Strictly Proper Scoring Rules, Prediction, and Estimation

Tilmann Gneiting and Adrian E. Raftery

Technical Report no. 463R
Department of Statistics, University of Washington

November 2005

Abstract

Scoring rules assess the quality of probabilistic forecasts, by assigning a numerical score based on the forecast 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 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 scientific problem at hand.

This paper 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 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, random-fold cross-validated likelihood.

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.

Report Documentation Page		*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **NOV 2005**	2. REPORT TYPE	3. DATES COVERED **00-11-2005 to 00-11-2005**

4. TITLE AND SUBTITLE **Strictly Proper Scoring Rules, Prediction, and Estimation**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **38**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

1   Introduction

One of the major purposes of statistical analysis is to make forecasts for the future, and to 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 applications such as weather and climate prediction (Palmer 2002; Gneiting and Raftery 2005), stochastic finance (Duffie and Pan 1997) and macroeconomic forecasting (Garratt, Lee, Pesaran and Shin 2003). In the statistical literature, advances in Markov chain Monte Carlo methodology (see, for example, 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. Gneiting, Raftery, Balabdaoui and Westveld (2003) and Gneiting, Balabdaoui and Raftery (2005) contend 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 forecast 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, thereby 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, her reward is $S(P, x)$. The function $S(P, \cdot)$ takes values 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 judgement is the distributional forecast Q. The forecaster has no incentive to predict any $P \neq Q$, and is encouraged to quote 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 the scoring rule is said to be *proper*. Propriety is essential in scientific and operational forecast evaluation, and our case study below provides a striking example of some of the difficulties resulting 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, 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 scientific problem at hand, maximize $\mathcal{S}_n(\theta)$ over the parameter space, and take $\hat{\theta}_n = \arg\max_\theta \mathcal{S}_n(\theta)$ as the *optimum score estimator* based on the scoring rule S. 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. The appeal of optimum score estimation lies in the potential adaptation of the scoring rule to the problem at hand. Apparently, this approach has only very recently been explored (Buja, Stuetzle and Shen 2005; Gneiting, Raftery, Westveld and Goldman 2005).

This paper reviews and develops the theory of proper scoring rules on general probability spaces, proposes and discusses examples thereof, and supplies case studies. The remainder of the paper is organized as follows. Section 2 states a fundamental characterization theorem, reviews the links between proper scoring rules, information measures, entropy functions and Bregman divergences, and introduces skill scores. Section 3 turns to scoring rules for categorical variables. We provide a rigorous version of the Savage (1971) representation and relate to a more recent characterization of Schervish (1989). 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* has lately attracted the attention of meteorologists, 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*. Section 5 introduces an even more general construction that is based on negative definite functions and inequalities of Hoeffding type, with side results on expectation inequalities and positive definite kernels that are of interest in their own right. Section 6 studies scoring rules for quantile and interval forecasts. We show the class of proper scoring rules for quantile forecasts to be larger than conjectured by Cervera and Muñoz (1996) and introduce 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, random-fold cross-validated likelihood. Section 8 presents the case study on the use of scoring rules in the evaluation of probabilistic weather forecasts. Section 9 turns to optimum score estimation and closes the paper. We discuss point, quantile and interval estimation, and propose the use of the interval score as a utility function that addresses width as well as coverage. Scoring rules show a superficial analogy to statistical depth functions, as we briefly discuss in the Appendix.

2 Characterizations of proper scoring rules

We introduce notation, provide characterizations of (strictly) proper scoring rules, and relate them to information measures and Bregman divergences. The discussion is more technical than in the remainder of the paper, 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 $\mathcal{P}$-*quasiintegrable* if it is measurable with respect to $\mathcal{A}$ and is quasiintegrable 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 $\mathcal{P}$-quasiintegrable for all $P \in \mathcal{P}$. Hence, 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) or Section 3.1 of Grünwald and Dawid (2004). 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

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

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. Clearly, finite sums of (strictly) proper scoring rules and $\mathcal{P}$-integrable functions are (strictly) proper. The term was apparently coined by Winkler and Murphy (1968, p. 754), but the general idea dates back at least to 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

$$G((1 - \lambda)P_0 + \lambda P_1) \leq (1 - \lambda) \, G(P_0) + \lambda \, G(P_1) \quad \text{for all} \quad \lambda \in (0, 1), \ P_0, P_1 \in \mathcal{P}. \tag{2}$$

It is *strictly convex* if (2) 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, quasiintegrable with respect to all $Q \in \mathcal{P}$, and

$$G(Q) \geq G(P) + \int G^*(P, \omega) \, \mathrm{d}(Q - P)(\omega) \tag{3}$$

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

Definition 2.1 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 2.2 *A regular scoring rule $S : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ is (strictly) proper relative to the class $\mathcal{P}$ if and only if there exists a (strictly) convex, real-valued function G on $\mathcal{P}$ such that*

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

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 $P \in \mathcal{P}$.

Proof. If the scoring rule S is of the stated form, the subtangent inequality (3) 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 therefore is convex on $\mathcal{P}$. Furthermore, the subtangent inequality (3) holds with $G^*(P, \omega) = S(P, \omega)$. This implies the representation (4) and proves the claim for propriety. By analogy to 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 $P \neq Q$, and this is equivalent to G being strictly convex on $\mathcal{P}$. ∎

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

2.2 Information measures and Bregman divergences

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

$$G(P) = \sup_{Q \in \mathcal{P}} S(Q, P), \qquad P \in \mathcal{P}, \tag{5}$$

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

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

the associated *divergence function*. 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, the divergence function becomes the *Bregman divergence* (Bregman 1967). Bregman divergences play major roles in optimization, and recently have 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 (6) for a proper scoring rule S, and to describe principled ways of finding such a scoring rule. The landmark paper 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 $(p - q)^2$.

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

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

Nau (1985) calls 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 through 5 provide numerous examples of proper scoring rules on general sample spaces with the associated entropy functions and divergence functions. For instance, the classical logarithmic score is linked to Shannon entropy and Kullback-Leibler divergence. Grünwald and Dawid (2004) and Buja et al. (2005) give further examples of proper scoring rules, entropy and divergence functions on finite sample spaces, and discuss the connections to the Bregman distance in detail.

2.3   Skill scores

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

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

over a fixed set of forecast situations. We give examples of this in case studies in Sections 6 and 8 below. Recommendations for choosing a scoring rule can be found in Section 6 of Winkler (1996) and throughout this paper.

Scores for competing forecast procedures are directly comparable if they refer to exactly the same set of situations. If scores for distinct sets of situations are compared, considerable care needs to be exercised to separate the confounding effects of intrinsic predictability and predictive performance. For example, there is substantial spatial and temporal variability in the predictability of weather and climate elements (Langland et al. 1999; Campbell and Diebold 2005). Hence, 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

$$\mathrm{SS}_n = \frac{S_n^{\mathrm{fcst}} - S_n^{\mathrm{ref}}}{S_n^{\mathrm{opt}} - S_n^{\mathrm{ref}}}, \tag{7}$$

where S_n^{fcst} is the forecaster's average score, S_n^{opt} is the mean score for a hypothetical ideal or optimal forecast, and S_n^{ref} is the average score for a reference strategy (Murphy 1973; Wilks 1995, p. 237; Potts 2003, p. 27; Briggs and Ruppert 2005). Skill scores are standardized in that (7) 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 the 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 temperature. Climatological forecasts are independent of the forecast horizon; they are calibrated by construction, but often lack sharpness.

