Electronic Journal of Statistics
Vol. 10 (2016) 1098–1108
ISSN: 1935-7524
DOI: 10.1214/16-EJS1132

Extensive scoring rules

Matthew Parry

Dept of Mathematics & Statistics
University of Otago
P.O. Box 56
Dunedin 9054
New Zealand
e-mail: mparry@maths.otago.ac.nz

Abstract: Scoring rules evaluate the performance of probabilistic forecasts. A scoring rule is said to be local if it assigns a score based on the observed outcome and on outcomes that are in some sense "close" to the observed outcome. All scoring rules can be derived from a concave entropy functional and the property of locality follows when the entropy is 1-homogeneous (up to an additive constant). Consequently, except for the log score, a local scoring rule has the remarkable property that it is 0-homogeneous; in other words, it assigns a score that is independent of the normalization of the quoted probability distribution. In many statistical applications, it is not plausible to treat observed outcomes as independent, e.g. time series data or multicomponent measurements. We show that local scoring rules can be easily extended to multidimensional outcome spaces. We also introduce the notion of an extensive scoring rule, i.e. a scoring rule that ensures the score of independent outcomes is a sum of independent scores. We construct local scoring rules that are extensive and show that a scoring rule is a extensive if and only if it is derived from an extensive entropy.

MSC 2010 subject classifications: Primary 62C99; secondary 62A99.
Keywords and phrases: Additivity, homogeneity, concavity, entropy, sequential, score matching.

Received February 2016.

Contents

1. Introduction

In many instances, the predictive value of a statistical model or analysis is of vital importance. Typically, a prediction or forecast takes the form of a probability distribution for some quantity or quantities of interest. The probabilistic nature of the forecasts reflects the fact that there is an inherent uncertainty in any prediction we make (Dawid (1984)). Probabilistic forecasts were first adopted in weather forecasting but have now become common in many areas, including climate prediction and macroeconomic forecasting. The crucial question, therefore, is how to assess the performance of a forecast as events unfold. Scoring rules were devised to answer this exact question.

A desirable feature of a scoring rule is that the forecaster should be motivated to state their true belief about future events (Bröcker and Smith (2007)). In the language of decision theory, the forecaster's expected score under their true belief should be optimized by actually using that belief to make their predictions. In other words, we don't want the scoring rule to be gamed. A scoring rule that cannot be gamed is said to be proper. Proper scoring rules have been used for evaluating weather and climate predictions (McCarthy (1956); Jolliffe and Stephenson (2003); Gneiting et al. (2005); Suckling and Smith (2013)), economic and financial forecasts (Boero, Smith and Wallis (2011); Gneiting and Ranjan (2011)), and sports results modelling (Constantinou and Fenton (2012)).

Formally, a scoring rule $S(x,Q)$ is the loss on observing x having quoted the probability distribution Q for the random variable X. We have $S : (\mathcal{X}, \mathcal{P}) \to \overline{\mathbb{R}} = [-\infty, \infty]$, where $\mathcal{X}$ is the outcome space and $\mathcal{P}$ is a set of probability distributions on $\mathcal{X}$. The defining properties of a *proper* scoring rule pertain to its expectation. Letting $S(P,Q) \equiv \mathbb{E}_{X\sim P}\, S(X,Q)$, where $P \in \mathcal{P}$, we require that *(i)* $S(P,Q)$ is affine in P and *(ii)* $S(P,Q) \geq S(P,P)$ for all $P, Q \in \mathcal{P}$. The first condition means we can take $\mathcal{P}$ to be convex; the second condition means ones expected score is minimized by quoting ones true belief. If $S(P,Q) > S(P,P)$ for $Q \neq P$, we say the scoring rule is *strictly proper*. A classical example of a scoring rule is the *log score*: $S(x,Q) = -\ln q(x)$, where $q(x)$ may be a probability density or mass function. Strict propriety is then equivalent to the statement that the Kullback-Leibler divergence $d(P,Q)$ is positive for $Q \neq P$.

