Strictly Proper Scoring Rules, Prediction, and Estimation

Tilmann Gneiting and Adrian E. Raftery

Technical Report no. 463

Department of Statistics, University of Washington

September 2004

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 strictly 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 any $G \neq F$. In prediction problems, strictly 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 characterizes strictly proper scoring rules on general probability spaces, and proposes and discusses examples of such. In the case of categorical and binary variables, a rigorous version of the Savage representation is established. Examples of scoring rules for probabilistic forecasts in the form of predictive densities include the spherical, pseudospherical, logarithmic and quadratic score. 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. Proper scoring rules for quantile and interval forecasts are also discussed. We relate proper scoring rules to Bayes factors and to cross-validation, and show that a particular form of cross-validation, random-fold cross-validated likelihood, corresponds to a proper scoring rule. This also allows us to define proper scoring rules when parameters defining the rule are estimated from the data.

A case study on probabilistic weather forecasts in the North American Pacific Northwest illustrates the importance of strict propriety. Optimum score approaches to point estimation are noted, and the intuitively appealing interval score is proposed as a utility function in interval estimation that addresses width as well as coverage.

Supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745 and by the National Science Foundation under Award 0134264. In Academic Year 2004/05, Tilmann Gneiting is on sabbatical leave at the Soil Physics Group, Universität Bayreuth, Universitätsstr. 30, 95440 Bayreuth, Germany.

AMS 2000 subject classifications. Primary 62A01; secondary 60G25, 62M20.

Key words and phrases. Bayes factor; Brier score; Continuous ranked probability score; Cross-validation; Energy score; Loss function; Minimum contrast estimation; Prediction interval; Predictive density; Predictive distribution; Probability assessor; Quantile forecast; Utility function; Weather prediction.

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 events (Dawid 1984). Indeed, over the past two decades probabilistic forecasting has become routine in applications such as weather prediction (Palmer 2002; Gel, Raftery and Gneiting 2004) and macroeconomic forecasting (Garratt, Lee, Pesaran and Shin 2003). Gneiting, Raftery, Balabdaoui and Westveld (2003) contend that the goal of probabilistic forecasting is to *maximize sharpness subject to calibration.* Calibration refers to statistical consistency between the distributional forecasts and the observations and is a joint property of the forecasts and the events 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. In terms of evaluation, scoring rules measure the quality of the probabilistic forecasts and reward assessors for forecasting jobs. 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 the case study in Section 7 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 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 2004; Gneiting, Westveld, Raftery and Goldman 2004).

The remainder of the paper is organized as follows. In Section 2 we prove a fundamental characterization theorem for strictly proper scoring rules on general probability spaces. Section 3 turns to scoring rules for categorical variables. The landmark paper of Savage (1971, p. 793) gave an elegant characterization of proper scoring rules that Savage described as "figurative." Theorem 3.2 provides a rigorous version thereof, and Theorem 3.4 relates to a more recent representation 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 and discuss the spherical, pseudospherical, logarithmic and quadratic score. The *continuous ranked probability score* has lately attracted the attention of meteorologists and forms a special case of a novel and very general class of scoring rules, the *energy score.* Section 5 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 central prediction intervals that is proper and has intuitive appeal. In Section 6 we relate proper scoring rules to Bayes factors and to cross-validation, and show that a particular form of cross-validation, random-fold cross-validated likelihood, corresponds to a proper scoring rule. This provides one way of defining proper scoring rules when parameters are being estimated. Section 7 presents the aforementioned case study that concerns the use of scoring rules in the assessment of probabilistic weather forecasts. Section 8 turns to optimum score estimation and closes the paper. We discuss both point estimation and interval estimation and propose the use of the interval score as a utility function that addresses width as well as coverage.

2 Characterization of strictly proper scoring rules

This section introduces notation and characterizes strictly proper scoring rules. The discussion is technical and readers with more applied interests might skip ahead without significant loss of continuity. 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 on Ω is $\mathcal{P}$-*quasiintegrable* if it is measurable with respect to $\mathcal{A}$ and 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 define

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

as 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.

A function $G : \mathcal{P} \to \overline{\mathbb{R}}$ 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 $\mathcal{P}$-quasiintegrable and

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

for all $Q \in \mathcal{P}$. The following theorem generalizes previous results by McCarthy (1956) and Hendrickson and Buehler (1971).

Theorem 2.1 *The scoring rule $S : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ is (strictly) proper if and only if there exists a (strictly) convex function $G : \mathcal{P} \to \overline{\mathbb{R}}$ such that $G(P) = S(P, P)$ for $P \in \mathcal{P}$ and*

$$S(P, \omega) = G^*(P, \omega) \tag{4}$$

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 $S(P, \omega) = G^*(P, \omega)$ is of the stated form the subtangent inequality (3) implies (1), that is, propriety. Conversely, suppose that S is a proper scoring rule. Define $G : \mathcal{P} \to \overline{\mathbb{R}}$ by $G(P) = S(P, P)$. Then the subtangent inequality (3) holds with $G^*(P, \omega) = S(P, \omega)$, which is a $\mathcal{P}$-quasiintegrable function. Furthermore,

$$G(P) = \sup\nolimits_{Q \in \mathcal{P}} S(Q, P)$$

is the pointwise supremum over a class of convex functions and therefore convex on $\mathcal{P}$. This proves the claim for propriety. In 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, the scoring rule S is (strictly) proper if and only if the expected score function $G(P) = S(P, P)$ is (strictly) convex on $\mathcal{P}$ and $S(P, \cdot\,)$ is a subtangent of G at the point P, for all $P \in \mathcal{P}$. A comparison of Theorem 2.1 to a more direct extension of the McCarthy (1956) and Hendrickson and Buehler (1971) characterization is given in the appendix.