Unfortunately, skill scores of the form (7) 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) recent claim of the propriety of the Brier skill score rests upon 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 we give examples of proper scoring rules.

3.1 Savage representation

We consider probabilistic forecasts of a categorical variable. Hence, 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

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

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 p and the event i materializes, her reward is $S(p, i)$. Theorem 3.2 below is a special case of Theorem 2.2 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, in the rigorous treatment, and in

the introduction of appropriate tools of convex analysis (Rockafellar 1970, Sections 23–25). Specifically, let $G : \mathcal{P}_m \to \mathbb{R}$ be a convex function. A vector $G'(p) = (G'_1(p), \ldots, G'_m(p))$ is a *subgradient* of G at the point $p \in \mathcal{P}_m$ if

$$G(q) \geq G(p) + \langle G'(p), q - p \rangle \tag{8}$$

for all $q \in \mathcal{P}_m$, where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product. We assume that the components of $G'(p)$ are real-valued, except that we permit $G'_i(p) = -\infty$ if $p_i = 0$.

Definition 3.1 A scoring rule S for categorical forecasts is *regular* if $S(\cdot, i)$ is real-valued for $i = 1, \ldots, m$, except possibly that $S(p, i) = -\infty$ if $p_i = 0$.

Theorem 3.2 (McCarthy, Savage) *A regular scoring rule S for categorical forecasts is (strictly) proper if and only if*

$$S(p, i) = G(p) - \langle G'(p), p \rangle + G'_i(p) \qquad for \qquad i = 1, \ldots, m, \tag{9}$$

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

Put slightly differently, a regular scoring rule S is (strictly) proper if and only if the expected score function $G(p) = S(p, p)$ is (strictly) convex on $\mathcal{P}_m$ and the vector with components $S(p, i)$ for $i = 1, \ldots, m$ is a subgradient of G at the point p, for all $p \in \mathcal{P}_m$. In view of these results, every bounded (strictly) convex function G on $\mathcal{P}_m$ generates a regular (strictly) proper scoring rule. This function G becomes the expected score function, information measure or entropy function (5) associated with the score, and the divergence function (6) is the respective Bregman distance.

We now give a number of examples. The scoring rules in Examples 3.3 through 3.5 are strictly proper, and the score in Example 3.6 is proper but not strictly proper.

Example 3.3 (quadratic or Brier score) If $G(p) = \sum_{j=1}^m p_j^2 - 1$ then (9) yields the quadratic score or Brier score,

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

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

Example 3.4 (spherical score) Let $\alpha > 1$ and consider the generalized entropy function $G(p) = (\sum_{j=1}^m p_j^\alpha)^{1/\alpha}$. This corresponds to the pseudospherical score,

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

which reduces to the traditional spherical score when $\alpha = 2$. The associated Bregman divergence is $d(p, q) = (\sum_{j=1}^m q_j^\alpha)^{1/\alpha} - \sum_{j=1}^m p_j q_j^{\alpha-1} / (\sum_{j=1}^m q_j^\alpha)^{(\alpha-1)/\alpha}$.

Example 3.5 (logarithmic score) Negative Shannon entropy, $G(p) = \sum_{j=1}^{m} p_j \log p_j$, corresponds to the logarithmic score, $S(p,i) = \log p_i$. The associated Bregman distance is the Kullback-Leibler divergence, $d(p,q) = \sum_{j=1}^{m} q_j \log(q_j/p_j)$. This scoring rule is classical and dates back at least to Good (1952). Detailed information-theoretic perspectives and interpretations in terms of gambling returns can be found in Roulston and Smith (2002) and Daley and Vere-Jones (2004).

Example 3.6 (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,

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

where $M(p) = \{i : p_i = \max_{j=1,\ldots,m} p_j\}$ denotes the set of modes of 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, for example, Toth, Zhu and Marchok 2001). The associated expected score or generalized entropy function (5) is $G(p) = \max_{j=1,\ldots,m} p_j$, and the divergence function (6) becomes

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

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

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

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

for all $p \in \mathcal{P}_m$, for all permutations π on m elements and for all events $i = 1,\ldots,m$. Winkler (1994; 1996, Section 5) argued that symmetric rules do not always appropriately reward forecasting skill and called for asymmetric ones, particularly in situations in which traditionally skills scores have been employed. Asymmetric (strictly) proper scoring rules can be generated by applying Theorem 3.2 to (strictly) convex entropy functions G that are not invariant under coordinate permutation.

3.2 Schervish representation

The classical case of *yes* or *no* forecasts for a dichotomous event suggests further discussion. We follow the literature in considering the sample space $\Omega = \{1,0\}$. A probabilistic forecast is a quoted probability $p \in [0,1]$ for *yes* or 1. A scoring rule S can be identified with a pair of functions $S(\,\cdot\,,1) : [0,1] \to \overline{\mathbb{R}}$ and $S(\,\cdot\,,0) : [0,1] \to \overline{\mathbb{R}}$. Hence, $S(p,1)$ is the forecaster's reward if she quotes p and the event materializes, and $S(p,0)$ is the reward if 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 for binary events.

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 3.2 shows that every regular (strictly) proper scoring rule is of the form

$$S(p,1) = G(p) + (1-p)G'(p), \qquad S(p,0) = G(p) - pG'(p), \tag{10}$$

where $G : [0,1] \to \mathbb{R}$ is a (strictly) 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 subgradient $G'(p)$ is real-valued, except that we permit $G'(0) = -\infty$ and $G'(1) = \infty$. 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).

The Savage representation (10) implies various interesting properties of regular (strictly) proper scoring rules. For instance, we conclude from Theorem 24.2 of Rockafellar (1970) that

$$S(p,1) = \lim_{q \to 1} G(q) - \int_p^1 \left(G'(q) - G'(p) \right) \mathrm{d}q \tag{11}$$

for $p \in (0,1)$, and since $G'(p)$ is (strictly) increasing, $S(p,1)$ is (strictly) increasing, too. Similarly, $S(p,0)$ is (strictly) decreasing, as one intuitively expects. Alternative proofs of these and other results can be found in the appendix of Schervish (1989).

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 (10) the Savage representation applies to a larger class of scoring rules than that of Schervish.

