Multidimensional local scoring rules

Matthew Parry
Dept of Mathematics & Statistics, University of Otago, Dunedin, New Zealand
e-mail: mparry@maths.otago.ac.nz

Abstract

A scoring rule is a principled way of assessing a probabilistic forecast. The key requirement of a scoring rule is that it rewards honest statements of ones beliefs. 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. In practice, almost all scoring rules can be derived from a concave entropy functional. The property of locality then 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. By using an appropriate entropy, we show that local scoring rules can be easily extended to multidimensional outcome spaces. Furthermore, we are able construct local scoring rules that are extensive, i.e. the score of independent outcomes is a sum of independent scores. Previously, only the log score was known to have this property. We end with an application of multidimensional local scoring rules to sequential data.

Keywords: entropy, normalization, log score, extensive

1. Introduction

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 $S:(\mathscr{X},\mathscr{P})\to\overline{\mathbb{R}}=[-\infty,\infty]$, where $\mathscr{X}$ is the outcome space and $\mathscr{P}$ is a set of probability distributions on $\mathscr{X}$. 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 $P\in\mathscr{P}$, we require that *(i)* $S(P,Q)$ is affine in P and *(ii)* $S(P,Q)\geq S(P,P)$ for all $P,Q\in\mathscr{P}$. The first condition means we can take $\mathscr{P}$ 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 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 $\mathscr{X}$ 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 $\mathscr{X}$ is an interval of the real line, Parry et al. (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

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

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 $\mathscr{X}$ 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 et al. (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 pseudolikelihood is a local scoring rule. Unfortunately, space constraints do not allow us to give a parallel development of local scoring rules on discrete outcome spaces, 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 $\mathscr{X}$ 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.

2. Entropy and multidimensional scoring rules

Each proper scoring rule defines a concave entropy $H(P) := S(P,P)$. Gneiting and Raftery (2007) showed that the converse almost holds (see also McCarthy (1956), Hendrickson and Buehler (1971)): if $H(P)$ is strictly) concave and $H^\star(\cdot,P) : \mathscr{X} \to \overline{\mathbb{R}}$ is a subgradient to H at $P \in \mathscr{P}$ then

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

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.

In the case $n = 1$, Parry et al. (2012) and Ehm and Gneiting (2012) considered entropies of the form

$$H(Q) = \int \mathrm{d}x\, \phi(x, q(x), q'(x)), \tag{3}$$

where $\phi(x,y,y_1)$ is differentiable in x and, for almost all $x \in \mathscr{X}$, is twice differentiable, jointly concave and 1-homogeneous 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 $\mathscr{P}$ (see Ehm and Gneiting (2012)). We omit fuller discussion of this point here.

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 $x \in \mathscr{X}$ as x^i, where $i = 1,\ldots,n$, and write q_i for $\partial q/\partial x^i$. We also let D_i denote the total derivative with respect to x^i. It follows from eq. (2) that if $\phi[y] := \phi(x^1,\ldots,x^n,y,y_1,\ldots y_n)$ is differentiable in x, and twice differentiable, jointly

(strictly) concave and 1-homogeneous in $(y, y_1, \ldots, y_n)$ then, with a slight abuse of notation,

$$S(x,Q) = \left(-D_i \frac{\partial}{\partial q_i} + \frac{\partial}{\partial q} \right) \phi[q] \tag{4}$$

is a (strictly) proper local scoring rule of second order. Note that we use the Einstein summation convention, i.e. a sum is implied over repeated indices.

The following example includes all existing 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 $\phi[q] = -\frac{1}{2} q^{-1} G^{ij} q_i q_j$ generates the proper scoring rule

$$S(x,Q) = G^{ij} \left(\frac{q_{ij}}{q} - \frac{1}{2} \frac{q_i q_j}{q^2} \right) + G^{ij}{}_{,i} \frac{q_j}{q}, \tag{5}$$

where $q_{ij} = \partial^2 q / \partial x^i \partial x^j$ and $G^{ij}{}_{,i} = \partial G^{ij} / \partial x^i$. When, additionally, $\mathscr{X}$ has a metric structure, the above scoring rule affords a covariant formulation. If $g_{ij}(x)$ is the metric tensor on $\mathscr{X}$ 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 constant):

$$S(x,Q) = g^{ij} \left(\frac{\nabla_i \nabla_j \overline{q}}{\overline{q}} - \frac{1}{2} \frac{\nabla_i \overline{q} \nabla_j \overline{q}}{\overline{q}^2} \right), \tag{6}$$

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

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. The idea of extensivity is not new but to the best of our knowledge has not been formalized before now. To avoid trivial subcases, we will assume $n > 1$ from now on.