3 Scoring rules for categorical variables

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

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 take $\mathcal{P} = \mathcal{P}_m$ where

$$\mathcal{P}_m = \Big\{ p = (p_1, \ldots, p_{m-1}) : p_1, \ldots, p_{m-1} \geq 0, \; p_1 + \cdots + p_{m-1} \leq 1 \Big\}$$

denotes the unit simplex in $\mathbb{R}^{m-1}$. A scoring rule S can then be identified with a collection of m functions

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

that is, if the forecaster quotes the probability vector p and the event i materializes, her reward is $S(p, i)$. Theorem 3.2 provides a rigorous version of the Savage (1971) representation of (strictly) proper scoring rules on finite sample spaces. Our contribution lies in the rigorous treatment and in the introduction of appropriate tools of convex analysis. We recall from Rockafellar (1970, Sections 23–25) that any nonconstant convex function $G : \mathcal{P}_m \to \overline{\mathbb{R}}$ that is bounded above is bounded below, and is continuous except possibly at the boundary of $\mathcal{P}_m$. A vector $G'(p) = (G'_1(p), \ldots, G'_{m-1}(p)) \in \overline{\mathbb{R}}^{m-1}$ is said to be a *subgradient* of G at the point $p \in \mathcal{P}_m$ if

$$G(q) \geq G(p) + \langle G'(p), q - p \rangle \tag{5}$$

for all $q \in \mathcal{P}_m$, where $\langle \cdot\,, \cdot \rangle$ denotes a scalar product. The value of the subgradient $G'(p)$ is unique and equals the gradient at every point p in the interior of $\mathcal{P}_m$ at which G is differentiable.

Definition 3.1 A scoring rule S for categorical forecasts is *regular* if $S(\cdot\,, i)$ is bounded above and real-valued for $i = 1, \ldots, m$, except possibly that if $i \leq m - 1$ and $p_i = 0$ then $S(p, i) = -\infty$, and if $p_1 + \cdots + p_{m-1} = 1$ then $S(p, m) = -\infty$.

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

$$S(p, i) = G(p) - \langle G'(p), p \rangle + G'_i(p) \quad \textit{for} \quad i = 1, \ldots, m - 1, \tag{6}$$

and

$$S(p, m) = G(p) - \langle G'(p), p \rangle, \tag{7}$$

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

Proof. If the scoring rule S admits the representation (6) and (7) with a bounded convex function G, the subgradient inequality (5) implies propriety and S is regular. Conversely, suppose that S is a regular proper scoring rule. By Theorem 2.1, $S(p, \cdot\,) = G^*(p, \cdot\,)$ where $G^*(p, \cdot\,)$ is a subtangent of the bounded convex function

$$G(p) = S(p, p) = \sum_{j=1}^{m-1} p_j\, S(p, j) + \left(1 - \sum_{j=1}^{m-1} p_j\right) S(p, m). \tag{8}$$

The subtangent inequality (3) implies that $G(q) \geq G(p) + \langle G'(p), q - p \rangle$ for all $p, q \in \mathcal{P}_m$, thereby showing that $G'_i(p) = S(p, i) - S(p, m)$ is the ith component of a subgradient of G at p, for $i = 1, \ldots, m-1$ and for all $p \in \mathcal{P}_m$. Summing over $i = 1, \ldots, m-1$, we obtain the representation (7) and then (6), and this proves the claim for propriety. The claim for strict propriety is immediate from Theorem 2.1. ■

Corollary 3.3 *A regular scoring rule S for categorical forecasts 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 $S(p, i) - S(p, m)$ for $i = 1, \ldots, m-1$ is a subgradient of G at the point p, for all $p \in \mathcal{P}_m$.*

In view of Theorem 3.2, every bounded strictly convex function $G : \mathcal{P}_m \to \mathbb{R}$ generates a strictly proper scoring rule. We put $p_m = 1 - \sum_{j=1}^{m-1} p_j$ and note a number of examples. If $G(p) = -\sum_{j=1}^{m} p_j(1 - p_j)$ then (6) and (7) yield the *Brier score* or *quadratic score*, $S(p, i) = 2p_i - 1 - \sum_{j=1}^{m} p_j^2$. The negative of the entropy function, $G(p) = \sum_{j=1}^{m} p_j \log p_j$, corresponds to the logarithmic score, $S(p, i) = \log p_i$. These scores are classical and were proposed by Brier (1950) and Good (1952), respectively. Another well-known scoring rule is the *spherical score* which corresponds to the expected score function $G(p) = (p_1^2 + \cdots + p_m^2)^{1/2}$, so that $S(p, i) = p_i / (p_1^2 + \cdots + p_m^2)^{1/2}$. The quadratic, logarithmic and spherical score are *symmetric* in the sense that

$$S\left((p_1, \ldots, p_{m-1}), i\right) = S\left((p_{\pi_1}, \ldots, p_{\pi_{m-1}}), \pi_i\right)$$