Theorem 3.7 (Schervish) *Suppose S is a regular scoring rule. Then S is proper and such that $S(0,1) = \lim_{p \to 0} S(p,1)$, $S(0,0) = \lim_{p \to 0} S(p,0)$ and both $S(p,1)$ and $S(p,0)$ are left continuous if and only if there exists a measure ν on $(0,1)$ such that*

$$S(p,1) = S(1,1) - \int_{[p,1)} (1-q)\,\nu(\mathrm{d}q), \qquad S(p,0) = S(0,0) - \int_{[0,p)} q\,\nu(\mathrm{d}q), \tag{12}$$

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

Proof. Suppose S satisfies the assumptions of the theorem. To prove that $S(p,1)$ is of the form (12), consider the representation (11), identify the increasing function $G'(p)$ with the left continuous distribution function of a 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 above steps. The statement for strict propriety follows from well-known properties of convex functions. ∎

Pearl (1978) considered scoring rules from an economic perspective, and Schervish (1989) proposed a general method for comparing binary forecasters within the framework of two-decision problems. A two-decision problem can be characterized by a cost-loss ratio $q \in [0, 1]$ that reflects the relative costs of the two possible types of inferior decision. The measure $\nu(\mathrm{d}q)$ in the representation (12) assigns relevance to distinct cost-loss ratios. If the expected score function, G, is sufficiently smooth, then $\nu(\mathrm{d}q)$ has Lebesgue density $-G''(q)$ (Buja et al. 2005). For instance, the quadratic or Brier score has entropy function $G(p) = 2p(1-p)$ and corresponds to a uniform measure. 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 $(q(1-q))^{-1}$. Buja et al. (2005) took this approach a major step further. They gave a comprehensive discussion of scoring rules for dichotomous events and introduced a parametric family of proper scoring rules, which includes the quadratic or Brier score, the logarithmic score, a scoring rule that underlies boosting and a left-continuous version of the zero-one score as special cases.

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 $p(\omega) = \lim_{\rho \to 0} P(S_\rho(\omega))/\mu(S_\rho(\omega))$ where $S_\rho(\omega)$ is a sphere of radius ρ centered at ω.

We begin by discussing scoring rules that correspond to Examples 3.3, 3.4 and 3.5. The *quadratic score*,

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

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 is $d(p, q) = \|p - q\|_2^2$. 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*,

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

emerges as the limiting case $\alpha \to 1$ in suitably scaled pseudospherical scores. This scoring rule was proposed by Good (1952) and has been widely used since, sometimes under other 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 that are 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 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 of the form $S(p, \omega) = a \log p(\omega) + f(\omega)$ for some constant $a \geq 0$ and function f. 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 the standard normal distribution and the uniform distribution on $(-\epsilon, \epsilon)$, respectively. If $\epsilon < \sqrt{\log 2}$ then

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

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.

If Lebesgue densities on the real line are used to predict discrete observations, 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 the PASCAL Challenges Workshop in Southampton in April 2005 and is described at `www.kyb.tuebingen.mpg.de/bs/people/jqc/southampton`. It disappears if scores 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 frequently impractical. Probabilistic quantitative precipitation forecasts, for instance, involve distributions with a point mass at zero (Krzysztofowicz and Sigrest 1999; Bremnes 2004). This could be handled by considering densities with respect to a mixed dominating measure in place of Lebesgue measure, but 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. Sensitivity to distance seems particularly desirable when the predictive distributions tend to be multimodal.

To address this situation, let $\mathcal{P}$ consist of the Borel probability measures on $\mathbb{R}$. We identify a probabilistic forecast, that is, a member of the class $\mathcal{P}$, with its cumulative distribution function F, and we use standard notation for the elements of the sample space $\mathbb{R}$. Let $\mathbf{1}\{y \geq x\}$ denote the function that attains the value 1 if $y \geq x$ and the value 0 otherwise. The *continuous ranked probability score* is defined as

$$\mathrm{CRPS}(F, x) = -\int_{-\infty}^{\infty} \left(F(y) - \mathbf{1}\{y \geq x\}\right)^2 \, \mathrm{d}y \tag{13}$$

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

Applications of the continuous ranked probability score have been hampered by a lack of analytic expressions, and the use of numerical quadrature rules for the evaluation of (13) has been proposed instead (Staël von Holstein 1977; Unger 1985). However, analytic expressions can be derived in some cases using the following results. By Lemma 2.2 of Baringhaus and Franz (2004) or identity (17) of Székely and Rizzo (2005),

$$\mathrm{CRPS}(F, x) = \frac{1}{2} E_F |X - X'| - E_F |X - x|, \tag{14}$$

where X and X' are independent copies of a random variable with distribution function F and finite first moment. For normal predictive distributions, it follows readily that

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

where φ and Φ denote the probability density function and the cumulative distribution function of a standard normal random variable, respectively. Similarly, analytical expressions can be given for other distributions. If a closed form expression is not available but random numbers with distribution F can be generated, the right-hand side of (14) can be evaluated by Monte Carlo techniques.

The continuous ranked probability score 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,

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

13

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

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

is of the Cramér-von Mises type.

The continuous ranked probability score has lately 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 2005). It is typically used in negative orientation, say $\mathrm{CRPS}^*(F, x) = -\mathrm{CRPS}(F, x)$. The representation (14) can then be written as

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

and this sheds new light on the score. In negative orientation, the continuous ranked probability score 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 continuous ranked probability score provides a direct way of comparing deterministic and probabilistic forecasts.

4.3   Energy score

We introduce a generalization of the continuous ranked probability score 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 $\mathbb{R}^m$ which are such that $E_P\|X\|^\beta$ is finite, where $\|\cdot\|$ denotes the Euclidean norm. We define the *energy score*,

$$\mathrm{ES}(P, x) = \frac{1}{2} E_P\|X - X'\|^\beta - E_P\|X - x\|^\beta, \tag{15}$$

where X and X' are independent copies of a random vector with distribution $P \in \mathcal{P}_\beta$. This generalizes the continuous ranked probability score, to which (15) reduces when $\beta = 1$ and $m = 1$, by allowing for an index $\beta \in (0, 2)$, and by applying it to distributional forecasts of a vector-valued quantity. The evaluation of (15) is straightforward if P is discrete, and Monte Carlo techniques can be used otherwise. The energy score has the potentially desirable property of invariance under joint translation and/or rotation of P and x. In negative orientation it can be interpreted as a generalization of the absolute error of order β. By Theorem 1 of Székely (2003), the energy score is strictly proper relative to the class $\mathcal{P}_\beta$. For a different and more general argument, see Section 5.1 below.

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

$$\mathrm{ES}(P, x) = -\frac{\beta \, 2^{\beta-2} \, \Gamma(\frac{m}{2} + \frac{\beta}{2})}{\pi^{m/2} \, \Gamma(1 - \frac{\beta}{2})} \int_{\mathbb{R}^m} \frac{|\varphi(y) - e^{i\langle x, y\rangle}|^2}{\|y\|^{m+\beta}} \, \mathrm{d}y, \tag{16}$$

where φ denotes the characteristic function of P. 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. This is akin to the metric studied by Eaton, Giovagnoli and Sebastiani (1996, p. 124). If P belongs to $\mathcal{P}_\beta$, Theorem 1 of Székely (2003) implies the equality of the right-hand sides in (15) and (16). In the limiting case $\beta = 2$, the right-hand side of (15) reduces to the squared Euclidean distance between x and the mean of P. This score is proper but not strictly proper relative to the class $\mathcal{P}_2$ of the Borel probability measures P for which $E_P\|X\|^2$ is finite.

