Proper local scoring rules

Philip Dawid

Joint work with Matthew Parry and Steffen Lauritzen

University of Cambridge
University of Oxford

Outline

[0]

Proper Scoring Rules

- Definition
- Examples
- Statistical inference
- Hyvärinen scoring rule

Local scoring rules

- Definition
- Variational analysis
- Characterisation
- Homogeneity
- Propriety
- Transformation of the data

Second order scoring rules

Discrete case

Scoring rules

X a random variable, values in $\mathcal{X}$.

- A *scoring rule* $S = S(x, Q)$ measures the loss You suffer if You quote a distribution Q over $\mathcal{X}$ to represent uncertainty about X, and then observe $X = x$.

- If You believe $X \sim P$, Your *expected score*, if you quote Q, is

$$S(P, Q) := \mathrm{E}_{X \sim P}\{S(X, Q)\}.$$

S is proper (w.r.t. suitable class $\mathcal{P}$ of distributions over $\mathcal{X}$) if, for $P, Q \in \mathcal{P}$, the expected score $S(P, Q)$ is minimised in Q at $Q = P$, and strictly proper if $S(P, Q) > S(P, P)$ for $Q \neq P$.

When S is proper, honesty is the best policy: If You believe $X \sim P$, You minimise Your expected score by quoting $Q = P$.

- $H(P) := S(P, P)$ is the (generalised) entropy of P
- $d(P, Q) := S(P, Q) - H(P)$ is the discrepancy/divergence between P and Q

S is proper iff $d(P, Q) \geq 0$.

Locally, $d(P, P + dP)$ defines a Riemannian metric on the set $\mathcal{P}$ of distributions over $\mathcal{X}$—decision geometry.

Log score

- $q(\cdot)$ the density of Q w.r.t. underlying measure μ
- $S(x, Q) = -\ln q(x)$
- $H(P) = -\int d\mu(y) p(y) \ln p(y)$ is the *Shannon entropy* of P
- $d(P, Q) = \int d\mu(y) p(y) \ln\{p(y)/q(y)\}$ is the *Kullback-Leibler discrepancy* $K(P, Q)$.

So S is strictly proper.

Decision metric = Fisher information metric.

NOTE: The log score has form $S(x, Q) = \xi\{x, q(x)\}$.
When $\#(\mathcal{X}) > 2$ it is essentially the only such "strictly local" proper scoring rule.

Statistical inference

- IID observations $(x_1, \ldots, x_N)$ from Q_θ: empirical distribution P_N.
- The *minimum discrepancy estimate* minimises $d(P_N, Q_\theta)$.
- Since $d(P_N, Q_\theta) = S(P_N, Q_\theta) - H(P_N)$, we can instead minimise the *total empirical score*

$$N\, S(P_N, Q_\theta) = \sum_{t=1}^{N} S(x_t, Q_\theta).$$

- This yields the *unbiased estimating equation*

$$\sum_{t=1}^{N} s(x_t, \theta) = 0$$

(where $s(x, \theta) := \partial S(x, Q_\theta)/\partial\theta$).

- Often we only know $q_\theta(\cdot)$ up to a multiplier $Z(\theta)$ that is hard to compute.
- Computation of $s(x, \theta)$ typically requires $Z(\theta)$.

Hyvärinen scoring rule

$\mathcal{X} = \mathbb{R}^k$, $\nabla := (\partial/\partial x^j)$, $\Delta = \sum_{j=1}^k \partial^2/(\partial x^j)^2$

$$S(x, Q) = \Delta \ln q(x) + \frac{1}{2} \left|\nabla \ln q(x)\right|^2 = \frac{\Delta\sqrt{q(x)}}{\sqrt{q(x)}}$$

On integrating by parts, and requiring boundary terms to vanish,

$$S(P, Q) = \frac{1}{2} \int \mathrm{d}\mu(x) \langle \nabla \ln q(x) - 2\nabla \ln p(x), \nabla \ln q(x)\rangle.$$

So

$$\begin{aligned} H(P) &= -\frac{1}{2} \int \mathrm{d}\mu(x) \left|\nabla \ln p(x)\right|^2 \\ d(P, Q) &= \frac{1}{2} \int \mathrm{d}\mu(x) \left|\nabla \ln p(x) - \nabla \ln q(x)\right|^2 \geq 0 \end{aligned}$$