A unique feature of the log score is that it depends on the quoted distribution only at the observed point x. *Local scoring rules* are scoring rules that come close to obtaining this property: they depend on the quoted distribution only in a neighbourhood of the point x. When $\mathcal{X}$ is continuous, the neighbourhood is infinitesimal and the scoring rule depends on the derivatives of the probability density. The *order* of a scoring rule is the order of the highest derivative of $q(x)$. In the case when $\mathcal{X}$ is an interval of the real line, Parry, Dawid and Lauritzen (2012) characterized the form of such local scoring rules and showed that only even order scoring rules are possible. Remarkably, the local scoring rules they found were also independent of the normalization of the quoted probability density. As an example, the simplest second order scoring rule is

$$S(x,Q) = \frac{q''(x)}{q(x)} - \frac{1}{2}\left(\frac{q'(x)}{q(x)}\right)^2 = \left(\frac{q'(x)}{q(x)}\right)' + \frac{1}{2}\left(\frac{q'(x)}{q(x)}\right)^2, \tag{1}$$

and was independently discovered by Almeida and Gidas (1993) and Hyvärinen (2005). General second order scoring rules were fully developed in Ehm and Gneiting (2012).

When $\mathcal{X}$ is discrete, a neighbourhood structure is defined via a graph on the outcome space: the edge xy indicates $S(x, Q)$ depends on $q(y)$. The resulting scoring rules are also termed local and were characterized in Dawid, Lauritzen and Parry (2012), who showed that the graph is undirected. Furthermore, like their continuous counterparts, such local scoring rules are independent of the normalization of the quoted probability distribution. As an example – essentially pointed out by Hyvärinen (2007) – the negative logarithm of Besag's pseudo-likelihood is a local scoring rule. A parallel development of local scoring rules on discrete outcome spaces will be given elsewhere, but it is worth noting that most of what follows can be extended, with appropriate modification, to the discrete outcome case.

From now on we let $\mathcal{X}$ be a simply connected subset of $\mathbb{R}^n$ and let $q(x)$ be a strictly positive density with respect to the Lebesgue measure. For simplicity, we will only consider local scoring rules of second order so that $q(x)$ is assumed to be twice differentiable. In section 2, we recap the connection between entropy and scoring rules, and use this to generalize local scoring rules to multidimensional outcome spaces. In section 3, we introduce the notion of extensivity and prove that a scoring rule is extensive if and only if its associated entropy is extensive. We then construct two classes of extensive scoring rules, one that takes advantage of an (arbitrary) ordering of the data termed the sequential class, and one that is inherently a local class. Finally, in section 4, we use an example of the local class to carry out inference for an otherwise intractable Markov model.

2. Entropy and multidimensional scoring rules

Each (strictly) proper scoring rule defines a (strictly) concave entropy $H(P) := S(P, P)$. The proof[1] relies on the fact that $S(P, Q)$ is affine in P. Gneiting and Raftery (2007) showed that the converse holds (see also McCarthy (1956), Hendrickson and Buehler (1971)): if $H(P)$ is (strictly) concave and $H^\star(\cdot, P) : \mathcal{X} \to \overline{\mathbb{R}}$ is a supergradient to H at $P \in \mathcal{P}$ then

$$S(x, Q) = H(Q) + H^\star(x, Q) - H^\star(Q, Q), \tag{2}$$

where $H^\star(Q, Q) \equiv \mathbb{E}_{X \sim Q}\, H^\star(X, Q)$, is a (strictly) proper scoring rule. In practice, $H^\star(\cdot, Q)$ is often a gradient and then $H(Q)$ defines a unique scoring rule. In this construction, locality is seen to be the statement that $H(Q) = H^\star(Q, Q)$, up to an additive constant.

As an example, the entropy that generates the log score is the Shannon entropy: $H(Q) = -\int \mathrm{d}x\, q(x) \ln q(x)$. Its supergradient is $H^\star(x, Q) = -\ln q(x) - 1$, so that $H^\star(Q, Q) = H(Q) - 1$.

[1] For $\lambda \in [0, 1]$, $H((1-\lambda)P + \lambda Q) = S((1-\lambda)P + \lambda Q, (1-\lambda)P + \lambda Q) = (1-\lambda)S(P, (1-\lambda)P + \lambda Q) + \lambda S(Q, (1-\lambda)P + \lambda Q) \geq (1-\lambda)S(P, P) + \lambda S(Q, Q) = (1-\lambda)H(P) + \lambda H(Q)$.

