Proper Scoring Rules in Bayesian Model Selection

Padova, February 3 2016

Monica Musio$^1$
joint work with Philip Dawid$^2$

---

$^1$Università degli Studi di Cagliari
$^2$University of Cambridge
Outline

Proper scoring rules

▶ Definition
▶ Examples
▶ Properties and characterizations

Model Selection

▶ General problem
▶ Bayesian Model Selection
▶ Alternative approach
▶ Examples

Conclusions
Example: Weather Forecaster

\[ X = \{ \text{rain tomorrow?} \} \hspace{1cm} (0/1) \]

- You quote a probability \( q \) for \( X = 1 \) that can be different from what you actually believe, \( p \).

- Your penalty is assessed according to the rule

\[ S(x, q) = (x - q)^2. \]

- Your expected loss is

\[ S(p, q) = E_p \{(X - q)^2\} = (p - q)^2 + p(1 - p) \]

- Uniquely minimised for \( q = p \) → strictly proper scoring rule.
**Ex ante: to motivate honesty** On each occasion, You will minimise your expected penalty by honestly divulging what You actually believe.

**Ex post: to measure performance** If you have made a sequence of forecasts, \((q_i : i = 1, \ldots, n)\), and the outcomes are \((x_i)\), Your performance can be assessed by your total realised penalty score:

\[
S_+ := \sum_{i=1}^{n} (x_i - q_i)^2.
\]

We can use this to compare different forecasting systems.

The use of scoring rules has no connection with the source of probabilities.
A scoring rule indicates how much the participant is to be penalised depending on the “distance” between the assessed probabilities and the actual result. The simplest and most practical scoring rule is Brier’s rule; if someone indicates as his own opinion $P(E) = p$, the score (i.e. the penalisation) is the square of the distance between forecast and result: $p^2 = (p - 0)^2$ if the result is 0 ($E$ does not happen), and $(1 - p)^2$ if the result is 1 ($E$ does happen). The fact that Brier’s rule is a proper one is proved since for a person indicating as his probability assessment $\bar{p}$ different from his own effective opinion $p$, expected penalisation is increased by $(p - \bar{p})^2$. 
Scoring rules

\( X \) a random variable, values in \( \mathcal{X} \)
\( \mathcal{P} \) family of distributions over \( \mathcal{X} \)

- A scoring rule is a function

\[
S : \mathcal{X} \times \mathcal{P} \rightarrow \mathbb{R}
\]

\( S = S(x, Q) \) measures the quality of the probability distribution \( Q \) for the random variable \( X \) given its observed result \( x \).

- The expected score w.r.t. \( P \) is:

\[
S(P, Q) := E_{X \sim P}\{S(X, Q)\}
\]

- \( S \) 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 \( Q \neq P \)
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 $Q$
  - $D(P, Q) \geq 0$
  - $D(P, Q) - D(P, Q_0)$ an affine function of $P$
Example: Brier score (1950)

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

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

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 \( Q = P \to \text{strictly proper scoring rule} \)

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 $K(P, Q)$.

So $S$ is strictly proper.

**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 (Bernardo 1979).
Example: Hyvärinen score (2005)

\[ \mathcal{X} = \mathbb{R}^k, \mu = \text{Lebesgue} \]

For \( k = 1 \)

\[ S(x, Q) := \frac{q''(x)}{q(x)} - \frac{1}{2} \left\{ \frac{q'(x)}{q(x)} \right\}^2 \]

In general

\[ \nabla = (\partial/\partial x^i), \quad \Delta = \sum_{j=1}^{k} \partial^2 / (\partial x^j)^2 \]

\[ S(x, Q) := \Delta \ln q(x) + \frac{1}{2} | \nabla \ln q(x) |^2 = \frac{\Delta \sqrt{q(x)}}{\sqrt{q(x)}} \]
Example: Hyvärinen score

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:** $S(\cdot, Q)$ 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.
Multivariate Normal distribution

\[ \mathbf{X} \sim \mathcal{N}_k(\mathbf{\mu}, \Phi^{-1}) \]

with density

\[ q(\mathbf{x}) \propto \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{\mu})^T \Phi (\mathbf{x} - \mathbf{\mu}) \right\} \]

We have

\[ \nabla \log q = -\Phi (\mathbf{x} - \mathbf{\mu}) \]
\[ \Delta \log q = -\text{tr} \Phi \]

so that

\[ S_H(\mathbf{x}, Q) = \| \Phi (\mathbf{x} - \mathbf{\mu}) \|^2 - 2 \text{ tr} \Phi. \]

- don’t need to compute \( \text{det} \Phi \)
- works even if \( \Phi \) is singular!
Multivariate Normal distribution

E. g., $\Phi$ a projection matrix of rank $r$:

$$S_H(x, Q) = (x - \mu)^T \Phi (x - \mu) - 2r$$

For the univariate case $Q = \mathcal{N}(\mu, \sigma^2)$ we have

$$S_H(x, Q) = \frac{1}{\sigma^2} \left\{ \left( \frac{x - \mu}{\sigma} \right)^2 - 2 \right\}$$
Local scoring rules

What other proper scoring rules are local and/or 0-homogeneous?