- Local: $S(x, Q)$ depends only on behaviour of $q(\cdot)$ in neighborhood of realised point x
- Homogeneous: Only need $q(\cdot)$ up to scale-factor

Generalization

Carries over to a general Riemannian manifold $\mathcal{X}$:

- $p, q \mapsto$ densities with respect to natural volume measure
- $\nabla \mapsto$ natural gradient
- $\Delta \mapsto$ Laplace-Beltrami operator
- $\langle \cdot, \cdot \rangle$, $| \cdot |^2 \mapsto$ metric inner product
- integration by parts $\mapsto$ Stokes's theorem

When $\mathcal{X}$ is itself the parameter-space of a statistical model endowed with the Fisher information metric, the associated decision metric over the space of prior distributions is that arising as a limiting form of Kullback-Leibler predictive loss (Komaki, Sweeting).

Works even for improper priors!

Local scoring rules

What other proper scoring rules are local and/or homogeneous?

A scoring rule $S(x, Q)$ is local of order m if depends on the density $q(\cdot)$ of Q only through its its value and those of its first m derivatives at the realized value x of X:

$$S(x, Q) = s\left(x, q(x), q'(x), \ldots, q^{(m)}(x)\right).$$

The log score is local of order 0. It is not homogeneous

The Hyvärinen scoring rule is local of order 2. It is homogeneous.

In sequel, $\mathcal{X} = \mathbb{R}$, s is a function of $(x, q_0, q_1, \ldots, q_m)$,
$s_k := \partial s / \partial q_k$, $S_k(x, Q) := s_k\left(x, q(x), q'(x), \ldots, q^{(m)}(x)\right)$.

Variational analysis

We develop conditions on s under which, at $Q = P$, $S(P, Q)$ is stationary under arbitrary infinitesimal variations $\delta q(\cdot)$ of $q(\cdot)$ — weak propriety. This yields:

$$0 \equiv \int \mathrm{d}x \left\{ \sum_{k=0}^{m} p(x) s_k\{x, p(x), p'(x), \ldots, p^{(m)}(x)\} \delta q^{(k)}(x) + \lambda_P \delta q(x) \right\}$$

($\lambda_P =$ Lagrange multiplier for normalisation constraint).
Integrate k'th term by parts k times, assume boundary terms vanish:

$$0 \equiv \int \mathrm{d}x \, \delta q(x) \left[\sum_{k=0}^{m} (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} \{q(x) S_k(x, Q)\} + \lambda_Q \right].$$

So we want

$$\sum_{k=0}^{m} (-1)^{k+1} \frac{\mathrm{d}^k}{\mathrm{d}x^k} \{q(x) S_k(x, Q)\} \equiv \lambda_Q.$$

Operator algebra

Introduce linear differential operators

$$D := \frac{\partial}{\partial x} + \sum_{j\geq 0} q_{j+1}\frac{\partial}{\partial q_j}$$

(corresponds to total derivative d/dx); and

$$L := \sum_{k\geq 0}(-1)^{k+1} D^k\, q_0\, \partial/\partial q_k$$

If f is of order m then Df is of order $m+1$ and Lf is (potentially) of order $2m$.

Sufficient condition for weak propriety is

$$Ls \equiv \lambda.$$

Characterisation

Re-express s as a function of $(x, \ell_0, \ell_1, \dots, \ell_m)$ where generating functions $Q(z) := \sum_{k=0}^{\infty} q_k\, z^k/k!$, $L(z) := \sum_{k=0}^{\infty} \ell_k\, z^k/k!$ satisfy $L(z) = \ln Q(z)$.

Then $S(x, Q) = s\{x, \ell(x), \ell'(x), \dots, \ell^m(x)\}$, with $\ell(x) := \log q(x)$.

Note: S is homogeneous iff $\partial s/\partial \ell_0 = 0$.

