Local Scoring Rules: A Versatile Tool For Inference

Monica Musio

Joint work with Philip Dawid

University of Cagliari
University of Cambridge

59th WSC
August 25-30 2013, Hong Kong
Scoring rules

- Suppose **You** are required to quote a probability distribution \( Q \) for a quantity \( X \)

- **Nature** revels the value \( x \in X \)
  
  A scoring rule \( S = S(x, Q) \) is a loss function, intended to measure the quality of your quoted distribution \( Q \) over \( X \) in the light of the realised outcome \( X = x \)

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

\[
S(P, Q) = \mathbb{E}_{X \sim P} \{ S(X, Q) \}
\]

- \( S \) is proper 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 \( Q \neq P \)

- In such case **honesty is the best policy**: You are motivated to quote what **You** actually believe
Related concepts

- $H(P) = S(P, P)$ is the Entropy of $P$
  - $H(P)$ a concave functional of $P$

- $D(P, Q) = S(P, Q) - H(P)$ is the Discrepancy/Divergence between $P$ and $Q$
  - $S$ is proper iff $D(P, Q) \geq 0$
Example: Brier score (Brier 1950)

- \( \mathcal{X} \) discrete and \( q(x) = Q(X = x) \)

\[
S(x, Q) = \sum_y (\delta_{x,y} - q(y))^2 = -2q(x) + \sum_y q(y)^2 + 1
\]

where \( \delta_{x,y} = 1 \) if \( x = y \) and \( \delta_{x,y} = 0 \), otherwise

Then

\[
S(P, Q) = \sum_y p(y)\{1 - p(y)\} + \sum_y \{q(y) - p(y)\}^2
\]

\[
= H(P) + D(P, Q)
\]

— minimised for \( Q = P \) \( \rightarrow \) strictly proper scoring rule

The divergence function is the squared Euclidean distance
Example: log score (Good 1952)

- $q(\cdot)$ the density of $Q$ w.r.t. underlying measure $\mu$
- $S(x, Q) = -\ln q(x)$

- $H(P) = -\int d\mu(y) \cdot p(y) \ln p(y)$ is the Shannon Entropy of $P$

- $D(P, Q) = \int d\mu(y) \cdot p(y) \ln \{p(y)/q(y)\}$ is the Kullback-Leibler discrepancy $K(P, Q)$

$S$ is local of order 0 $\rightarrow S(x, Q)$ depends on $q(\cdot)$ only through its value at the observation $x$
Example: Hyvärinen score (2005)

\[ \chi = \mathbb{R}^k, \nabla := (\partial / \partial x^j), \Delta = \sum_{j=1}^{k} \partial^2 / (\partial x^j)^2 \]

\[ S_H(x, Q) = \Delta \ln q(x) + \frac{1}{2} \| \nabla \ln q(x) \|^2 = \frac{\Delta \sqrt{q(x)}}{\sqrt{q(x)}} \]

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

\[ S(P, Q) = \frac{1}{2} \int d\mu(y) \cdot \langle \nabla \ln q(y) - 2\nabla \ln p(y), \nabla \ln q(y) \rangle \]

\[ H(P) = -\frac{1}{2} \int d\mu(y) \cdot \| \nabla \ln p(y) \|^2 \]

\[ D(P, Q) = \frac{1}{2} \int d\mu(y) \cdot \| \nabla \ln p(y) - \nabla \ln q(y) \|^2 \geq 0 \]

- **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 — can ignore normalising constant
Example: Composite scores

- Observable variable $X$ (typically a vector)

- $\{X_k\}$ collection of marginal and/or conditional variables, and $S_k$ a proper scoring rule for $X_k$

$$S(x, Q) = \sum_k S_k(x_k, Q_k)$$

$X_k \sim Q_k$ when $X \sim Q$

We term a scoring rule of this form a **composite score**

- When each $S_k$ is the log score, we obtain the log composite-likelihood (see e.g. Statistica Sinica (2011))

- We can treat composite likelihood as a proper scoring rule, rather than an approximation to the true likelihood
Example: Pseudo-likelihood

- **Spatial process** \( X = (X_v : v \in V) \), where \( V \) is a set of lattice sites
- **\( Q \) joint distribution for \( X \)**
- \( \{Q_v : v \in V\} \) family of conditional distributions for \( X_v \), given the values of \( X \backslash v \)
- Suppose \( Q \) is Markov \( \rightarrow Q_v \) depends only on \( X_{ne(v)} \)

\[
S(x, Q) = \sum_v S_0(x_v, Q_v)
\]

(Dawid et al. 2012, Dawid and Musio 2013)

- \( S_0 \) proper scoring rule for the state at a single site
- \( S_0 \) log score \( \rightarrow \) (negative log) pseudo-likelihood of Besag (1975)
- \( X_v \) binary and \( S_0 \) quadratic (“Brier”) score \( \rightarrow \) method of ratio matching (Hyvärinen (2005))
Statistical inference