In the case of $n=1$ local scoring rules, Parry, Dawid and Lauritzen (2012) and Ehm and Gneiting (2012) considered entropies of the form $H(Q)=\int \mathrm{d}x\; \phi(x,q(x), q'(x))$, where $\phi(x,y,y_1)$ is differentiable in x and, for almost all $x \in \mathcal{X}$, is twice differentiable, jointly concave and 1-homogeneous[2] in (y,y_1). For example, eq. (1) is obtained with the choice $\phi(x,y,y_1) = -\frac{1}{2}\frac{y_1^2}{y}$. It is the condition of 1-homogeneity that ensures $H(Q) = H^\star(Q,Q)$ and, consequently, $S(x,Q)$ is 0-homogeneous. In fact, the form of the gradient $H^\star(\cdot,Q)$ depends crucially on assuming the boundary terms that arise in integration by parts are zero; this puts important constraints on $\mathcal{P}$ (see Ehm and Gneiting (2012)).

The advantage of the entropy construction is that it suggests obvious generalizations, first to the multidimensional case, i.e. $n > 1$, and second to the multidimensional local scoring rules found by Almeida and Gidas (1993), Hyvärinen (2005), and Dawid and Lauritzen (2005). We denote the components of $x \in \mathcal{X}$ as x^i, where $i = 1,\dots,n$, and write q_i for $\partial q/\partial x^i$. We also let D_i denote the total derivative with respect to x^i. We now have the first key theoretical result of this paper:

Theorem 1. *If $\phi[y] := \phi(x^1,\dots,x^n,y,y_1,\dots y_n)$ is differentiable in x, and twice differentiable, jointly (strictly) concave and 1-homogeneous in $(y, y_1,\dots,y_n)$, then, with a slight abuse of notation,*

$$S(x,Q) = \sum_{i=1}^{n}\left(-D_i\frac{\partial}{\partial q_i} + \frac{\partial}{\partial q}\right)\phi[q] \tag{3}$$

is a (strictly) proper local scoring rule of second order.

Proof. $H(Q) = \int \mathrm{d}x\;\phi[q]$ is (strictly) concave. The proof then follows from eq. (2). □

Remark. Note that $\phi[y]$ being 1-homogeneous implies $\partial\phi/\partial y$ and $\partial\phi/\partial y_i$ are 0-homogeneous. This is a key result that we will use later.

The following example includes all previously known multidimensional scoring rules as special cases. Let $G_{ij}(x)$ be (the components of) a positive definite symmetric matrix and let G^{ij} be its inverse. Then $\phi[q] = -\frac{1}{2}q^{-1}\sum_{ij} G^{ij}q_iq_j$ generates the proper scoring rule

$$S(x,Q) = \sum_{i,j=1}^{n}\left\{G^{ij}\left(\frac{q_{ij}}{q} - \frac{1}{2}\frac{q_iq_j}{q^2}\right) + G^{ij}{}_{,i}\frac{q_j}{q}\right\}, \tag{4}$$

where $q_{ij} = \partial^2 q/\partial x^i\partial x^j$ and $G^{ij}{}_{,i} = \partial G^{ij}/\partial x^i$. When, additionally, $\mathcal{X}$ has a metric structure, the above scoring rule affords a covariant formulation. If $g_{ij}(x)$ is the metric tensor on $\mathcal{X}$ then $\overline{q}(x) = g^{-1/2}q(x)$ is the probability density with respect to the measure $g^{1/2}\mathrm{d}x$, where $g := \det[g_{ij}]$. Setting $G_{ij} = g_{ij}$, gives the scoring rule of Dawid and Lauritzen (2005) (up to an irrelevant additive

[2] A function $f(\mathbf{x})$ is k-homogeneous if, for all $\mathbf{x}$ and $\lambda > 0$, $f(\lambda\mathbf{x}) = \lambda^k f(\mathbf{x})$.

constant):

$$S(x,Q) = \sum_{i,j=1}^{n} g^{ij} \left(\frac{\nabla_i \nabla_j \bar{q}}{\bar{q}} - \frac{1}{2} \frac{\nabla_i \bar{q} \nabla_j \bar{q}}{\bar{q}^2} \right), \tag{5}$$

where ∇_i is the covariant derivative with respect to the Levi-Civita connection.

