New York: Wiley.

Jeffreys, H. (1939), *Theory of Probability*, Oxford, U.K.: Oxford University Press.

Jolliffe, I. T. (2006), "Uncertainty and Inference for Verification Measures," *Weather and Forecasting*, in press.

Jolliffe, I. T., and Stephenson, D. B. (eds.) (2003), *Forecast Verification: A Practicioner's Guide in Atmospheric Science*, Chichester, U.K.: Wiley.

Kabaila, P. (1999), "The Relevance Property for Prediction Intervals," *Journal of Time Series Analysis*, 20, 655–662.

Kabaila, P., and He, Z. (2001), "On Prediction Intervals for Conditionally Heteroscedastic Processes," *Journal of Time Series Analysis*, 22, 725–731.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Knorr-Held, L., and Rainer, E. (2001), "Projections of Lung Cancer in West Germany: A Case Study in Bayesian Prediction," *Biostatistics*, 2, 109–129.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Machado, J. A. F. (1999), "Goodness-of-Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310.

Kohonen, J., and Suomela, J. (2006), "Lessons Learned in the Challenge: Making Predictions and Scoring Them," in *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment*, eds. J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché-Buc, Berlin: Springer-Verlag, pp. 95–116.

Koldobskiĭ, A. L. (1992), "Schoenberg's Problem on Positive Definite Functions," *St. Petersburg Mathematical Journal*, 3, 563–570.

Krzysztofowicz, R., and Sigrest, A. A. (1999), "Comparative Verification of Guidance and Local Quantitative Precipitation Forecasts: Calibration Analyses," *Weather and Forecasting*, 14, 443–454.

Langland, R. H., Toth, Z., Gelaro, R., Szunyogh, I., Shapiro, M. A., Majumdar, S. J., Morss, R. E., Rohaly, G. D., Velden, C., Bond, N., and Bishop, C. H. (1999), "The North Pacific Experiment (NORPEX-98): Targeted Observations for Improved North American Weather Forecasts," *Bulletin of the American Meteorological Society*, 90, 1363–1384.

Laud, P. W., and Ibrahim, J. G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 57, 247–262.

Lehmann, E., and Casella, G. (1998), *Theory of Point Estimation* (2nd ed.), New York: Springer.

Liu, R. Y. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414.

Ma, C. (2003), "Nonstationary Covariance Functions That Model Space–Time Interactions," *Statistics & Probability Letters*, 61, 411–419.

Mason, S. J. (2004), "On Using Climatology as a Reference Strategy in the Brier and Ranked Probability Skill Scores," *Monthly Weather Review*, 132, 1891–1895.

Matheron, G. (1984), "The Selectivity of the Distributions and the 'Second Principle of Geostatistics,' " in *Geostatistics for Natural Resources Characterization*, eds. G. Verly, M. David, and A. G. Journel, Dordrecht: Reidel, pp. 421–434.

Matheson, J. E., and Winkler, R. L. (1976), "Scoring Rules for Continuous Probability Distributions," *Management Science*, 22, 1087–1096.

Mattner, L. (1997), "Strict Definiteness via Complete Monotonicity of Integrals," *Transactions of the American Mathematical Society*, 349, 3321–3342.

McCarthy, J. (1956), "Measures of the Value of Information," *Proceedings of the National Academy of Sciences*, 42, 654–655.

Murphy, A. H. (1973), "Hedging and Skill Scores for Probability Forecasts," *Journal of Applied Meteorology*, 12, 215–223.

Murphy, A. H., and Winkler, R. L. (1992), "Diagnostic Verification of Probability Forecasts," *International Journal of Forecasting*, 7, 435–455.

Nau, R. F. (1985), "Should Scoring Rules Be 'Effective'?," *Management Science*, 31, 527–535.

Palmer, T. N. (2002), "The Economic Value of Ensemble Forecasts as a Tool for Risk Assessment: From Days to Decades," *Quarterly Journal of the Royal Meteorological Society*, 128, 747–774.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford, U.K.: Oxford University Press.

Perlman, M. D. (1972), "On the Strong Consistency of Approximate Maximum Likelihood Estimators," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability I*, eds. L. M. Le Cam, J. Neyman, and E. L. Scott, Berkeley, CA: University of California Press, pp. 263–281.

Pfanzagl, J. (1969), "On the Measurability and Consistency of Minimum Contrast Estimates," *Metrika*, 14, 249–272.

Potts, J. (2003), "Basic Concepts," in *Forecast Verification: A Practicioner's Guide in Atmospheric Science*, eds. I. T. Jolliffe and D. B. Stephenson, Chichester, U.K.: Wiley, pp. 13–36.