4.4   Predictive model choice criterion

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 data $x_1, \ldots, x_n$. The predictive model choice criterion (PMCC) assesses the model fit through the quantity

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

where μ_i and σ_i^2 denote the expected value and the variance, respectively, 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

$$S(F, x) = -(E_F X - x)^2 - \text{Var}_F(X),$$

where X is a random variable with distribution F and finite variance. This is not a proper scoring rule: if the forecaster's true belief is F and if she wishes to maximize the expected score, she will quote the point measure at $E_F X$ — that is, a deterministic forecast — rather than the predictive distribution F.

One might also be tempted to consider an alternative to the continuous ranked probability score (13) in terms of F^{-1}, say

$$S(F, x) = -\int_0^1 \left(F^{-1}(u) - x\right)^2 \, \mathrm{d}u = -E_F(X - x)^2.$$

This relates to the Mallows (1972) distance and the Wasserstein metric and does not define a proper scoring rule either, with a hedging strategy that is identical to the above.

5   Proper scoring rules, negative definite functions and inequalities of Hoeffding type

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

5.1 Scoring rules associated with negative definite kernels

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 $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j\, g(x_i, x_j) \leq 0$ for all positive integers n, all $a_1, \ldots, a_n \in \mathbb{R}$ that sum to zero, and all $x_1, \ldots, x_n \in \Omega$. Numerous examples of negative definite kernels can be found in Berg, Christensen and Ressel (1984) and the references therein.

We now give the key result in this section.

Theorem 5.1 *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*

$$S(P, x) = \frac{1}{2} E_P\, g(X, X') - E_P\, g(X, x) \tag{17}$$

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, respectively. We need to show that

$$\frac{1}{2} E_Q\, g(Y, Y') \geq \frac{1}{2} E_P\, g(X, X') - E_{P,Q}\, g(X, Y). \tag{18}$$

If the expectation $E_{P,Q}\, g(X, Y)$ is infinite, the inequality is trivially satisfied; if it is finite, Theorem 2.1 in Berg, Christensen and Ressel (1984, p. 235) implies (18). ∎

Next we give examples of scoring rules that admit the representation in Theorem 5.1. In each case, we equip the sample space with the standard topology. Note that scores of the type (17) are straightforward to evaluate if P is discrete and has a moderate number of atoms only, as is typically true for ensemble forecasts and Markov chain Monte Carlo samples.

Example 5.2 (quadratic or Brier score) Let $\Omega = \{0, 1\}$ and suppose that $g(0, 0) = g(1, 1) = 0$ and $g(0, 1) = g(1, 0) = 2$. Then (17) recovers the quadratic or Brier score.

Example 5.3 (continuous ranked probability score) If $\Omega = \mathbb{R}$ and $g(x, x') = |x - x'|$ for $x, x' \in \mathbb{R}$ in Theorem 5.1, we obtain the continuous ranked probability score (14).

Example 5.4 (energy score) If $\Omega = \mathbb{R}^m$, $\beta \in (0, 2)$ and $g(x, x') = \|x - x'\|^\beta$ for $x, x' \in \mathbb{R}^m$, Theorem 5.1 recovers the energy score (15).

Example 5.5 (continuous ranked probability score 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. Hence,

$$S(P, \theta) = \frac{1}{2} E_P \, \alpha(\Theta, \Theta') - E_P \, \alpha(\Theta, \theta) \tag{19}$$

defines a proper scoring rule relative to the class of the Borel probability measures on the circle. Grimit et al. (2005) introduced (19) as an analogue of the continuous ranked probability score (14) that applies to directional variables, and used Fourier analytic tools to prove the propriety of the score.

We now turn to a far-reaching generalization of the energy score. For $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and $\alpha \in (0, \infty]$ define $\|x\|_\alpha = (\sum_{i=1}^m |x_i|^\alpha)^{1/\alpha}$ if $\alpha \in (0, \infty)$ and $\|x\|_\alpha = \max_{1 \le i \le m} |x_i|$ if $\alpha = \infty$. Schoenberg's theorem (Berg, Christensen and Ressel 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

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

is negative definite if and only if the following holds.

Assumption 5.6 Suppose that either (i) $m = 1$, $\alpha \in (0, \infty]$ and $\beta \in (0, 2]$; or (ii) $m \ge 2$, $\alpha \in (0, 2]$ and $\beta \in (0, \alpha]$; or (iii) $m = 2$, $\alpha \in (2, \infty]$ and $\beta \in (0, 1]$.

Example 5.7 (non-Euclidean energy score) Under Assumption 5.6, the scoring rule

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

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

The following result sharpens Theorem 5.1 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 possesses derivatives $\eta^{(k)}$ of all orders and $(-1)^k \, \eta^{(k)}(t) \ge 0$ for all nonnegative integers k and all $t > 0$.

Theorem 5.8 *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 X and X' be independent random vectors with distribution P. Then the scoring rule*

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

is strictly proper relative to the class of the Borel probability measures P on $\mathbb{R}^m$ for which $E_P \, \psi(\|X - X'\|_2^2)$ 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 $\beta \in (0,2)$, Theorem 5.8 assures the strict propriety of the energy score relative to the class of the Borel probability measures P on $\mathbb{R}^m$ for which $E_P \|X\|_2^\beta$ is finite.

5.2 Inequalities of Hoeffding type and positive definite kernels

A number of side results seem 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 (18) can be written as a Hoeffding type inequality,

$$2\, E_{P,Q}\, g(X, Y) - E_P\, g(X, X') - E_Q\, g(Y, Y') \geq 0. \tag{20}$$

See Theorem 1 of Székely and Rizzo (2005) for a nearly identical result and a converse: if g is not negative definite, then there are counterexamples to (20), and the respective scoring rule is improper. If furthermore Ω is a group, and the negative definite function g satisfies $g(x, x') = g(-x, -x')$ for $x, x' \in \Omega$, a special case of (20) can be stated as

$$E_P\, g(X, -X') \geq E_P\, g(X, X'). \tag{21}$$

In particular, if $\Omega = \mathbb{R}^m$ and Assumption 5.6 holds, inequalities (20) and (21) apply and reduce to

$$2\, E \,\|X - Y\|_\alpha^\beta - E\, \|X - X'\|_\alpha^\beta - E\, \|Y - Y'\|_\alpha^\beta \geq 0 \tag{22}$$

and

$$E\, \|X - X'\|_\alpha^\beta \leq E\, \|X + X'\|_\alpha^\beta, \tag{23}$$

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

In the above 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

$$h(x, x') = g(x, -x') - g(x, x'), \qquad x, x' \in \Omega, \tag{24}$$

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 5.6,

$$h(x, x') = \|x + x'\|_\alpha^\beta - \|x - x'\|_\alpha^\beta, \qquad x, x' \in \mathbb{R}^m, \tag{25}$$

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

6 Scoring rules for quantile and interval forecasts

Occasionally, full predictive distributions are difficult to specify, and the forecaster might quote predictive quantiles or prediction intervals instead. Christoffersen (1998) and Bremnes (2004) gave examples of this type of situation.