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

3.2 Schervish representation

The classical case $m = 2$ of binary "yes" or "no" forecasts calls for 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, and 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 is she quotes p and the event materializes, and $S(p, 0)$ is the compensation if she quotes p and the event does not materialize. By Theorem 3.2, every regular (strictly) proper scoring rule for binary variables is of the form

$$S(p, 1) = G(p) + (1 - p)G'(p), \qquad S(p, 0) = G(p) - pG'(p) \tag{9}$$

where $G : [0, 1] \to \mathbb{R}$ is a bounded (strictly) convex function and $G'(p)$ is a subgradient of G at the point p, for all $p \in [0, 1]$. If G is differentiable at an interior point $p \in (0, 1)$ then $G'(p)$ is simply the derivative of G at p. Putting $G(p) = -2p(1 - p)$ and $G(p) = p \log p + (1-p) \log(1-p)$ in (9) recovers the Brier score and the logarithmic score, respectively. The expected score function $G(p) = -p^{1/2}(1 - p)^{1/2}$ has been associated with boosting (Buja et al. 2004). Any strictly proper scoring rule for binary variables is regular. Furthermore, any proper binary scoring rule that is not regular satisfies either $\max\{S(p, 1), S(p, 0)\} = +\infty$ for all $p \in (0, 1)$, or $\min\{S(p, 1), S(p, 0)\} = -\infty$ for all $p \in (0, 1)$, and therefore is uninteresting.

The Savage representation (9) implies various interesting properties of (strictly) proper regular scoring rules. For instance, Theorem 24.2 of Rockafellar (1970) implies that

$$S(p, 1) = \lim_{q \to 1} G(q) - \int_p^1 \left(G'(q) - G'(p)\right) \mathrm{d}q \tag{10}$$

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 the rigorous form of Theorem 3.2 the representation of Savage applies to a larger class of scoring rules than that of Schervish.

Theorem 3.4 (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*

$$S(p, 1) = S(1, 1) - \int_{[p,1)} (1 - q)\, \nu(\mathrm{d}q), \qquad S(p, 0) = S(0, 0) - \int_{[0,p)} q\, \nu(\mathrm{d}q) \tag{11}$$

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 (11) consider the representation (10), 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 and the statement for strict propriety we refer to Schervish (1989, pp. 1876–1877). ■

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 (11) assigns relevance to distinct cost-loss ratios. For instance, the Brier score corresponds to a uniform measure, and the logarithmic score corresponds to the infinite measure with Lebesgue density $(q(1-q))^{-1}$. Buja et al. (2004) took this approach a major step further. They suggested a taxonomy of scores and introduced a parametric family of strictly proper scoring rules that includes the Brier score, the logarithmic score and the aforementioned scoring rule that underlies boosting.

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.

Good (1971) proposed the *pseudospherical score*,

$$\mathrm{PseudoS}(p, \omega) = \frac{1}{\alpha - 1} \left(\left(\frac{p(\omega)}{\|p\|_\alpha} \right)^{\alpha - 1} - 1 \right), \tag{12}$$

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 $\mathcal{L}_\alpha$, as noted by Good, and the representation (4) holds where

$$G(p) = \frac{1}{\alpha - 1} \left(\|p\|_\alpha - 1 \right).$$

The strict convexity of G and the associated subtangent inequality follow from the Hölder and Minkowski inequalities, respectively. Taking the limit as $\alpha \to 1$ in (12) yields the *logarithmic score*, $\mathrm{LogS}(p, \omega) = \log p(\omega)$, which was proposed by Good (1952) and has been widely used since. Roulston and Smith (2002) gave an information theoretic perspective and an interpretation in terms of gambling returns. The logarithmic score is strictly proper relative to the convex class $\mathcal{L}_1$ of the probability measures that are dominated by μ. The associated expected score function is negative entropy. Well-known properties of the Kullback-Leibler divergence imply its strict convexity and the associated subtangent inequality. 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 some function f.

Consequently, the *linear score*, $\mathrm{LinS}(p, \omega) = p(\omega)$, is not a proper scoring rule, despite its intuitive appeal. For instance, if $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the σ-algebra of Borel sets and μ is Lebesgue measure, let ϕ and u_ϵ denote the density of the standard normal distribution and the uniform distribution on $(-\epsilon, \epsilon)$, respectively. If $\epsilon < \sqrt{\log 2}$ then

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

in violation of propriety. Essentially, the linear score encourages overprediction at the modes of an assessor's true predictive density. Alternatives to the linear score that are strictly proper relative to the class $\mathcal{L}_2$ include the spherical score, defined above, and the *quadratic score*,

$$\mathrm{QS}(p, \omega) = 2\, p(\omega) - \int (p(\cdot))^2 \, \mathrm{d}\mu(\cdot),$$

which has expected score function $G(p) = \|p\|_2^2$ and corresponds to the Brier score when Ω is finite. 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.

4.2 Continuous ranked probability score

The restriction to absolutely continuous predictive distributions 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 rather than Lebesgue measure, but the resulting scores seem difficult to interpret. Furthermore, the scores discussed in Section 4.1 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 distribution is multimodal.