In terms of the (ℓ_k),

$$
\begin{aligned}
D &= \frac{\partial}{\partial x} + \sum_{p \geq 0} \ell_{p+1} \frac{\partial}{\partial \ell_p} \\
L &= \sum_{k \geq 0} (-1)^{k+1} e^{-\ell_0} D^k e^{\ell_0} \frac{\partial}{\partial \ell_k}.
\end{aligned}
$$

We want to solve $Ls \equiv \lambda$.

Key Theorem

Theorem

$\left(L + \frac{\partial}{\partial \ell_0}\right)(1 - L) = 0.$

Corollary

If $Ls \equiv \lambda$, *then* s *is of the form*

$$s(x, \ell_0, \ldots, \ell_m) = -\lambda \ell_0 + h(x, \ell_1, \ldots, \ell_m)$$

where $Lh = 0$.

Proof.

In this case

$0 = \left(L + \frac{\partial}{\partial \ell_0}\right)(s - \lambda) = \left(L + \frac{\partial}{\partial \ell_0}\right) s = \lambda + \partial s / \partial \ell_0.$ □

Homogeneous case

From Key Theorem, any solution of $Ls = 0$ is homogeneous.
Confine attention to this case.

Theorem

$Ls = 0$ *iff* $s = (L - 1)f$ *for some homogeneous* f*.*

Proof.

Restricted to act on homogeneous functions, $L^2 = L$: so L and $1 - L$ are complementary projections. □

Corollary (Main result)

A homogeneous weakly proper local scoring rule arises iff

$$s = (L - 1)f$$

for some homogeneous f*.*

Theorem

In this case s *must be of even order.*

Propriety

Write $\phi = q_0 f$ (homogeneous of degree 1). Then $s = \Lambda\phi$ with

$$\Lambda := (L-1)q_0^{-1}\times = q_0^{-1}\sum_{k\geq 0}(-1)^{k+1}D^k\partial/\partial\ell_k$$

Integrating by parts and ignoring boundary terms yields

$$S_0(P,Q) = -\int \mathrm{d}x \sum_k p_k(x)\phi_k(x,\boldsymbol{q})$$

with $\boldsymbol{q} = (q_0, q_1, \ldots) = (q(x), q'(x), \ldots)$, which gives

$$\begin{aligned} S_0(P,P) &= -\int \mathrm{d}x\, \phi(x,\boldsymbol{p}), \\ d_0(P,Q) &= \int \mathrm{d}x\, [\phi(x,\boldsymbol{p}) - \{\phi(x,\boldsymbol{q}) + (\boldsymbol{p}-\boldsymbol{q})\nabla\phi(x,\boldsymbol{q})\}]\,. \end{aligned}$$

So long as $\phi(x,\boldsymbol{q})$ is, for each x, a [strictly] convex function of $\boldsymbol{q}$, $d_0(P,Q)$ will be [strictly] positive ($P \neq Q$). Metric is given by:

$$g(\theta) = \int \mathrm{d}x \sum_{j=1}^{m}\sum_{k=1}^{m}\phi_{jk}\,\dot{q}_{\theta,j}\,\dot{q}_{\theta,k}$$

Transformation of the data

Let $k : \mathcal{X} \to \overline{\mathcal{X}}$ be a (differentiable, invertible) transformation.

If X has distribution Q over $\mathcal{X}$ with Lebesgue density $q(\cdot)$, then the induced distribution $\overline{Q}$ for $\overline{X} := f(X)$ has density $\overline{q}(\overline{x}) = q(x)/k'(x)$ over $\overline{\mathcal{X}}$. We can define operators $\overline{D}$, $\overline{L}$ for $\overline{\mathcal{X}}$ exactly as D, L for $\mathcal{X}$.

Theorem

L is a *scalar operator*, i.e., *if $f(x, \boldsymbol{q})$ transforms as a scalar: $(\overline{f}(\overline{x}, \overline{\boldsymbol{q}}) = f(x, \boldsymbol{q}))$, then so does Lf $(\overline{Lf} = Lf)$.*

Corollary

*If $s = (L - 1)f$ defines a scoring rule S over $\mathcal{X}$, and $\overline{s} = (\overline{L} - 1)\overline{f}$ defines a scoring rule $\overline{S}$ over $\overline{\mathcal{X}}$, then $S(x, Q) = \overline{S}(\overline{x}, \overline{Q})$ (*i.e.*, the scoring rule determined by scalar f is the same, no matter how the data are expressed).*

Need deeper understanding!