In what follows, $\phi[q]$ plays central role and we will often say it *generates* its associated local scoring rule.

3. Extensive scoring rules

At an intuitive level, *extensivity* of a scoring rule means that independent data can be taken individually or all together yet yield the same score. Extensivity, also known as additivity, has previously been applied to entropies: when the outcomes are independent, the entropy of the joint distribution becomes a sum of the entropies of the marginal distributions. As we shall see, extensivity of scoring rules and entropy are inextricably linked. To avoid trivial subcases, we assume $n > 1$ from now on.

Let $Q \in \mathcal{P}$ be a joint distribution on $\mathcal{X}$ and let $\mathcal{M}_i$ be the operation of marginalizing over all variables except x^i. In other words, $Q_i := \mathcal{M}_i Q$ is the marginal distribution for X^i. It follows that $\mathcal{P}_i := \mathcal{M}_i \mathcal{P}$ is a set of distributions on $\mathcal{X}_i := \{x^i \,|\, x \in \mathcal{X}\}$ and that $\mathcal{P}_i$ inherits convexity from $\mathcal{P}$. We now define the operator $\mathcal{I}$ by

$$\mathcal{I}Q = \prod_{i=1}^{n} \mathcal{M}_i Q = \prod_{i=1}^{n} Q_i, \tag{6}$$

i.e. $\mathcal{I}Q$ is a distribution that treats the (X^i) as independent. It is straightforward to show $\mathcal{I}^2 = \mathcal{I}$, hence $\mathcal{I}$ is a projection operator. We call the range of $\mathcal{I}$ the *center* of $\mathcal{P}$ and denote it $\mathcal{C} = \mathcal{I}\mathcal{P} \subseteq \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$. We say $\mathcal{P}$ is *centered* if $\mathcal{C} \subset \mathcal{P}$. Note that $\mathcal{C}$ is not convex and so $\mathcal{C} \neq \mathcal{P}$. Further, we call $\mathcal{R}(C) = \{Q \in \mathcal{P} \,|\, \mathcal{I}Q = C\}$ the *ray* at $C \in \mathcal{C}$. More generally, we can identify a ray by any distribution it "passes through"; we define $\mathcal{R}(Q) \equiv \mathcal{R}(\mathcal{I}Q)$.

We are now in a position to define extensivity for scoring rules. Let $\mathcal{P}$ be a convex and centered set of distributions on $\mathcal{X}$. We say a scoring rule $S(x,Q)$ on $(\mathcal{X},\mathcal{P})$ is *extensive* if it is strictly proper and if for all $Q \in \mathcal{C}$,

$$S(x,Q) = \sum_{i=1}^{n} S_i(x^i, Q_i), \tag{7}$$

where $S_i(x^i, Q_i)$ are strictly proper scoring rules on $(\mathcal{X}_i, \mathcal{P}_i)$. It follows that, for $Q \in \mathcal{C}$, $S(P,Q) = \sum_{i=1}^{n} S_i(P_i, Q_i)$. Note that the requirement of strict propriety means eq. (7) cannot be lifted to $Q \in \mathcal{P}$, for we would have $S(P,Q) = S(P,P)$ for all $Q \in \mathcal{R}(P)$. Therefore, eq. (7) represents a simplification of $S(x,Q)$ only in the case where $Q \in \mathcal{C}$.

Remark. It is worth pointing out that eq. (7) is often used when $Q \in \mathcal{P}$ and is referred to as the *observed* or *empirical score*, but it is perhaps not widely appreciated that such a definition sacrifices strict propriety.