To address this situation, let $\mathcal{P}$ consist of all probability measures on $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B})$. 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). By Proposition 1 of Székely (2003),

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

where X and X' are independent copies of a random variable with distribution function F and finite first moment. When the predictive distribution is normal, $\mathcal{N}(\mu, \sigma^2)$, it follows readily from (13) or (14) that

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

where erf denotes the error function (Gneiting et al. 2004). Similarly, analytical expressions can be derived for many 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 but not strictly proper relative to $\mathcal{P}$. It is strictly proper relative to the class $\mathcal{P}_1$ of probability measures on $(\mathbb{R}, \mathcal{B})$ that have finite first moment.

The continuous ranked probability score has lately found renewed interest in the atmospheric sciences community (Hersbach 2000; Gneiting et al. 2004). It is typically used in negative orientation, say $\mathrm{CRPS}^*(F, x) = -\,\mathrm{CRPS}(F, x)$. The representation (14) can then be written as

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

which 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.

Matheson and Winkler (1976) proposed a univariate generalization of the continuous ranked probability score. We generalize further and propose a multivariate version thereof. If $\mathcal{P}$ denotes the class of probability measures on $(\mathbb{R}^m, \mathcal{B}^m)$, we identify a probabilistic forecast $P \in \mathcal{P}$ with its cumulative distribution function F. The multivariate continuous ranked probability score is defined as

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

where the integral is taken with respect to some measure W on $(\mathbb{R}^m, \mathcal{B}^m)$. This is a weighted integral of the Brier scores at all m-variate thresholds, and the measure W can be chosen to encourage the forecaster to concentrate her efforts on the important ones. In analogy to Eq. (24) of Matheson and Winkler (1976),

$$\mathrm{CRPS}(F, F_0) = -\int_{\mathbb{R}^m} (F(y) - F_0(y))^2 \, \mathrm{d}W(y) - \int_{\mathbb{R}^m} F(y)\,(1 - F(y)) \, \mathrm{d}W(y).$$

The multivariate continuous ranked probability score is proper relative to $\mathcal{P}$. If W is a finite measure that dominates Lebesgue measure, the expected score function is real-valued and strictly convex, and the score is strictly proper relative to $\mathcal{P}$.

4.3 Energy score

This section introduces another generalization of the continuous ranked probability score that draws on Székely's (2003) statistical energy perspective. Let $\mathcal{P}_\alpha$, $\alpha \in (0, 2)$, denote the class of probability measures P on $(\Omega, \mathcal{A}) = (\mathbb{R}^m, \mathcal{B}^m)$ which are such that $E_P \|X\|^\alpha$ is finite, where $\| \cdot \|$ denotes the Euclidean norm. We define the *energy score*

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

where X and X' are independent copies of an m-variate random vector with distribution P. This generalizes the univariate continuous ranked probability score, to which (15) reduces when $\alpha = 1$ and $m = 1$, by allowing for an index $\alpha \in (0, 2)$ and by applying to distributional forecasts of a vector-valued quantity. The evaluation of (15) is straightforward using Monte Carlo techniques. To prove that the energy score is strictly proper relative to $\mathcal{P}_\alpha$, recall Theorem 2.1 and note from Theorem 1 of Székely (2003) that $\mathrm{ES}(P, x)$ is a subtangent of

$$G(P) = -\frac{1}{2} E_P \left\| X - X' \right\|^\alpha,$$

which is a strictly convex function on $\mathcal{P}_\alpha$. The 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 α.

The energy score applies to the class $\mathcal{P}_0$ of all probability measures on $(\mathbb{R}^m, \mathcal{B}^m)$, by defining

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

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 $\mathcal{P}_\alpha$, Theorem 1 of Székely (2003) implies the equality of the right-hand sides in Eqs. (15) and (16), respectively. Relative to the full class $\mathcal{P}_0$, the energy score is proper but not strictly proper.

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

$$S(P, x) = -\,(E_P X - x)^2 - \text{Var}_P(X),$$

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

5 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. Bremnes (2004) gave an example of this type of situation. That said, specifying a predictive distribution is equivalent to specifying all predictive quantiles; and we can build scoring rules for predictive distributions from scoring rules for quantiles. Matheson and Winkler (1976) and Cervera and Muñoz (1996) suggested ways of doing this. For instance, we might sum the interval score (19), introduced below, over prediction intervals with equidistant or representative probability content.

5.1 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 probability measures on $(\mathbb{R}, \mathcal{B})$ 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 (17) grow at most polynomially in x. We write $\mathbf{1}\{x \leq r\}$ for the function that attains the value 1 if $x \leq r$ and the value 0 otherwise. Theorem 5.1 addresses the prediction of a single quantile; Corollary 5.2 turns to the general case.

Theorem 5.1 *If s is nondecreasing and h is arbitrary, the scoring rule*

$$S(r; x) = \alpha s(r) + (s(x) - s(r))\, \mathbf{1}\{x \leq r\} + h(x) \tag{17}$$

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. ■

Corollary 5.2 *If s_i is nondecreasing for $i = 1, \ldots, k$ and h is arbitrary the scoring rule*

$$S(r_1, \ldots, r_k; x) = \sum_{i=1}^{k} \Big(\alpha_i s_i(r) + (s_i(x) - s_i(r_i))\, \mathbf{1}\{x \leq r_i\} \Big) + h(x) \tag{18}$$

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 5.2 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 (17) and (18), respectively, provide the general form of proper scoring rules for quantiles.