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 the quantiles $r_1, \ldots, r_k$ and x materializes, 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, the forecaster who wishes to maximize the expected score is encouraged to be honest and to volunteer 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 (26) and (28) be $\mathcal{P}$-measurable and grow at most polynomially in x. We write $\mathbf{1}\{x \leq r\}$ for the function that takes the value 1 if $x \leq r$ and the value 0 otherwise. Theorem 6.1 addresses the prediction of a single quantile; Corollary 6.2 turns to the general case.

Theorem 6.1 *If s is nondecreasing and h is arbitrary, the scoring rule*

$$S(r; x) = \alpha s(r) + (s(x) - s(r)) \, \mathbf{1}\{x \leq r\} + h(x) \tag{26}$$

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

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

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

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

$$S(r; x) = (x - r)(\mathbf{1}\{x \geq r\} - \alpha), \tag{27}$$

which was proposed by Koenker and Machado (1999), Taylor (1999) and Theis (2005, p. 232) for measuring in-sample goodness-of-fit and out-of-sample forecast performance, respectively.

Corollary 6.2 *If s_i is nondecreasing for $i = 1, \ldots, k$ and h is arbitrary, the scoring rule*

$$S(r_1, \ldots, r_k; x) = \sum_{i=1}^{k} \Big(\alpha_i s_i(r) + (s_i(x) - s_i(r_i)) \, \mathbf{1}\{x \le r_i\}\Big) + h(x) \tag{28}$$

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 6.2 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 (26) and (28) 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, whose lower and upper endpoints are given by the predictive quantile 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 stand for the quoted $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantile, respectively. Hence, if the forecaster quotes the $(1 - \alpha) \times 100\%$ central prediction interval $[l, u]$ and x materializes, 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 $h(x) = -\frac{x}{\alpha}$ in (28) yields the *interval score*,

$$S_\alpha^{\text{int}}(l, u; x) = -\left((u - l) + \frac{2}{\alpha}(l - x)\mathbf{1}\{x < l\} + \frac{2}{\alpha}(x - u)\mathbf{1}\{x > u\}\right). \tag{29}$$

This scoring rule has intuitive appeal and — in the form of a utility function — can be traced back at least to Dunsmore (1968) and Winkler (1972). The forecaster is rewarded for narrow prediction intervals, and she avoids an additional penalty whose size depends on α if the interval covers the observation. In the particular case $\alpha = 0.50$, Hamill and Wilks (1995, p. 622) used a score that is negatively oriented yet equivalent to the interval score. They noted that "a strategy for gaming [...] was not obvious" which is confirmed by the propriety of the score.

6.3   Prediction intervals for a conditionally heteroscedastic process

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 a 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 thereby maximize the expected value of all proper scores.

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

$$X_{t+1} = \frac{1}{2}X_t + \frac{1}{2}X_t \epsilon_t + \epsilon_t, \tag{30}$$

where the ϵ_t are independent standard normal 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

$$I = \left[\frac{1}{2}X_t - c \left| 1 + \frac{1}{2}X_t \right|, \frac{1}{2}X_t + c \left| 1 + \frac{1}{2}X_t \right| \right], \tag{31}$$

where $c = \Phi^{-1}(0.975)$. 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, respectively. Kabaila and He considered two alternative prediction intervals, namely

$$J = \left[F^{-1}(0.025), F^{-1}(0.975) \right], \tag{32}$$

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

$$K = \left[\frac{1}{2}X_t - \gamma \left(\left| 1 + \frac{1}{2}X_t \right| \right), \frac{1}{2}X_t + \gamma \left(\left| 1 + \frac{1}{2}X_t \right| \right) \right], \tag{33}$$

where

$$\gamma(y) = \begin{cases} \left(2 \log \left(\frac{7.36}{y} \right) \right)^{1/2} y & \text{if } 0 < y \leq 7.36, \\ 0 & \text{if } y \geq 7.36. \end{cases}$$

This choice of γ minimizes the expected width of the prediction interval under the constraint of nominal coverage. However, the interval forecast (33) seems misguided; it collapses to a point forecast when the conditional predictive variance is highest.

We generated a sample path of length $100\,001$ from the bilinear process (30) and considered the interval forecasts (31), (32) and (33), respectively. Table 1 summarizes the results of this experiment. All three interval forecasts showed close to nominal coverage, and the prediction interval (33) showed the smallest average width. Nevertheless, the classical prediction interval (31) performed best in terms of the interval score.

6.4   Scoring rules for distributional forecasts

Specifying a predictive cumulative distribution function is equivalent to specifying all predictive quantiles; hence, one 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. In particular, 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

$$S(F, x) = \int_0^1 S_\alpha(F^{-1}(\alpha); x) \, \nu(\mathrm{d}\alpha) \tag{34}$$

Table 1: One-step ahead 95% prediction intervals for the stationary bilinear process (30). Results of a simulation study using 100 000 interval forecasts each.

Interval Forecast	Empirical Coverage	Average Width	Average Interval Score
I (31)	95.01%	4.00	$-$ 9.55
J (32)	95.08%	5.45	-16.09
K (33)	94.98%	3.79	-10.64

is proper, subject to regularity and integrability constraints.

Similarly, one 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

$$S(F, x) = \int_{-\infty}^{\infty} S(F(y), \mathbf{1}\{y \geq x\})\, \nu(\mathrm{d}y) \tag{35}$$

is proper, subject to integrability constraints (Matheson and Winkler 1976; Gerds 2002). The continuous ranked probability score (13) corresponds to the special case in (35) in which S is the quadratic or Brier score and ν is Lebesgue measure. This construction carries over to the multivariate case. If $\mathcal{P}$ denotes the class of the Borel probability measures on $\mathbb{R}^m$, we identify a probabilistic forecast $P \in \mathcal{P}$ with its cumulative distribution function F. A multivariate analogue of the continuous ranked probability score can be defined as

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

This is a weighted integral of the Brier scores at all m-variate thresholds, and the Borel measure ν can be chosen to encourage the forecaster to concentrate her efforts on the important ones. If ν is finite and dominates Lebesgue measure, this score 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-validated likelihood.