We can also define extensivity for entropies in a similar way. We say an entropy $H(Q)$ on $\mathcal{P}$ is extensive if for all $Q \in \mathcal{C}$,

$$H(Q) = \sum_{i=1}^{n} H_i(Q_i), \tag{8}$$

where $H_i(Q_i)$ are entropies on $\mathcal{P}_i$. We now have the second key theoretical result of this paper:

Theorem 2. *If $H(Q)$ defines the scoring rule $S(x,Q)$ via eq. (2), then $S(x,Q)$ is extensive iff $H(Q)$ is extensive.*

Proof. Forwards is immediate. The reverse follows because, when $Q \in \mathcal{C}$, $H^\star(x,Q) = \sum_{i=1}^{n} H_i^\star(x^i, Q_i)$. □

3.1. Sequential class

We define the *sequential class* of extensive scoring rules as follows. Writing the joint probability density as a product of nested conditional densities (the ordering of outcomes is arbitrary), we have

$$q(x) = q(x^n|x^{1:n-1})q(x^{n-1}|x^{1:n-2}) \cdots q(x^2|x^1)q(x^1),$$

where we have introduced the shorthand notation $x^{1:j} = (x^1, \dots, x^j)$ and used the convention that the argument of $q(\cdot)$ implicitly specifies the outcome space for the density. Then the following scoring rule is extensive:

$$S(x,Q) = \sum_{i=1}^{n} S_i(x^i, Q_{i|1:i-1}), \tag{9}$$

where the S_i are strictly proper scoring rules on $(\mathcal{X}_i, \mathcal{P}_i)$.

Proof. When $Q \in \mathcal{C}$, this reduces to eq. (7) since then $Q_{i|1:i-1} = Q_i$, and strict propriety follows from the fact that

$$S(P,Q) = \sum_{i=1}^{n} \mathbb{E}_{X^{1:i-1} \sim P_{1:i-1}} S_i(P_{i|1:i-1}, Q_{i|1:i-1}). \tag{10}$$

□

Remark. The logarithmic scoring rule is clearly a member of the sequential class since $\ln q(x) = \sum_{i=1}^{n} \ln q(x^i|x^{1:i-1})$.

Result for separable Bregman scores. Separable Bregman scores are of the form

$$S(x,Q) = \psi'(q(x)) + \int \mathrm{d}y \, \{\psi(q(y)) - q(y)\psi'(q(y))\},$$

and are (strictly) proper when $\psi(s)$ is a (strictly) concave function of $s \geq 0$. (Note that eq. (2) holds with $H(Q) = \int \mathrm{d}x\, \psi(q(x))$.) For example, the well-known Brier score is given by $\psi(s) = -\frac{1}{2}s^2$. In many applications, the $\mathcal{X}_i$ are the same (possibly infinite) interval of the real line, i.e. $\mathcal{X} = \mathcal{X}_1^n$. In this case, the S_i will typically be taken to have the same functional form. The only separable Bregman score in this restricted class of sequential extensive scoring rules is the log score.

Proof. For extensivity, a necessary condition is $\psi'(s_1 \dots s_n) = f(s_1) + \cdots + f(s_n)$, for some function f. Treating this as a functional equation, we see this implies $\psi'(s) = f(s) + (n-1)f(1)$. But then the original expression becomes $f(s_1 \dots s_n) - f(1) = [f(s_1) - f(1)] + \cdots + [f(s_n) - f(1)]$ and the only solution to this is $f(s) - f(1) = \ln s$, up to irrelevant additive and multiplicative constants. □

3.2. *Local class*

We now introduce the *local class* of extensive scoring rules. Specifically, if

$$\phi[y] = \sum_{i=1}^{n} \phi_i(x^i, y, y_i), \tag{11}$$

where for all i, $\phi_i(x^i, y, y_i)$ is differentiable in x^i and twice differentiable, jointly strictly concave and 1-homogeneous in (y, y_i), then

$$S(x, Q) = \sum_{i=1}^{n} \left(-D_i \frac{\partial}{\partial q_i} + \frac{\partial}{\partial q} \right) \phi_i(x^i, q, q_i) \tag{12}$$

is an extensive scoring rule.