5.2 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 $\alpha_1 = \frac{\alpha}{2}$, $\alpha_2 = 1-\frac{\alpha}{2}$, $s_1(x) = s_2(x) = 4x$ and $h(x) = -2x$ in (18) yields the *interval score*,

$$S_\alpha(l, u; x) = \begin{cases} -2\alpha(u-l) - 4(l-x) & \text{if } \ x \leq l, \\ -2\alpha(u-l) & \text{if } \ l \leq x \leq u, \\ -2\alpha(u-l) - 4(x-u) & \text{if } \ x \geq u. \end{cases} \tag{19}$$

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 a penalty 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 but equivalent to the interval score. They noted that "a strategy for gaming [. . .] was not obvious" which is confirmed by the propriety of the score.

5.3 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

$$X_t = \frac{1}{2} X_{t-1} + \frac{1}{2} X_{t-1} \epsilon_t + \epsilon_t \tag{20}$$

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

$$I = \left[\frac{1}{2} X_t - c \left| 1 + \frac{1}{2} X_t \right|, \frac{1}{2} X_t + c \left| 1 + \frac{1}{2} X_t \right| \right], \tag{21}$$

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

Table 1: One-step ahead 95% prediction intervals for the stationary bilinear process (20). Results of a simulation study using 100,000 interval forecasts each.

Interval Forecast		Empirical Coverage	Average Width	Average Interval Score
I	(21)	95.01%	4.00	-0.48
J	(22)	95.08%	5.45	-0.79
K	(23)	94.98%	3.79	-0.52

97.5% percentiles of the true conditional distribution function, respectively. Kabaila and He considered two alternative prediction intervals, namely

$$J = \left[F^{-1}(0.025), F^{-1}(0.975) \right], \tag{22}$$

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

$$K = \left[\frac{1}{2} X_t - \gamma\left(\left| 1 + \frac{1}{2} X_t \right| \right), \frac{1}{2} X_t + \gamma\left(\left| 1 + \frac{1}{2} X_t \right| \right) \right], \tag{23}$$

where

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

This choice of γ minimizes the expected width of the prediction interval under the constraint of nominal coverage. However, the interval forecast (23) 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 (20) and considered the interval forecasts (21), (22) and (23), respectively. Table 1 summarizes the results of this experiment. All three interval forecasts showed close to nominal coverage, and the prediction interval (23) showed the smallest average width. Nevertheless, the classical prediction interval (21) showed the largest value of the interval score.

6 Scoring rules, Bayes factors and random-fold cross-validation

6.1 Proper scoring rules 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, i.e. 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

$$B = \frac{P(X|H_1)}{P(X|H_2)}, \tag{24}$$

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

$$\log(B) = \mathrm{LogS}(H_1, X) - \mathrm{LogS}(H_2, X). \tag{25}$$

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 log 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 (24), 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:

$$P(X|H_k) = \prod_{t=1}^{n} P(X_t|X^{t-1}, H_k), \tag{26}$$

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

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

This is a proper score with respect to the Bayesian predictive density

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

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

$$S_{k,D} = \sum_{t=1}^{n} \log P(X_t|X^{t-1}, \hat{\theta}_k^{t-1}), \tag{27}$$

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 an approximation to the BIC score,

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

where $d_k = \dim(\theta_k)$, in the same sense, namely $S_{k,\mathrm{BIC}}/S_{k,B} \to 1$ 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 log 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.

6.2 Scoring rules and random-fold cross-validation

Suppose now that the data are unordered. Then equation (26) holds for any ordering of the data. We can replace (26) by

$$S_{k,B} = S^*_{k,B} = \sum_{t=1}^{n} E_D[\log p(X_t|X^{(D)}, H_k], \tag{28}$$

where D is a random sample from $\{1, \ldots, t-1, t+1, \ldots, n\}$, whose size is a random variable that has a discrete uniform distribution on $\{0, 1, \ldots, n-1\}$. Dawid's result (27) implies that this is asymptotically equivalent to the plug-in maximum likelihood version,

$$S^*_{k,D} = \sum_{t=1}^{n} E_D[\log p(X_t|X^{(D)}, \hat{\theta}_k^{(D)}, H_k], \tag{29}$$

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

The formulations (28) and (29) 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.

These results suggest that random-fold cross-validation corresponds to a proper scoring rule, asymptotically as the amount of simulation gets large. 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 as scoring rules may shed some light on this issue, for example by determining whether or not they are proper. We are not aware of results showing that other versions of cross-validation correspond to proper scoring rules.

We have seen by (25) that when there are no parameters being estimated, the Bayes factor is equivalent to the log score. Thus one could replace the log score by another proper score, and the difference in scores could be viewed as a kind of "predictive Bayes factor" with a non-log 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 log score) by another proper score, such as the continuous ranked probability score, and we conjecture that the resulting scores are also proper. Then we would have a way of generating proper scoring rules when there are parameters being estimated.

7 Case study: Probabilistic forecasts of sea-level pressure over the North American Pacific Northwest

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).

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.