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 we have a sample $X = (X_1, \ldots, X_n)$ 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 paper we have concentrated on the situation where the forecasting rule is completely specified before any of the X_i is 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

$$B = \frac{P(X|H_1)}{P(X|H_2)}, \tag{36}$$

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

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

$$\log B = \text{LogS}(H_1, X) - \text{LogS}(H_2, X). \tag{37}$$

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

So far in this paper 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 (36), but now $P(X|H_k)$ is the *integrated likelihood*,

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

where $p(X|\theta_k, H_k)$ 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:

$$P(X|H_k) = \prod_{t=1}^{n} P(X_t|X^{t-1}, H_k), \tag{38}$$

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

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

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

Let us denote by $S_{k,B}$ the log integrated likelihood, viewed now as a scoring rule. It helps to view it as a scoring rule to rewrite it as

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

23

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

$$S_{k,D} = \sum_{t=1}^{n} \log P(X_t | X^{t-1}, \hat{\theta}_k^{t-1}), \tag{39}$$

where $\hat{\theta}_k^{t-1}$ is the maximum likelihood estimator (MLE) of θ_k based on the past observations, X^{t-1}, in the sense that $S_{k,D}/S_{k,B} \to 1$ as $n \to \infty$. He also showed that $S_{k,B}$ is asymptotically equivalent to the BIC score,

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

where $d_k = \dim(\theta_k)$, in the same sense, namely $S_{k,\text{BIC}}/S_{k,B} \to 1$ as $n \to \infty$. This justifies the use of 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, and 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 (38) by

$$S_{k,B}^* = \sum_{t=1}^{n} E_D[\log p(X_t | X^{(D)}, H_k)], \tag{40}$$

where D is a random sample from $\{1, \dots, t-1, t+1, \dots, n\}$, whose size is a random variable that has a discrete uniform distribution on $\{0, 1, \dots, n-1\}$. Dawid's result (39) implies that this is asymptotically equivalent to the plug-in maximum likelihood version,

$$S_{k,D}^* = \sum_{t=1}^{n} E_D[\log p(X_t | X^{(D)}, \hat{\theta}_k^{(D)}, H_k)], \tag{41}$$

where $\hat{\theta}_k^{(D)}$ is the MLE of θ_k based on $X^{(D)}$.

The formulations (40) and (41) may be useful because they turn a score that was a sum of non-identically 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 holdout sample size. This general approach can be traced back at least to Geisser and Eddy (1975). One issue in cross-validation generally is how much data to leave out, and 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 (37) that when there are no parameters being estimated, the Bayes factor is equivalent to the difference in the logarithmic score. Thus one 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 continuous ranked probability score, and we conjecture that similar asymptotic equivalences 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 that consists of 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 six-month period. Each record consists of the five ensemble member forecasts and the associated verifying observation. The root-mean-square 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 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). To obtain calibrated predictive probability distributions, it thus 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 normal, centered at the ensemble mean forecast, and with predictive standard
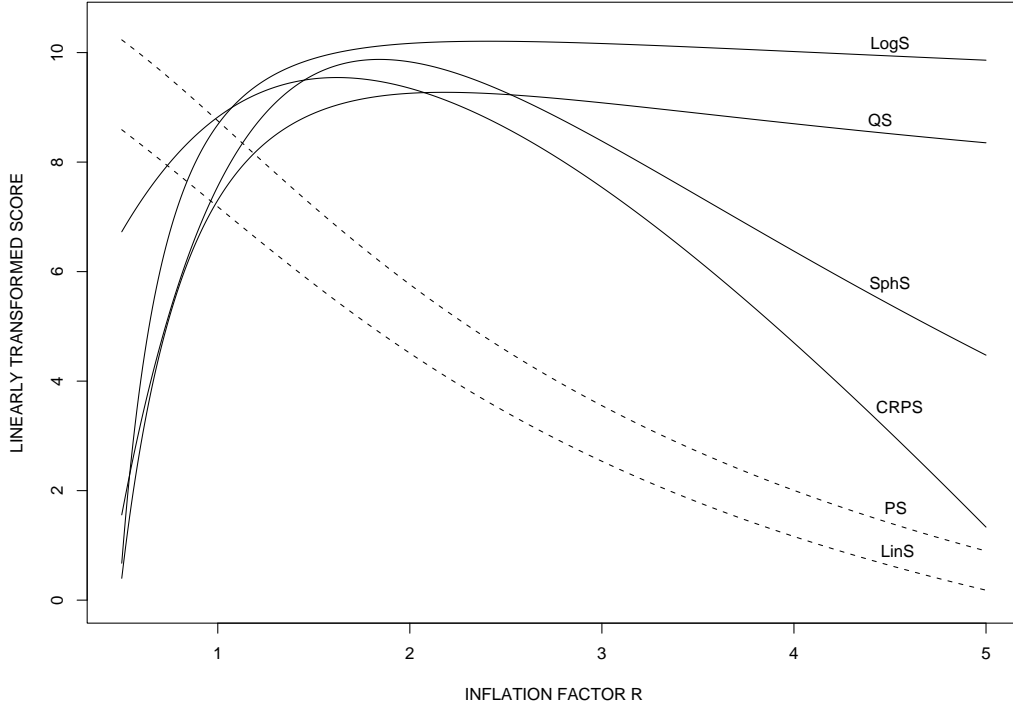
Figure 1: Probabilistic sea-level pressure forecasts over the North American Pacific Northwest in January–July 2000. The scores are shown as a function of the inflation factor r, where the predictive density is taken to be normal, centered at the ensemble mean forecast, and with predictive standard deviation equal to r times the standard deviation of the forecast ensemble. The scores were subject to linear transformations as detailed in Table 2.

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,

$$s(r) = \frac{1}{16\,015} \sum_{i=1}^{16\,015} S(\varphi_{\mu_i,r\sigma_i}, x_i), \qquad r > 0, \tag{42}$$

26

Table 2: Probabilistic sea-level pressure forecasts over the North American Pacific Northwest in January–July 2000. The predictive density is taken to be normal, centered at the ensemble mean forecast, and with predictive standard deviation equal to r times the standard deviation of the forecast ensemble.

Score	$\arg\max_r s(r)$ in Eqn. (42)	Linear Transformation in Figure 1
Quadratic score (QS)	2.18	$40s + 6$
Spherical score (SphS)	1.84	$108s - 22$
Logarithmic score (LogS)	2.41	$s + 13$
Continuous ranked probability score (CRPS)	1.62	$10s + 8$
Linear score (LinS)	0.05	$105s - 5$
Probability score (PS)	0.02	$60s - 5$

as a function of the inflation factor r. The index i refers to the i-th record in the verification data base, 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 $r = 1.55$ of the root-mean-square error of the ensemble mean forecast over the square root of the average ensemble variance.