Proof. The proof is straightforward though somewhat hampered by the limitations of notation. When $Q \in \mathcal{C}$, $q(x) = q(x^1) \cdots q(x^n) = q(x^{-i})q(x^i)$ and $q_i(x) = q(x^{-i})q'(x^i)$, where $x^{-i} := (x^j \,|\, j \neq i)$. Then $\phi_i(x^i, q, q_i) = q(x^{-i})\, \phi_i(x^i, q(x^i), q'(x^i))$ by 1-homogeneity, and

$$\frac{\partial}{\partial q}\phi_i(x^i, q, q_i) := \frac{\partial}{\partial y}\phi_i(x^i, y, y_i)\bigg|_{y=q, y_i=q_i} = \frac{\partial}{\partial y}\phi_i(x^i, y, y_i)\bigg|_{y=q(x^i), y_i=q'(x^i)}$$

and similarly with $(\partial/\partial q_i)\phi_i(x^i, q, q_i)$, by 0-homogeneity. Crucially, when $Q \in \mathcal{C}$, $\partial\phi_i/\partial q$ and $\partial\phi_i/\partial q_i$ no longer depend on x^{-i}. Consequently, $S(x, Q)$ becomes a sum of strictly proper one-dimensional scoring rules, as required. □

As an example, a sufficient condition for local scoring rules of the form given in eq. (4) to be extensive is that $G^{ij} = h_i(x^i)\delta^{ij}$ and $h_i(\cdot) > 0$. (δ^{ij} is the Kronecker delta, which is 1 when $i = j$ and 0 otherwise.) As with the sequential class, in many applications we can expect the ϕ_i to have the same functional form. This would imply $h_i(\cdot) = h(\cdot)$.

3.3. Multivariate Normal data

We can use the equivalence of independence and vanishing covariance in jointly Normally-distributed data to illustrate both the behaviour of extensive entropies and the form of their associated scoring rules. As previously indicated, the entropy associated with the log score – the canonical example of a sequential class extensive scoring rule – is the Shannon entropy. If $X \sim \mathcal{N}(\mu, \Sigma)$, then the Shannon entropy is $H(Q) = \frac{1}{2} \ln\det 2\pi \mathrm{e}\Sigma$. When the off-diagonal terms of Σ vanish, this simplifies to $\frac{1}{2}\sum_{i=1}^{n} \ln 2\pi\mathrm{e}\,\sigma_i^2$, which is clearly the sum of the Shannon entropies of the marginal distributions. The log score is

$$S(x, Q) = \frac{1}{2}\left\{(x-\mu)^\top \Sigma^{-1}(x-\mu) + \ln\det 2\pi\Sigma\right\}. \tag{13}$$

The simplest extensive local scoring rule arises when $G^{ij} = \delta^{ij}$ in eq. (4). We can make the extensivity of its associated entropy explicit by noting that $q_i = -\sum_j (\Sigma^{-1})_{ij}(x^j - \mu^j)q$. Then $H(Q) = -\frac{1}{2}\mathrm{tr}\,\Sigma^{-1}$. When the off-diagonal terms vanish, this becomes $-\frac{1}{2}\sum_{i=1}^{n}\sigma_i^{-2}$. The extensive local scoring rule is

$$S(x, Q) = \mathrm{tr}\left\{-\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x-\mu)(x-\mu)^\top\Sigma^{-1}\right\}. \tag{14}$$

4. Application to statistical inference

If $Q = Q_\theta$ is a distribution from a parametric family, then scoring rules lead to a straightforward inferential method: $\widehat{\theta} = \arg\min_\theta S(x, Q_\theta)$. Indeed, the use of the log score is equivalent to maximum likelihood estimation. Typically, the estimating equation becomes

$$\frac{\partial S(x, Q_\theta)}{\partial \theta} = 0 \tag{15}$$

and Dawid and Lauritzen (2005) have shown that it is unbiased and leads to a consistent estimator for θ.