The underdispersive behavior — 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 part of the sources of uncertainty (Raftery, Balabdaoui,

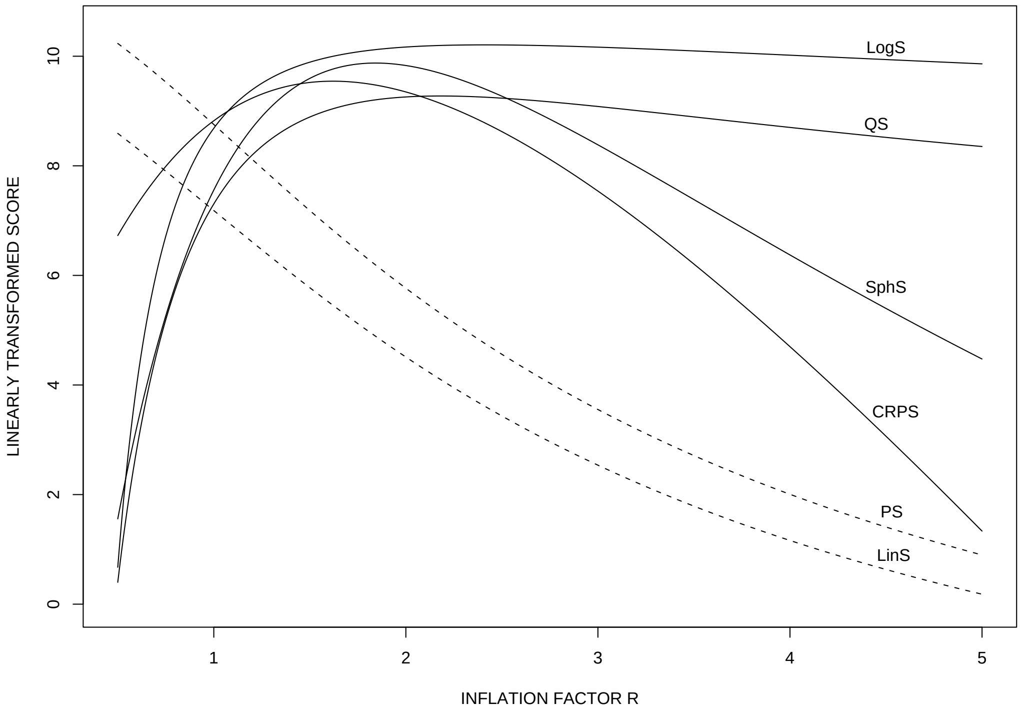

Figure 1: Probabilistic sea-level pressure forecasts over the North American Pacific Northwest in January–July 2000. The scores are shown as a function of the inflation factor r, where the predictive density is taken to be normal, centered at the ensemble mean forecast, and with predictive standard deviation equal to r times the standard deviation of the forecast ensemble. The scores were subject to linear transformations as detailed in Table 2.

Gneiting and Polakowski 2003, p. 4). 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 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*.

7.1 Evaluation of density forecasts

In the aforementioned approach the predictive density is Gaussian, say $\phi_{\mu,r\sigma}$: 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

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. (30)	Linear Transformation in Figure 1
Logarithmic score (LogS)	2.41	$s + 13$
Spherical score (SphS)	1.84	$108s + 86$
Quadratic score (QS)	2.18	$40s + 6$
Continuous ranked probability score (CRPS)	1.62	$10s + 8$
Linear score (LinS)	0.05	$105s - 5$
Probability score (PS)	0.02	$60s - 5$

considered various scoring rules S and computed the average score,

$$s(r) = \frac{1}{16\,015} \sum_{i=1}^{16\,015} S(\phi_{\mu_i, r\sigma_i}, x_i)\,, \qquad r > 0, \tag{30}$$

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 (30) for inflation factors $r \in (0, 5)$ and for the logarithmic score (LogS), spherical score (SphS), quadratic score (QS), continuous ranked probability score (CRPS), linear score (LinS), and probability score (PS), as defined in Sections 4.1 and 4.2. Briefly, if p denotes the predictive density and x stands for the observed value, then

$$\begin{aligned}
\mathrm{LogS}(p,x) &= \log p(x), \\
\mathrm{SphS}(p,x) &= \frac{p(x)}{(\int_{-\infty}^{\infty} (p(y))^2 \,\mathrm{d}y)^{1/2}} - 1, \\
\mathrm{QS}(p,x) &= 2\,p(x) - \textstyle\int_{-\infty}^{\infty} (p(y))^2 \,\mathrm{d}y, \\
\mathrm{CRPS}(p,x) &= \tfrac{1}{2}\, E_p|X - X'| - E_p|X - x|, \\
\mathrm{LinS}(p,x) &= p(x), \\
\mathrm{PS}(p,x) &= \textstyle\int_{x-1}^{x+1} p(y) \,\mathrm{d}y.
\end{aligned}$$

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 spherical score, for instance, we plotted the sum of 108 times the value in (30) and 86. Clearly, propriety is preserved under the transformation. The logarithmic score, spherical score, quadratic 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 ignorable forecast uncertainty and almost 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.

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 a more resistant alternative.

7.2 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

$$s_\alpha(r) = \frac{1}{16\,015} \sum_{i=1}^{16\,015} S_\alpha(l_i, u_i; x_i), \qquad r > 0, \tag{31}$$

where S_α denotes the interval score (19). 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 (31) 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.