Quiñonero-Candela, J., Rasmussen, C. E., Sinz, F., Bousquet, O., and Schölkopf, B. (2006), "Evaluating Predictive Uncertainty Challenge," in *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment*, eds. J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché-Buc, Berlin: Springer, pp. 1–27.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005), "Using Bayesian Model Averaging to Calibrate Forecast Ensembles," *Monthly Weather Review*, 133, 1155–1174.

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton, NJ: Princeton University Press.

Roulston, M. S., and Smith, L. A. (2002), "Evaluating Probabilistic Forecasts Using Information Theory," *Monthly Weather Review*, 130, 1653–1660.

Savage, L. J. (1971), "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66, 783–801.

Schervish, M. J. (1989), "A General Method for Comparing Probability Assessors," *The Annals of Statistics*, 17, 1856–1879.

Schumacher, M., Graf, E., and Gerds, T. (2003), "How to Assess Prognostic Models for Survival Data: A Case Study in Oncology," *Methods of Information in Medicine*, 42, 564–571.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Selten, R. (1998), "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1, 43–62.

Shuford, E. H., Albert, A., and Massengil, H. E. (1966), "Admissible Probability Measurement Procedures," *Psychometrika*, 31, 125–145.

Smyth, P. (2000), "Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood," *Statistics and Computing*, 10, 63–72.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion and rejoinder), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–616.

Staël von Holstein, C.-A. S. (1970), "A Family of Strictly Proper Scoring Rules Which Are Sensitive to Distance," *Journal of Applied Meteorology*, 9, 360–364.

——— (1977), "The Continuous Ranked Probability Score in Practice," in *Decision Making and Change in Human Affairs*, eds. H. Jungermann and G. de Zeeuw, Dordrecht: Reidel, pp. 263–273.

Székely, G. J. (2003), "$\mathcal{E}$-Statistics: The Energy of Statistical Samples," Technical Report 2003-16, Bowling Green State University, Dept. of Mathematics and Statistics.

Székely, G. J., and Rizzo, M. L. (2005), "A New Test for Multivariate Normality," *Journal of Multivariate Analysis*, 93, 58–80.

Taylor, J. W. (1999), "Evaluating Volatility and Interval Forecasts," *Journal of Forecasting*, 18, 111–128.

Tetlock, P. E. (2005), *Political Expert Judgement*, Princeton, NJ: Princeton University Press.

Theis, S. (2005), "Deriving Probabilistic Short-Range Forecasts From a Deterministic High-Resolution Model," unpublished doctoral dissertation, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany, Mathematisch-Naturwissenschaftliche Fakultät.

Toth, Z., Zhu, Y., and Marchok, T. (2001), "The Use of Ensembles to Identify Forecasts With Small and Large Uncertainty," *Weather and Forecasting*, 16, 463–477.

Unger, D. A. (1985), "A Method to Estimate the Continuous Ranked Probability Score," in *Preprints of the Ninth Conference on Probability and Statistics in Atmospheric Sciences, Virginia Beach, Virginia*, Boston: American Meteorological Society, pp. 206–213.

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics*, 20, 595–601.

Weigend, A. S., and Shi, S. (2000), "Predicting Daily Probability Distributions of S&P500 Returns," *Journal of Forecasting*, 19, 375–392.

Wilks, D. S. (2002), "Smoothing Forecast Ensembles With Fitted Probability Distributions," *Quarterly Journal of the Royal Meteorological Society*, 128, 2821–2836.

——— (2006), *Statistical Methods in the Atmospheric Sciences* (2nd ed.), Amsterdam: Elsevier.

Wilson, L. J., Burrows, W. R., and Lanzinger, A. (1999), "A Strategy for Verification of Weather Element Forecasts From an Ensemble Prediction System," *Monthly Weather Review*, 127, 956–970.

Winkler, R. L. (1969), "Scoring Rules and the Evaluation of Probability Assessors," *Journal of the American Statistical Association*, 64, 1073–1078.

——— (1972), "A Decision-Theoretic Approach to Interval Estimation," *Journal of the American Statistical Association*, 67, 187–191.

——— (1994), "Evaluating Probabilities: Asymmetric Scoring Rules," *Management Science*, 40, 1395–1405.

——— (1996), "Scoring Rules and the Evaluation of Probabilities" (with discussion and reply), *Test*, 5, 1–60.

Winkler, R. L., and Murphy, A. H. (1968), "'Good' Probability Assessors," *Journal of Applied Meteorology*, 7, 751–758.

——— (1979), "The Use of Probabilities in Forecasts of Maximum and Minimum Temperatures," *Meteorological Magazine*, 108, 317–329.

Zastavnyi, V. P. (1993), "Positive Definite Functions Depending on the Norm," *Russian Journal of Mathematical Physics*, 1, 511–522.

Zuo, Y., and Serfling, R. (2000), "General Notions of Statistical Depth Functions," *The Annals of Statistics*, 28, 461–482.