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

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

A function \( f : \mathcal{A} \to \mathbb{R} \) is called \( \alpha \)-homogeneous if

\[
f(\lambda x) = \lambda^\alpha f(x), \quad \text{for all } \lambda > 0
\]

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

The Hyvärinen scoring rule is local of order 2. It is 0-homogeneous.
Local scoring rules

\[ \mathcal{X} = \mathbb{R} \]

Characterization theorem for local proper scoring rules provided by Parry et al. (2012).

Excluding the log score, every proper local scoring rule is a linear combination of the log score and a key local scoring rule, having the form

\[
S(x, Q) = \sum_{k=0}^{t} (-1)^k \frac{d^k}{dx^k} \phi[k] \left\{ x, q(x), q'(x), \ldots, q^{(t)}(x) \right\}
\]

where \( \phi(x, q_0, \ldots, q_t) \) is a function 1-homogeneous and concave in \( (q_0, \ldots, q_t) \) for each fixed \( x \) and \( \phi[k] := \partial \phi / \partial q_k \).

Then (subject to boundary conditions) \( S \) is a homogeneous, local proper scoring rule of even order \( \leq 2t \).

For \( \phi = -q_1^2 / 2q_0 \) we recover the Hyvärinen rule.
Choose between (unparametrised) probability models $P_j$ for $X^n := (X_1, X_2, \ldots, X_n)$, based on observed data $x^n, p^n_j(x^n)$ joint density for $X^n$

Typically use likelihood: $p^n_j(x^n)$

Equivalently, use log score: $-\log p^n_j(x^n)$

- measures how badly $P_j$ did at forecasting $X^n = x^n$

$p^n$ can be decomposed into its sequence of recursive conditionals:

$$p^n(x^n) = p_1(x_1) \times p_2(x_2) \times \ldots \times p_n(x_n)$$

where $p_i$ is the density of $X_i$, given $X^{i-1} = x^{i-1}$. 
Model selection

- When we observe $X_{i+1} = x_{i+1}$, the associated logarithmic score is $S(x_{i+1}, P_{i+1}) = -\log p(x_{i+1} \mid x^i)$

- The cumulative score is:

$$-\log p_j^n(x^n) = -\log \prod_{i=1}^{n} p_j(x_i \mid x^{i-1}) = \sum_i -\log p_j(x_i \mid x^{i-1})$$

- measures prequential (predictive sequential) loss of forecasting each $X_i$ in turn.
Model selection

- Could use some other proper scoring rule in place of log score

- In such case we assume now that data either arrive in a time order stream (sequential forecasting system), or can be arranged into such form:

\[
X = (X_1, X_2, \ldots ).
\]

- **Multivariate**: \( S \left( x, p_j(\cdot) \right) \)

- **Prequential**: \( \sum_i S_i \left( x_i, p_j(\cdot \mid x^{i-1}) \right) \)

  apply a proper scoring rule \( S_i \) for the \( i \)-th term and cumulate the scores.
Model selection

- When $S$ is the log score the cumulate prequential score is just the overall multivariate log score;

- In such case it insensitive to the ordering of the data;

- Not generally equivalent;

- for other scores there may be some sensitivity to ordering.

- Under some regularity conditions, minimising the prequential score will be a consistent model selection criterion (the criterion will choose the correct model with probability 1) (Dawid 1992).
Bayesian Model Selection

- Set of candidate models $M_k$ with probability densities $p_k(x | \theta_k)$.

- $\pi_k(\theta_k)$ prior distribution for parameter vector $\theta_k$ under model $M_k$.

- Choose the model with the largest posterior probability for $x = (x_1, \cdots, x_n)$

$$P(M_k | x) \propto P(M_k) \int p_k(x | \theta_k) \pi_k(\theta_k) d\theta_k.$$ 

- Predictive density function $p_k(x) = \int p_k(x | \theta_k) \pi_k(\theta_k) d\theta_k$ of the data $x$ under the model $M_k$.