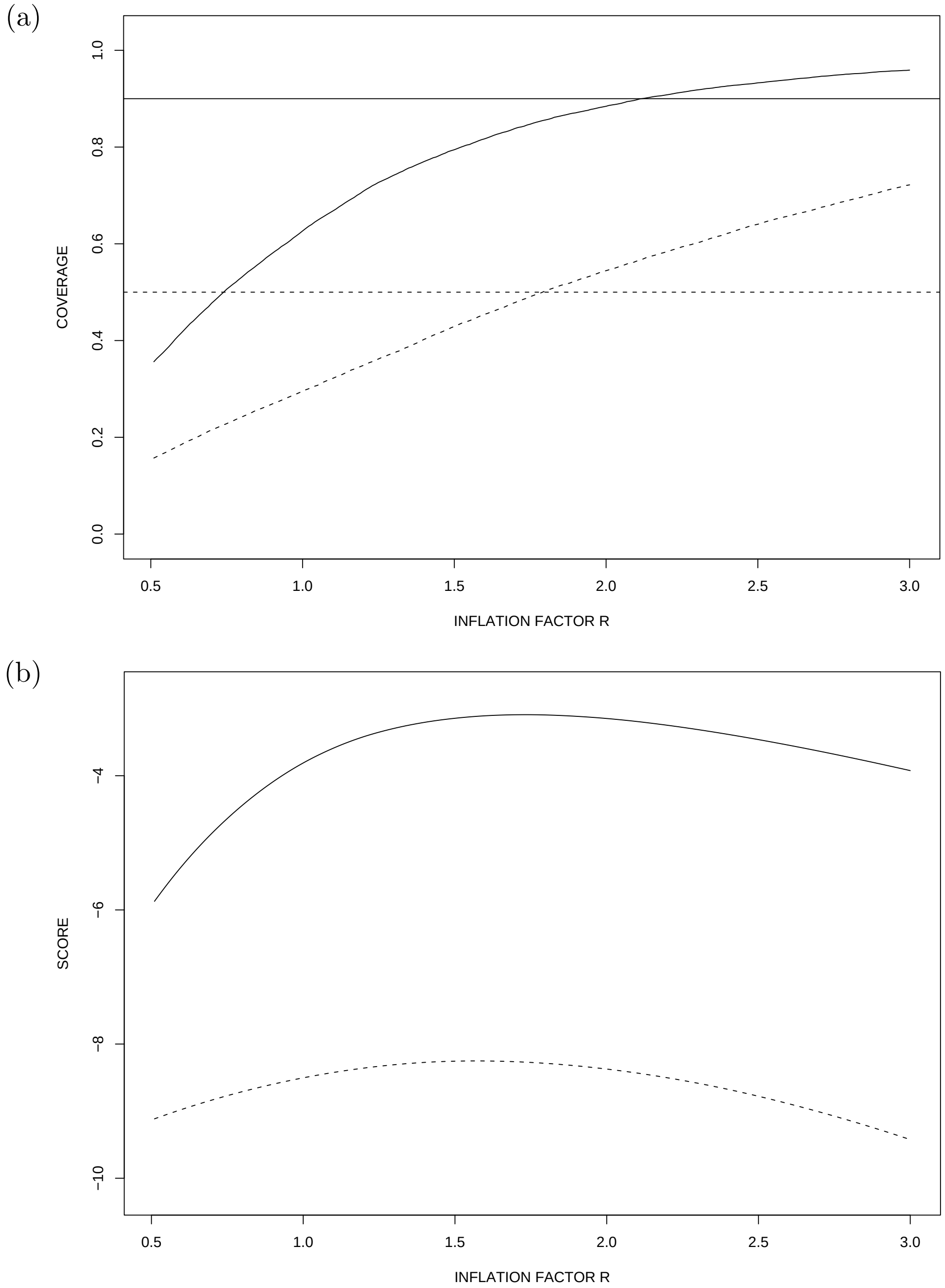

Figure 2: Interval forecasts of sea-level pressure over the North American Pacific Northwest in January–July 2000: (a) Nominal and actual coverage, and (b) the interval score (31), for 50% central prediction intervals ($\alpha = 0.50$, broken line) and 90% central prediction intervals ($\alpha = 0.10$, solid line). 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.

8 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.

8.1 Point estimation

We recall 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 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 with respect to a convex class of probability measures that contains the parametric model. If θ_0 denotes the true parameter value, asymptotic arguments indicate that

$$\arg\max\nolimits_\theta \mathcal{S}_n(\theta) \to \theta_0 \quad \text{as} \quad n \to \infty. \tag{32}$$

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 (32), 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 et al. (2004) 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 results 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. (2004) proposed the use of strictly proper scoring rules in classification and class probability estimation problems and drew links to Bayesian techniques and boosting.