- \{ P_\theta : \theta \in \Theta \} parametric family of distribution on \( \mathcal{X} \), \( \Theta \in \mathbb{R}^p \)
- i.i.d. observations \((x_1, \ldots, x_N)\) from \( P_\theta \): empirical distribution \( \hat{P}_N \)
- Estimate \( \theta \) by that value minimising \( D(\hat{P}_N, P_\theta) \)
- Since \( D(\hat{P}, P_\theta) = S(\hat{P}, P_\theta) - S(\hat{P}, \hat{P}) \), we minimise \( nS(\hat{P}, P_\theta) \) which is the total empirical score \( \sum_{i=1}^n S(x_i, P_\theta) \)

\[ \hat{\theta}_S = \arg\min_\theta \sum_{i=1}^n S(x_i, P_\theta) \]

\( \hat{\theta}_S \), the minimum score estimate of \( \theta \), is the root of the score equation

\[ \sum_{i=1}^n s(x_i, \theta) = 0 \]

where \( s(x, \theta) = \nabla_\theta S(x, P_\theta) = \left( \frac{\partial S(x, P_\theta)}{\partial \theta_j} : j = 1, \ldots, p \right) \)
Theory

- \( S(x, Q) = -\log q(x) \)
  - score equation = - likelihood equation
  - minimum score estimate = maximum likelihood estimate

- The score equation will always yield an unbiased estimating equation (Dawid and Lauritzen (2005), Dawid (2007))

\[
E_\theta s(X, \theta) = 0
\]

- We can apply standard results on unbiased estimating equations to describe the properties of the minimum score estimator \( \hat{\theta}_S \)

- \( \hat{\theta}_S \) is typically consistent in repeated i.i.d. sampling (but not necessarily efficient)
Theory

- Suppose $X_1, X_2, \ldots, X_n$ i.i.d. as $P_\theta$. Under some regularity conditions
  \[
  \hat{\theta}_S \approx N(\theta, \{nG(\theta)\}^{-1})
  \]
  where
  \[
  G(\theta) = H(\theta) J(\theta)^{-1} H(\theta)
  \]
is the Godambe information matrix

\[
\begin{align*}
J(\theta) &= \mathbb{E}_\theta \left\{ s(X, \theta) s(X, \theta)^T \right\} \\
H(\theta) &= \mathbb{E}_\theta \left\{ \nabla_\theta s(X, \theta)^T \right\}
\end{align*}
\]

- If $S(x, Q) = -\log q(x)$, we have $H(\theta) = J(\theta)$ the Fisher information matrix
Likelihood

Consider the unbiased estimating equation

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

where $s(x, \theta) = \partial S(x, P_\theta) / \partial \theta$

If we use the log score, we obtain the usual likelihood equation

$$\frac{\partial \ln p(x \mid \theta)}{\partial \theta} = 0$$

Often we only know $P_\theta$ up to a multiplier

$$p(x \mid \theta) \propto f(x \mid \theta)$$

In this case we need to compute and differentiate the normalising factor

$$Z(\theta) = \int f(x \mid \theta) \, dx$$

Likewise for Brier, etc. estimating equation

This can be problematic especially for highly complex models
Likelihood

- Two methods for dealing with this problem:
  - Use approximations to $Z(\theta)$
  - Avoid the use of likelihood function and introduce composite likelihood

- We propose to use homogeneous scoring rule:
  A function $f : \mathcal{A} \rightarrow \mathbb{R}$ is called homogeneous, if

$$ f(\lambda \alpha) = f(\alpha), \quad \text{for all} \quad \lambda > 0 $$

Characterization theorem for homogeneous proper scoring rules provided by Dawid et al. (2012)

- log score is not homogeneous
- Hyvärinen score is homogeneous: Only need $q(\cdot)$ up to scale-factor

Homogeneous scoring rules avoid the need to compute normalising constants!
Example: AR1 model

\[ Y \sim \mathcal{N}(\mathbf{0}, \Phi^{-1}) \]

\[ \Phi (N \times N) = \begin{pmatrix}
\alpha & \beta & 0 & 0 & \cdots & 0 \\
\beta & \alpha & \beta & 0 & \cdots & 0 \\
0 & \beta & \alpha & \beta & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & \beta & \alpha 
\end{pmatrix} \]

Full likelihood proportional to

\[ \text{det}(\Phi)^{1/2} \exp\left(-\frac{1}{2} \mathbf{y}^T \Phi \mathbf{y}\right) \]

with

\[ \text{det}(\Phi) = \beta^N \frac{\rho^{N+1} - \rho^{-(N+1)}}{\rho - \rho^{-1}} \]

where \( \rho + \rho^{-1} = \alpha/\beta \)

- Hard to maximise directly
The multivariate Hyvärinen score eliminates the problematic normalising constant, and is a simple quadratic

\[-N\alpha + \frac{1}{2} \sum_{r=1}^{N} (\alpha y_r + \beta z_r)^2\]

where \(z_r = y_{r-1} + y_{r+1}\) (taking \(y_{-1} = y_{N+1} = 0\))