Second order scoring rules

The general proper local scoring rule of order 2 has the form

$$S(x, Q) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial \phi}{\partial q_1} - \frac{\partial \phi}{\partial q_0}$$

where $\phi(x, q_0, q_1)$ is 1-homogeneous and convex in (q_0, q_1), and evaluations are at $q_0 = q(x)$, $q_1 = q'(x)$.

- Entropy: $H(P) = -\int \mathrm{d}x\, \phi(x, \boldsymbol{p})$
- Metric: $g(\theta) = \int \mathrm{d}x\, p_\theta(x)(\partial^2 F/\partial u^2)\, \dot{u}^2$

where $F = F(x, u) = \phi(x, 1, u)$ and evaluations are at $u = p'_\theta(x)/p_\theta(x)$.

For $\phi = q_1^2/2q_0$ ($F = u^2/2$) we recover the Hyvärinen rule.

Discrete case

Now let $\mathcal{X}$ be a discrete outcome space, $\mathcal{A}$ the set of positive real vectors $\boldsymbol{\alpha} = (\alpha_x : x \in \mathcal{X})$ and $\mathcal{P} = \{\boldsymbol{p} \in \mathcal{A} : \sum_x p_x = 1\}$ the set of strictly positive probability distributions on $\mathcal{X}$.

If S is a scoring rule, we can extend its domain to $\mathcal{X} \times \mathcal{A}$ by defining

$$S(x, \boldsymbol{\alpha}) := S(x, \boldsymbol{\alpha}/\alpha_+) \tag{1}$$

where $\alpha_+ := \sum_x \alpha_x$. Then S is 0-homogeneous in $\boldsymbol{\alpha}$.

Theorem

0-homogeneous S is proper if and only if it is the gradient of a concave 1-homogeneous function $H : \mathcal{A} \to \mathbb{R}$,

$$S(x, \boldsymbol{\alpha}) = [\boldsymbol{\nabla} H(\boldsymbol{\alpha})]_x.$$

Then $H(\boldsymbol{\alpha}) = \sum_x \alpha_x S(x, \boldsymbol{\alpha})$ (so $H(\boldsymbol{p}) = S(\boldsymbol{p}, \boldsymbol{p})$ is the generalised entropy of the distribution $\boldsymbol{p}$).

Locality

We describe locality in terms of an undirected graph $\mathcal{G}$. We write $x \sim y$ if $x = y$ or there is an edge between x and y, and require that $S(x, \boldsymbol{q})$ depend on $\boldsymbol{q}$ only through $(q_y : y \sim x)$.

Let $\mathcal{C}$ be the set of cliques of $\mathcal{G}$. For $C \in \mathcal{C}$, let $H_C : \mathcal{A} \to \mathbb{R}$ be a 1-homogeneous and concave function depending only on $\boldsymbol{\alpha}_C := (\alpha_j : j \in C)$. This generates a proper scoring rule $S_C(x, \boldsymbol{q})$, which will depend on $\boldsymbol{q}$ only through $\boldsymbol{q}_C$, and be non-zero only for $x \in C$. In particular it is local.

Since S_C is a 0-homogeneous function of $\boldsymbol{q}_C$, it can be computed without knowledge of the normalising constant of $\boldsymbol{q}$: at worst, we might need to compute $\sum_{j \in C} q_j$.

Extension

It follows that any scoring rule of the form

$$S(x, \boldsymbol{q}) = -\lambda \ln q_x + \sum_{C \in \mathcal{C}} S_C(x, \boldsymbol{q}) \qquad (2)$$

with $\lambda \geq 0$ and each S_C having the form described above, will be both proper and local. When $\lambda = 0$, $S(x, \boldsymbol{q})$ can be computed without knowledge of the normalising constant of $\boldsymbol{q}$.

Conjecture

Any local proper scoring rule must have the form of equation (2)*.*

Counterexample

$$\begin{aligned} \mathcal{G} &= 1—2—3 \\ S(1, q) = S(2, q) &= (1 - q_1 - q_2)^2 \\ S(3, q) &= (1 - q_3)^2 \end{aligned}$$