8.2 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

$$L(I;\theta) = c\lambda(I) - \mathbf{1}\{\theta \in I\}, \tag{33}$$

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 (33) applies to all set estimates, regardless of coverage and size, which seems unnecessarily ambitious. As an alternative, we restrict attention to interval estimators with equal nominal coverage and use the (negative of the) interval score (19). This loss function can be written as

$$L_\alpha(I;\theta) = 2\alpha\lambda(I) + 4\inf_{\eta\in I}|\theta - \eta|, \tag{34}$$

and applies to interval estimates with upper and lower exceedance probability $\frac{\alpha}{2} \times 100\%$. This approach can, again, be traced back to Dunsmore (1968) and Winkler (1972) and avoids paradoxes, as a consequence of the propriety of the interval score. When compared to (33), the loss function (34) provides a more flexible assessment of the coverage, by taking account of the distance between the interval estimate and the estimand.

Appendix

In this appendix, we compare Theorem 2.1 to a more direct extension of the McCarthy (1956) and Hendrickson and Buehler (1971) characterization. The results of McCarthy and Hendrickson and Buehler differ from Theorem 2.1 by considering functions on the convex cone $\mathcal{D} = \{\lambda P : P \in \mathcal{P}, \lambda > 0\}$. Furthermore, Hendrickson and Buehler (1971) assumed that the convex class $\mathcal{P}$ is dominated by a σ-finite measure μ on $(\Omega, \mathcal{A})$. This assumption can be disposed of as follows. A function $H : \mathcal{D} \to \overline{\mathbb{R}}$ is *homogeneous* if $H(\lambda P) = \lambda H(P)$ for all $P \in \mathcal{D}$. If there exists a $P \in \mathcal{D}$ and a $\mathcal{P}$-quasiintegrable function $H^*(P, \cdot\,) : \Omega \to \overline{\mathbb{R}}$ such that

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

for all $Q \in \mathcal{D}$, then $H^*(P, \cdot\,)$ is said to be a *subtangent of* H *relative to* $\mathcal{D}$ at the point $P \in \mathcal{D}$. Following Hendrickson and Buehler, it is easy to show that a scoring rule $S : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ is (strictly) proper if and only if

$$S(P,\omega) = H^*(P,\omega) \tag{35}$$

for $P \in \mathcal{P}$ and $\omega \in \Omega$, where $H : \mathcal{D} \to \overline{\mathbb{R}}$ is homogeneous, (strictly) convex on $\mathcal{P}$, and such that $H^*(P, \cdot\,)$ is a subtangent of H relative to $\mathcal{D}$ at P.

In the case of a finite sample space of size m, the subtangent $H^*(P, \cdot\,)$ in the Hendrickson-Buehler characterization (35) can be identified with the subgradient of a convex function on $\mathbb{R}^m$. The representation (4) suggests the characterization of Savage (1971) who instead considered convex functions on the unit simplex in $\mathbb{R}^{m-1}$, as detailed in Section 3. That said, the representations (4) and (35) are equivalent and closely related. If S is of the form (4) with an associated convex function $G : \mathcal{P} \to \overline{\mathbb{R}}$, define $H : \mathcal{D} \to \overline{\mathbb{R}}$ by $H(\lambda P) = \lambda G(P)$ for $P \in \mathcal{P}$. Then (35) holds, since $H^*(P, \cdot\,) = G^*(P, \cdot\,)$ is a subtangent of H relative to $\mathcal{D}$ at $P \in \mathcal{P}$. Conversely, suppose that S is of the form (35) with an associated homogeneous function $H : \mathcal{D} \to \overline{\mathbb{R}}$. Then $S(P, P) = H(P)$ for $P \in \mathcal{D}$ and the representation (4) holds where $G : \mathcal{P} \to \overline{\mathbb{R}}$ is the restriction of H to $\mathcal{P}$ and $G^*(P, \cdot\,) = H^*(P, \cdot\,)$ is a subtangent of G at $P \in \mathcal{P}$.

Acknowledgement

The authors are grateful to Mark Albright, Veronica J. Berrocal, William M. Briggs, Eric P. Grimit, Peter J. Huber, Hans Kuensch, Clifford F. Mass, Werner Stuetzle, Jon A. Wellner and Lawrence J. Wilson for helpful discussions and/or providing data.

References

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

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.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, **97**, 113–150.

Bremnes, J. B. (2004). Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Monthly Weather Review*, **132**, 338–347.

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

Buja, A., Stuetzle, W. and Shen, Y. (2004). Degrees of boosting — A study of loss functions for classification and class probability estimation. Unpublished manuscript. Available online at www-stat.wharton.upenn.edu/~buja/.

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.

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

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.

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, 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.

Gel, Y., Raftery, A. E. and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion). *Journal of the American Statistical Association*, **99**, in press.

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

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., Westveld, A., Raftery, A. E. and Goldman, T. (2004). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Technical Report no. 449, Department of Statistics, University of Washington. Available online at www.stat.washington.edu/tech.reports/.

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.

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, U.K.

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.

Krzysztofowicz, R. and Sigrest, A. A. (1999). Comparative verification of guidance and local quantitative precipitation forecasts: Calibration analyses. *Weather and Forecasting*, **14**, 443–454.

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

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, **42**, 654–655.

Murphy, A. H. and Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**, 435–455.

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.

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.

Raftery, A. E., Balabdaoui, F., Gneiting, T. and Polakowski, M. (2003). Using Bayesian model averaging to calibrate forecast ensembles. Technical Report no. 440, Department of Statistics, University of Washington. Available online at www.stat.washington.edu/tech.reports/.

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.

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.

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. (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. (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.