Consider a homogeneous discrete time Markov process as a model for a random process on the real line, observed at times $0\!:\!n$. If we model the transition probability density as

$$q(x|x') = \frac{\exp(\theta f(x - \rho x'))}{Z(\theta)}, \tag{16}$$

where ρ is assumed known, then the normalization $Z(\theta)$ is typically not computable. Nevertheless, local scoring rules enable inference in such cases. Conditional on x^0, the probability of the observations is

$$q(x^{1:n}|x^0) = q(x^n|x^{n-1})q(x^{n-1}|x^{n-2})\cdots q(x^1|x^0) = \prod_{i=1}^{n}\frac{\exp(\theta f(x^i - \rho x^{i-1}))}{Z(\theta)}. \tag{17}$$

Choosing the simplest extensive local scoring rule, namely that generated by $\phi_i(x^i, q, q_i) = -\frac{1}{2}\frac{q_i^2}{q}$, we obtain the score

$$S(x,Q) = \theta(1+\rho^2)\sum_{i=1}^{n} f''(x^i - \rho x^{i-1}) + \frac{1}{2}\theta^2 \left\{(1+\rho^2)\sum_{i=1}^{n} f'(x^i - \rho x^{i-1})^2 \right.$$
$$\left. -2\rho \sum_{i=1}^{n-1} f'(x^i - \rho x^{i-1}) f'(x^{i+1} - \rho x^i)\right\}. \quad (18)$$

Note that when $\rho = 0$, the states of the Markov chain are independent and the scoring rule reduces to a sum of independent scores, as expected. In the case of Gaussian diffusion, i.e. $f(x) = -\frac{1}{2}(x-\mu)^2$, θ is the precision parameter and $Z(\theta)$ can be computed. The resulting estimator for θ is

$$\widehat{\theta} = \left(\frac{1}{n}\left\{\sum_{i=1}^{n}(x^i - \rho x^{i-1} - \mu)^2 - \frac{2\rho}{1+\rho^2}\sum_{i=1}^{n-1}(x^i - \rho x^{i-1} - \mu)(x^{i+1} - \rho x^i - \mu)\right\}\right)^{-1}, \quad (19)$$

which differs from the maximum likelihood estimator when $\rho \neq 0$ due to the presence of the second sum. This illustrates the fact that tractability of the estimator is achieved at the cost of efficiency.

However, there is a deeper analysis of this example to be made. Under the Markov assumption, the $\overline{x}^i := x^i - \rho x^{i-1}$, for $i = 1, \ldots, n$, are independent increments. If we treat the increments (conditional on x_0) as the outcomes of interest, then $\overline{q}(\overline{x}^{1:n}|x^0) = Z(\theta)^{-1}\prod_{i=1}^{n}\exp(\theta f(\overline{x}^i))$ is the probability density on the new outcome space $\overline{\mathcal{X}}$. The scoring rule generated by $\overline{\phi}_i(\overline{x}^i, \overline{q}, \overline{q}_i) = -\frac{1}{2}\frac{\overline{q}_i^2}{\overline{q}}$ is

$$S(\overline{x}, \overline{Q}) = \theta\sum_{i=1}^{n} f''(\overline{x}^i) + \frac{1}{2}\theta^2\left\{\sum_{i=1}^{n} f'(\overline{x}^i)^2\right\}, \quad (20)$$

which is essentially eq. (18) with $\rho = 0$. In fact, we can obtain the same improved score in the original outcome space by "transforming" $-\frac{1}{2}\frac{\overline{q}_i^2}{\overline{q}}$. Specifically, noting that $(\partial/\partial\overline{x}^i) = \sum_j {J_i}^j(\partial/\partial x^j)$, where

$${J_i}^j = \begin{cases} \rho^{i-j}, & i \geq j \\ 0, & \text{otherwise,} \end{cases} \quad (21)$$

we arrive at a scoring rule of the form given in eq. (4) with $G^{ij} = \sum_{k\ell}\delta^{k\ell}{J_k}^i{J_\ell}^j = (\rho^{|i-j|} - \rho^{i+j})/(1-\rho^2)$. Note that this is no longer of the extensive class for the original outcome space but it is of the sequential class because $S(\overline{x}, \overline{Q})$ is equivalent to

$$\sum_{i=1}^{n}\left\{\left(\frac{q'(x^i|x^{i-1})}{q(x^i|x^{i-1})}\right)' + \frac{1}{2}\left(\frac{q'(x^i|x^{i-1})}{q(x^i|x^{i-1})}\right)^2\right\},$$

and this is a special case of eq. (9) with each S_i taking the form of eq. (1).