We computed the mean score (42) for inflation factors $r \in (0,5)$ and for the quadratic score (QS), spherical score (SphS), logarithmic score (LogS), continuous ranked probability score (CRPS), linear score (LinS), and probability score (PS), as defined in Section 4. Briefly, if p denotes the predictive density and x stands for the observed value, then

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

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

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

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

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

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

Figure 1 and Table 2 summarize the results of this experiment. The scores shown in the figure are linearly transformed, and the transformations are listed in the right-hand column of the table. In the case of the quadratic score, for instance, we plotted the sum of 40 times the value in (42) and 6. Clearly, propriety is preserved under the transformation. The quadratic score, spherical score, logarithmic score and continuous ranked probability score were maximized at values of r that were larger than 1, thereby confirming the underdispersive character of the ensemble. These scores are proper. The linear score and the probability score were maximized at $r = 0.05$ and $r = 0.02$, respectively, thereby suggesting

27

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, Burrows and Lanzinger 1999). However, they are improper and their use may result in misguided scientific inferences, as in this experiment. A similar comment applies to the scores discussed 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 therefore is highly sensitive to extreme cases. Our verification data base 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 continuous ranked probability score 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 $\alpha = 0.50$ and $\alpha = 0.10$, respectively. The associated lower and upper prediction bounds l_i and u_i are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the normal distribution with mean μ_i and standard deviation $r\sigma_i$, as described above. We assessed the resulting 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

$$s_\alpha(r) = \frac{1}{16\,015} \sum_{i=1}^{16\,015} S_\alpha(l_i, u_i; x_i), \qquad r > 0, \tag{43}$$

where S_α denotes the interval score (29). This scoring rule assesses both calibration and sharpness — the latter by rewarding narrow prediction intervals, and the former by penalizing prediction intervals that do not cover the observation. Figure 2(a) shows the empirical coverage of the prediction intervals. Clearly, the coverage increased with r. If $\alpha = 0.50$ and $\alpha = 0.10$ the nominal coverage was obtained at $r = 1.78$ and $r = 2.11$, respectively. This confirms the underdispersive character of the ensemble. Figure 2(b) shows the interval score (43) as a function of the inflation factor r. If $\alpha = 0.50$ and $\alpha = 0.10$ the score was maximized at $r = 1.56$ and $r = 1.72$, respectively.

9   Optimum score estimation

Strictly proper scoring rules are also 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 the introduction. Suppose that we wish to fit a parametric model P_θ based on a sample $X_1, \ldots, X_n$ of identically distributed
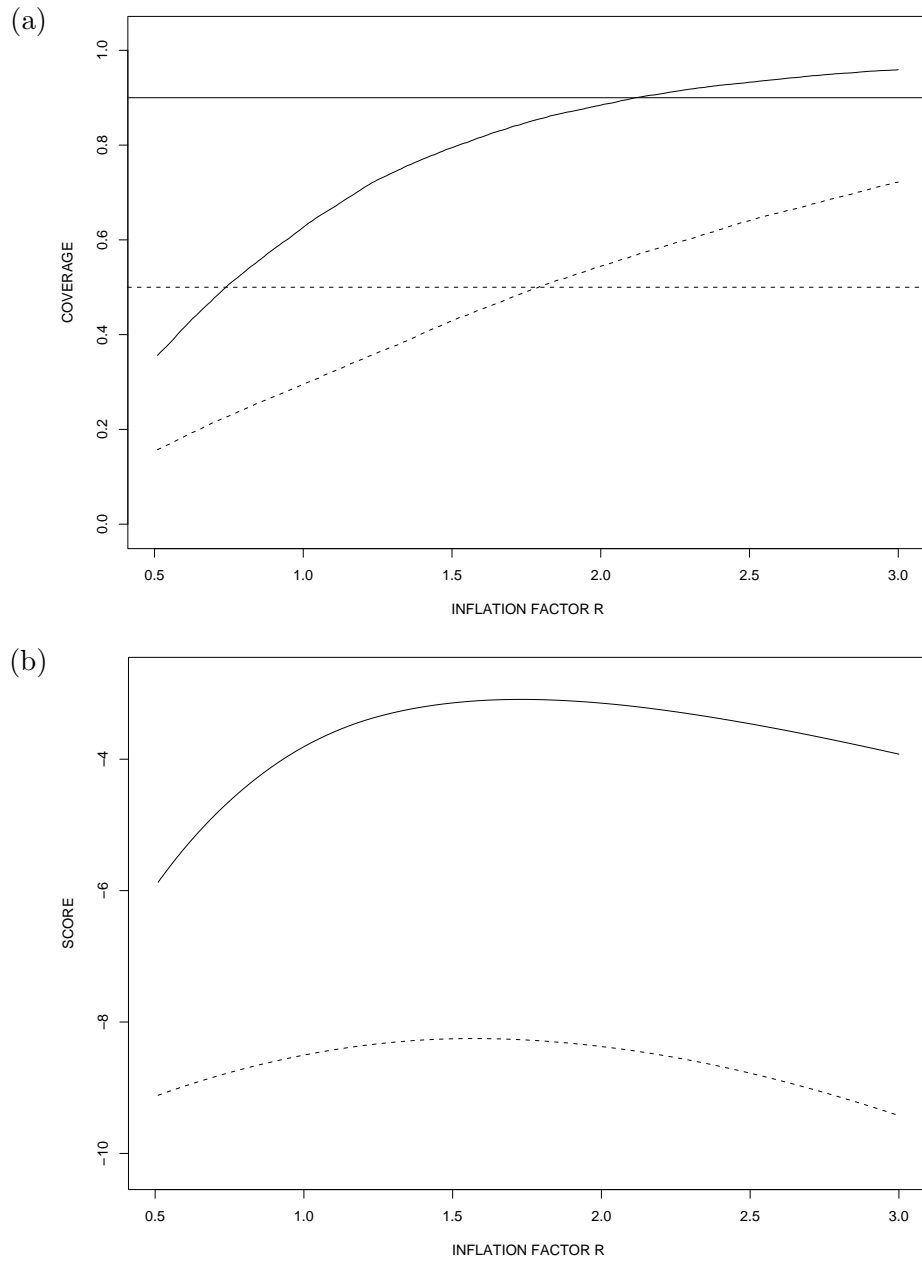
Figure 2: Interval forecasts of sea-level pressure over the North American Pacific Northwest in January–July 2000: (a) Nominal and actual coverage, and (b) the interval score (43), for the 50% central prediction interval ($\alpha = 0.50$, broken line) and the 90% central prediction interval ($\alpha = 0.10$, solid line, score scaled by a factor of 10). 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.

observations. To estimate θ, we can 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 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, asymptotic arguments indicate that

$$\arg\max_\theta \mathcal{S}_n(\theta) \to \theta_0 \quad \text{as} \quad n \to \infty. \tag{44}$$

This suggests a general approach to estimation: Choose a strictly proper scoring rule S that is tailored to the scientific problem at hand and take $\hat{\theta}_n = \arg\max_\theta \mathcal{S}_n(\theta)$ as the *optimum score estimator* based on the scoring rule S. The first four values of the arg max in Table 2, for instance, refer to the optimum score estimate for the inflation factor r based on the logarithmic score, spherical score, quadratic score and continuous ranked probability score, respectively. 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 maximum likelihood estimators, least squares and other estimators of regression models, and estimators for mixture models or deconvolution. Pfanzagl (1969) proved rigorous versions of the consistency result (44), 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 normal population with known variance, for example, the optimum score estimator based on the continuous ranked probability score 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 normal 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, too. Wald's (1949) classical proof of the consistency of maximum likelihood estimates 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. This approach has, apparently, only very recently been explored. Gneiting, Raftery, Westveld and Goldman (2005) estimated a predictive regression model using the optimum score estimator based on the continuous ranked probability score — a choice that was motivated by the meteorological problem at hand. 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) proposed the use of strictly proper scoring rules in classification and class probability estimation problems and drew links to Bayesian techniques as well as boosting.