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

$$\mathscr{I} Q = \prod_{i=1}^{n} \mathscr{M}_i Q = \prod_{i=1}^{n} Q_i, \tag{7}$$

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

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

$$S(x,Q) = \sum_{i=1}^{n} S_i(x^i, Q_i), \tag{8}$$

where $S_i(x^i, Q_i)$ are strictly proper scoring rules on $(\mathscr{X}_i, \mathscr{P}_i)$. It follows that, for $Q \in \mathscr{C}$, $S(P,Q) = \sum_{i=1}^n S_i(P_i, Q_i)$. Note that the requirement of strict propriety means eq. (8) cannot be lifted to $Q \in \mathscr{P}$, for we would have $S(P,Q) = S(P,P)$ for all $Q \in \mathscr{R}(P)$. Therefore, eq. (8) represents a simplification of $S(x,Q)$ only in the case where $Q \in \mathscr{C}$. It is worth pointing out that eq. (8) is often used when $Q \in \mathscr{P}$ 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 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 $x^{1:j} = (x^1, \ldots, x^j)$. 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 $S_i(x^i, Q_{i|1:i-1})$ are strictly scoring rules on $(\mathscr{X}_i, \mathscr{P}_i)$. *Proof*: When $Q \in \mathscr{C}$, this reduces to eq. (8) since then $Q_{i|1:i-1} = Q_i$, and strict propriety follows from the fact that

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

The logarithmic scoring rule is a member of this class since $\ln q(x) = \sum_{i=1}^n \ln q(x^i|x^{1:i-1})$. Indeed it is easy to see that the log score is essentially the only extensive score in the class of separable Bregman scores. Separable Bregman scores are of the form

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

and are (strictly) proper when $\psi(s)$ is a (strictly) concave function of $s \geq 0$. (Note that eq. (2) holds with $H(Q) = \int \mathrm{d}x\, \psi(q(x))$.) The well-known Brier score is given by $\psi(s) = -\frac{1}{2}s^2$, for example. For extensivity, we require $\psi'(s_1 \ldots s_n) = f(s_1) + \cdots + f(s_n)$, 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 $f(s_1 \ldots s_n) - f(1) = [f(s_1) - f(1)] + \cdots + [f(s_n) - f(1)]$ and the only solution to this is $f(s) - f(1) = \ln s$, up to irrelevant additive and multiplicative constants.

We now introduce the *local class* of extensive scoring rules, indeed the form of the local scoring rule eq. (4) is already reminiscent of an extensive scoring rule. 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*: When $Q \in \mathscr{C}$, $\phi_i(x^i, q, q_i) = q(x^{-i})\, \phi_i(x^i, q(x^i), q'(x^i))$, where $x^{-i} := (x^j \mid j \neq i)$, so that

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

and similarly with $(\partial/\partial q_i)\phi_i(x^i, q, q_i)$. Consequently, $S(x,Q)$ becomes a sum of strictly proper one-dimensional scoring rules, as required.

4. Application

Consider a homogeneous discrete time Markov process on the real line, observed at times $1:n$. If we model the transition probability as

$$q(x|y) = \frac{\exp(\theta f(x-\rho y))}{Z(\theta)} \tag{13}$$

then the normalization $Z(\theta)$ is typically not computable. Nevertheless, local scoring rules enable inference in such cases. Conditional on x^1, the probability of the observations is

$$q(x^{2:n}|x^1) = q(x^n|x^{n-1})q(x^{n-1}|x^{n-2})\cdots q(x^2|x^1) = \prod_{i=2}^{n} \frac{\exp(\theta f(x^i - \rho x^{i-1}))}{Z(\theta)}. \tag{14}$$

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

$$\begin{aligned} S(x,Q) &= \theta(1+\rho^2)\sum_{i=2}^{n} f''(x^i - \rho x^{i-1}) + \\ &\quad \tfrac{1}{2}\theta^2 \left\{ (1+\rho^2)\sum_{i=2}^{n} f'(x^i-\rho x^{i-1})^2 - 2\rho \sum_{i=2}^{n-1} f'(x^i-\rho x^{i-1}) f'(x^{i+1}-\rho x^i) \right\}. \end{aligned} \tag{15}$$

Then the estimating equation $(\partial/\partial\theta)S(x,Q) = 0$ is unbiased and leads to a consistent and very simple estimator for θ (see Dawid and Lauritzen (2005), for example). 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

$$\widehat{\theta} = \left(\tfrac{1}{n-1} \left\{ \sum_{i=2}^{n} (x^i - \rho x^{i-1} - \mu)^2 - \tfrac{2\rho}{1+\rho^2} \sum_{i=2}^{n-1} (x^i - \rho x^{i-1} - \mu)(x^{i+1} - \rho x^i - \mu) \right\} \right)^{-1}, \tag{16}$$

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

The author was partially supported by EPSRC Statistics Mobility Fellowship EP/E009670.

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.

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

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*, University of Tokyo, 22–28.

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

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

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

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

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

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

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