- Easy to minimise

With \(\lambda = -\beta / \alpha\), we have

\[-N\alpha + \frac{1}{2} \alpha^2 \sum (y_r - \lambda z_r)^2\]
The global minimum is at

\[
\hat{\lambda} = \frac{\sum y_r z_r}{\sum z_r^2}
\]

\[
\hat{\alpha}^{-1} = \frac{\sum y_r^2 - (\sum y_r z_r)^2}{\sum z_r^2 N}
\]

**Pseudo-likelihood:** The full conditionals are given by

\[Y_r|\text{rest} \sim \mathcal{N}\left(\lambda z_r, \alpha^{-1}\right) \quad (r = 1, \ldots, n)\]

The log pseudo-likelihood is thus, up to a constant:

\[
\frac{1}{2} N \log \alpha - \frac{1}{2} \alpha \sum (y_r - \lambda z_r)^2
\]

- It gives the same estimates as for Hyvärinen score
Wishart model

\( \nu \) independent vectors distributed as \( \mathcal{N}(\mathbf{0}, \Phi^{-1}) \)

- We could form an estimating equation by summing those for the individual vectors

- A more efficient method is as follows:

A sufficient statistic for this model is the sum-of-squares-and-products matrix

\[ S \sim W_N(\nu; \Phi^{-1}) \]

the multivariate Hyvärinen score based on \( \{s_{ij} : 1 \leq i \leq j \leq N\} \) is

\[
h(S, \Phi) = \sum_i \frac{\partial^2 l(S|\Phi)}{\partial s_{ii}^2} + \sum_{i<j} \frac{\partial^2 l(S|\Phi)}{\partial s_{ij}^2} + \frac{1}{2} \sum_i \left( \frac{\partial l(S|\Phi)}{\partial s_{ii}} \right)^2 + \\
+ \frac{1}{2} \sum_{i<j} \left( \frac{\partial l(S|\Phi)}{\partial s_{ij}} \right)^2\]
where \( l(S|\Phi) \) is the log-density, given, up to a constant independent of \( S \), by

\[
l(S|\Phi) = c \log \det(S) - \frac{1}{2} \text{tr}(\Phi S')
\]

\[
= c \log \det(S) - \frac{1}{2} \sum_i \phi_{ii}s_{ii} - \sum_{i<j} \phi_{ij}s_{ij},
\]

where \( c = \frac{1}{2}(\nu - N - 1) \)

After some algebra we get

\[
h(S|\Phi) = -c \left\{ \sum_{i,j} \left( s_{ij} \right)^2 + \left( \sum_i s_{ii} \right)^2 - \sum_i \left( s_{ii} \right)^2 \right\} +
\]

\[
+ \frac{1}{2} \left\{ \sum_i \left( cs_{ii} - \frac{1}{2} \phi_{ii} \right)^2 + \sum_{i<j} \left( 2cs_{ij} - \phi_{ij} \right)^2 \right\}
\]
The associated estimate of $\Phi$ is thus obtained by minimising

$$
\sum_i (2cs^{ii} - \phi^{ii})^2 + 4 \sum_{i<j} (2cs^{ij} - \phi^{ij})^2
$$

over the parameter-space for $\Phi$

- If $\Phi$ is totally unrestricted this gives the unbiased estimate
  \[ \hat{\Phi} = (\nu - N - 1)S^{-1} \]

- For $\Phi$ tridiagonal, we get the unbiased estimates
  \[ \hat{\alpha} = \frac{\nu - N - 1}{N} \sum_{i=1}^{N} s^{ii} \]
  \[ \hat{\beta} = \frac{\nu - N - 1}{N - 1} \sum_{i=1}^{N-1} s_{i,i+1} \]
Bayesian Model Selection

- Models $\mathcal{P} = \{P_\theta\}, \ Q = \{Q_\phi\}$.
- Within-model priors $\pi(\theta), \pi(\phi)$
- Bayes Factor $BF = p(x)/q(x)$ with

\[
\begin{align*}
p(x) &= \int p(x|\theta)\pi(\theta)d\theta \\
q(x) &= \int q(x|\phi)\pi(\phi)d\phi
\end{align*}
\]

- Some sensitivity to specification of $\pi(\theta), \pi(\phi)$
- $BF$ is not well-defined when improper prior distributions are used
  - sensitivity to overall scale
Alternative approach

- **logBF** compares log-scores for predictive distributions \( P \) and \( Q \)
- If we replace log-score by a homogeneous scoring rule — e.g., Hyvärinen score \( S_H \) — the problem with unspecified scale factors disappears!

**Example: Exponential family:**
Suppose the statistical model is an exponential family with natural statistic \( X \):

\[
p(x \mid \theta) = \exp \left\{ a(x) + b(\theta) + \theta^T x \right\}
\]

Let \( \mu(x) \) and \( \Sigma(x) \) be the posterior mean-vector and dispersion matrix of \( \Theta \), given \( X = x \). Then

\[
S_H(x, P) = 2 \Delta a(x) + \| \mu(x) + \nabla a(x) \|^2 + 2 \text{tr} \Sigma(x)
\]

- just requires posterior to be proper
Bibliography