9.2   Quantile estimation

Koenker and Bassett (1978) proposed quantile regression using an optimum score estimator that is based on the proper scoring rule (29).

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 normal population with unknown scale. Let $\mathbf{1}\{\cdot\}$ denote an indicator function. Under the loss function

$$L(I;\theta) = c\lambda(I) - \mathbf{1}\{\theta \in I\}, \tag{45}$$

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 we propose the use of strictly proper scoring rules to assess interval estimators using 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 (45) 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 (negative of the) interval score (29). This loss function can be written as

$$L_\alpha(I;\theta) = \lambda(I) + \frac{2}{\alpha} \inf_{\eta \in I} |\theta - \eta|, \tag{46}$$

and applies to interval estimates with upper and lower exceedance probability $\frac{\alpha}{2} \times 100\%$, respectively. 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. When compared to (45), the loss function (46) provides a more flexible assessment of the coverage, by taking account of the distance between the interval estimate and the estimand.

Appendix

Statistical depth functions (Zuo and Serfling 2000) provide useful tools in nonparametric inference for multivariate data. Specifically, if P is a Borel probability measure on $\mathbb{R}^m$, a *depth function* $D(P, x)$ gives a P-based center-outward ordering of points $x \in \mathbb{R}^m$. Formally, this resembles a scoring rule $S(P, x)$ that assigns a P-based numerical value to an event $x \in \mathbb{R}^m$. Liu (1990) and Zuo and Serfling (1999) list several 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 do not appear to be defendable requirements for scoring rules; conversely, propriety is irrelevant for depth functions.

Acknowledgement

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*. W. de Gruijter, Berlin.

BERG, C., CHRISTENSEN, J. P. R. AND RESSEL, P. (1984). *Harmonic Analysis on Semigroups*. Springer-Verlag, New York.

BERNARDO, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, **7**, 686–690.

BERNARDO, J. M. AND SMITH, A. F. M. (1994). *Bayesian Theory*. John Wiley, New York.

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** (3), 200–217.

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

BRIGGS, W. AND RUPPERT, D. (2005). Assessing the skill of yes/no predictions. *Biometrics*, **61**, 799–807.

BUJA, A., STUETZLE, W. AND SHEN, Y. (2005). Degrees of boosting — A study of loss functions for classification and class probability estimation. Unpublished manuscript dated 14 September 2005. Available online at www-stat.wharton.upenn.edu/~buja/.

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. *Annals of Statistics*, **22**, 406–438.

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*, Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., eds., pp. 513–519. Oxford University Press, Oxford.

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.

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.

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., 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.

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 UK 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. Ph.D. Thesis, Mathematische Fakultät, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

GNEITING, T. (1998). Simple tests for the validity of correlation function models on the circle. *Statistics & Probability Letters*, **39**, 119–122.

GNEITING, T. AND RAFTERY, A. E. (2005). Probabilistic weather forecasting using ensembles. *Science*, **310**, 248–249.

GNEITING, T., RAFTERY, A. E., BALABDAOUI, F. AND WESTVELD, A. (2003). Verifying probabilistic forecasts: Calibration and sharpness. In *Proceedings of the Workshop on Ensemble Forecasting, Val-Morin, Québec*. Available online at www.cdc.noaa.gov/people/tom.hamill/ef_workshop_2003_schedule.html.

GNEITING, T., BALABDAOUI, F. AND RAFTERY, A. E. (2005). Probabilistic forecasts, calibration and sharpness. Technical Report no. 483, Department of Statistics, University of Washington. Available online at www.stat.washington.edu/tech.reports/.

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.

GOOD, I. J. (1971). Comment on "Measuring information and uncertainty" by Robert J. Buehler. In *Foundations of Statistical Inference*, Godambe, V. P. and Sprott, D. A., eds., pp. 337–339. Holt, Rinehart and Winston, Toronto.

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.

GRIMIT, E. P., GNEITING, T., BERROCAL, V. J. AND JOHNSON, N. A. (2005). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Unpublished manuscript.

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

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. *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.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, Le Cam, L. M. and Neyman, J., eds., pp. 221–233. University of California Press, Berkeley.

HUBER, P. J. (1981). *Robust Statistics*. John Wiley, New York.

JEFFREYS, H. (1939). *Theory of Probability*. Oxford University Press, Oxford.

JOLLIFFE, I. T. AND STEPHENSON, D. B., EDS. (2003). *Forecast Verification: A Practicioner's Guide in Atmospheric Science*. Wiley, Chichester.

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.

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. Springer-Verlag, New York.

LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, **18**, 405–414.

MA, C. (2003). Nonstationary covariance functions that model space-time interactions. *Statistics & Probability Letters*, **61**, 411–419.

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Annals of Mathematical Statistics*, **43**, 508–515.

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*, Verly, G., David, M. and Journel, A. G., eds., pp. 421–434. Reidel, Dordrecht.

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.

PEARL, J. (1978). An economic basis for certain methods of evaluating probabilistic forecasts. *International Journal of Man-Machine Studies*, **10**, 175–183.

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*, Vol. I, Le Cam, L. M., Neyman, J. and Scott, E. L., eds., pp. 263–281. University of California Press, Berkeley.

PFANZAGL, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, **14**, 249–272.

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

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 University Press, Princeton.

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. *Annals of Statistics*, **17**, 1856–1879.

SCHWARZ, G. (1978). Estimating the dimension of a model. *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.

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.

STAËL VON HOLSTEIN, C.-A. S. (1977). The continuous ranked probability score in practice. In *Decision Making and Change in Human Affairs*, Jungermann, H. and de Zeeuw, G., eds., pp. 263–273. D. Reidel, Dordrecht.

SZÉKELY, G. J. (2003). $\mathcal{E}$-Statistics: The energy of statistical samples. Technical Report no. 2003–16, Department of Mathematics and Statistics, Bowling Green State University, Ohio.

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.

THEIS, S. (2005). Deriving probabilistic short-range forecasts from a deterministic high-resolution model. Ph.D. Thesis, Mathematisch-Naturwissenschaftliche Fakultät, Rheinische Friedrich-Wilhelm-Universität Bonn, Bonn, Germany.

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*, pp. 206–213. American Meteorological Society, Boston.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *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. (1995). *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego.

WILKS, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, **128**, 2821–2836.

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.

WINKLER, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, **67**, 187–191.

WINKLER, R. L. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Science*, **40**, 1395–1405.

WINKLER, R. L. (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.

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. *Annals of Statistics*, **28**, 461–482.