The preceding result exemplifies a property of local scoring rules first pointed out in Parry, Dawid and Lauritzen (2012), namely that they transform as scalars under transformation of the data. More precisely, if $\overline{x} = F(x)$ is an invertible transformation, then $\overline{S}(\overline{x}, \overline{Q}) = S(x, Q)$ is a scoring rule for the appropriately transformed distribution $\overline{Q}$, defined for outcomes in $\overline{\mathcal{X}} = F(\mathcal{X})$. This was proved in Parry, Dawid and Lauritzen (2012) for the case $n = 1$; the proof in the case $n > 1$ will be developed elsewhere.

4.1. Numerical example

Let $f(z) = 0.4\,(z - 0.4)^2 - 0.08\,z^4$, which is an example popularized by Mackay (2003). We take $n = 100$ and, for simplicity, fix $\rho = 1$, though it is straightforward to jointly estimate ρ and θ. Supposing $\theta = 1$, we generate samples using rejection sampling, and then numerically compare the performance of the estimators arising from eq. (18) and eq. (20). We find the bias and variance of local class estimator are both approximately 0.083, whereas the bias and variance of the sequential class estimator are both approximately 0.07.

5. Conclusions

All scoring rules can be derived from an appropriate entropy functional. We have used this fact to show that it is possible to usefully generalize local scoring rules in multidimensional settings. Local or not, we have also shown that a scoring rule has the extensive property if and only if its associated entropy does. Extensivity means that the score of independent outcomes is the sum of independent scores. Previously, only the log score was known to have this property.

In the context of multidimensional local scoring rules, an important technical question that remains is how to specify $\mathcal{P}$, the set of distributions on $\mathcal{X}$, for which the boundary entropy vanishes Ehm and Gneiting (2012); Parry, Dawid and Lauritzen (2012). A more interesting problem, however, and one with particular practical significance (e.g. Yang et al. (2014)), is how to devise local scoring rules for mixed continuous and discrete outcomes.

Acknowledgments

I am particularly grateful for discussions with Shun-ichi Amari, Philip Dawid, David Fletcher and Ting Wang. Portions of this paper appear in Parry (2013).

References

Almeida, M. P. and Gidas, B. (1993). A Variational Method for Estimating the Parameters of MRF from Complete or Incomplete Data. *The Annals of Applied Probability* **3** 103–136. MR1202518

Boero, G., Smith, J. and Wallis, K. F. (2011). Scoring rules and survey density forecasts. *International Journal of Forecasting* **27** 379–393.

Bröcker, J. and Smith, L. A. (2007). Scoring Probabilistic Forecasts: The Importance of Being Proper. *Weather and Forecasting* **22 (2)** 382–388.

Constantinou, A. and Fenton, N. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports* **8** 1.

Dawid, A. P. (1984). Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society, Ser. A,* **147** 278–292. MR0763811

Dawid, A. P. and Lauritzen, S. L. (2005). The Geometry of Decision Theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications* 22–28. University of Tokyo.

Dawid, A. P., Lauritzen, S. and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *Annals of Statistics* **40** 593–608. MR3014318

Ehm, W. and Gneiting, T. (2012). Local proper scoring rules of order two. *Annals of Statistics* **40** 609–637. MR3014319

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378. MR2345548

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29** 411–422. MR2848512

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.

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *Ann. Math. Statist.* **42** 1916–1921. MR0314430

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning* **6** 695–709. MR2249836

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis* **51** 2499–2512. MR2338984

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

Mackay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press. MR2012999

McCarthy, J. (1956). Measures of the value of information. *Proc. Nat. Acad. Sci.* **42** 654–655.

Parry, M. (2013). Multidimensional local scoring rules. In *Proceedings of the 59th ISI World Statistics Congress* 1453–1458.

Parry, M., Dawid, A. P. and Lauritzen, S. (2012). Proper local scoring rules. *Annals of Statistics* **40** 561–592. MR3014317

Suckling, E. B. and Smith, L. A. (2013). An evaluation of decadal probability forecasts from state-of-the-art climate models. *Journal of Climate* **26** 9334–9347.

Yang, E., Ravikumar, P., Allen, G. I., Baker, Y., Wan, Y. W. and Liu, Z. (2014). A General Framework for Mixed Graphical Models. *arXiv:1411.0288*.