- It can be seen as a hybrid between an objective component, $p_k(x | \theta_k)$, and a subjective component $\pi_k(\theta_k)$. 
Bayesian Model Selection

- Models $M_1 = \{P_\theta\}$, $M_2 = \{Q_\phi\}$.
- Within-model priors $\pi(\theta)$, $\pi(\phi)$
- Bayes Factor $BF = p(x)/q(x)$ with

\[
p(x) = \int p(x | \theta)\pi(\theta) d\theta
\]
\[
q(x) = \int q(x | \phi)\pi(\phi) d\phi
\]

- $P(M_1 | x) / P(M_2 | x) = BF \times \frac{P(M_1)}{P(M_2)}$

- with equals prior model probabilities, $BF = \text{Posterior odds.}$
Bayesian Model Selection

- Some sensitivity to specification of $\pi(\theta)$, $\pi(\phi)$.

- But problems if we try to use improper priors

\[
\pi(\theta) \propto h(\theta) \quad \text{where} \quad \int h(\theta)d\theta = \infty
\]

You can also use $w(\theta) = C\pi(\theta)$ as a prior, $C$ arbitrary positive constant

sensitivity to overall scale — which is not determined

In the literature a variety of pseudo-Bayes factors have been proposed to convert improper objective priors into proper distributions (Intrinsic, Partial and Fractional BF) (Berger and Pericchi 1996, O’Hagan 1995, Moreno Bertolino Racugno (1998)).
Alternative approach

- log $BF$ compares log-scores for predictive distributions $P$, $Q$ in the light to the observed data $x$:

$$\log BF = \log p(x) - \log q(x)$$

- $S(x, Q) = -\log p(x)$ can be interpreted as a measure of how badly $Q$ dis at forecasting the outcome $x$.

- If we replace log-score by a homogeneous scoring rule — e.g., Hyvärinen score $S_H$ in the continuous case — the problem with unspecified scale factors disappears!

- OK in general so long as predictive density is finite (not necessarily proper) — then we have proper posterior
Example: Exponential family

Suppose the statistical model is an exponential family with natural statistic $X$:

$$p(x | \theta) = \exp \left\{ a(x) + b(\theta) + \theta^T x \right\}$$

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

Then

$$S_H(x, P) = 2 \Delta a(x) + \| \mu(x) + \nabla a(x) \|^2 + 2 \text{tr} \Sigma(x)$$
Normal linear model: Variance known

- Y ~ \mathcal{N}(X\theta, \sigma^2 I),

- Uniform prior for \theta

- Multivariate Hyvärinen score:

\[
S_H = \frac{1}{\sigma^2} \left( \frac{R}{\sigma^2} - 2\nu \right) \quad (\nu \geq 0)
\]

where R [resp., \nu] is the residual sum-of-squares [resp., degrees of freedom]

- For fixed \sigma^2, proportional to AIC
  - not consistent
Prequential application

Apply univariate Hyvärinen score to sequence of predictive distributions:

\[ Y_i \mid (Y_1, \ldots, Y_{i-1}) \sim \mathcal{N}(\eta_i, k_i^2 \sigma^2) \quad (i > p) \]

with

\[ \eta_i := x_i^T \hat{\theta}_{i-1} \]
\[ k_i^2 := 1 + x_i^T \left\{ (X^{i-1})^T (X^{i-1}) \right\}^{-1} x_i \]

The incremental score is

\[ S_i = \frac{1}{k_i^2 \sigma^2} \left( \frac{Z_i^2}{\sigma^2} - 2 \right) \quad (i > p) \]

where \( Z_i = (Y_i - \eta_i)/k_i \sim \mathcal{N}(0, \sigma^2) \) (so \( \sum_{i=p}^{n} Z_r^2 = R \))

- Minimise cumulative prequential score
  - consistent model choice
- To avoid start-up problems, assess first \( p_{\text{max}} \) observations with multivariate score, then proceed incrementally
Normal linear model: Variance unknown

With standard improper prior \( \pi(\theta, \sigma^2) \propto 1/\sigma^2 \), we get:

- **Multivariate score:**
  \[
  S_H = -\frac{(\nu - 4)}{\hat{\sigma}^2}
  \]
  where \( \hat{\sigma}^2 = R/\nu \)

- **Incremental prequential score:**
  \[
  S_i = \frac{\nu_i \{(4 + \nu_i) Z_i^2 - 2R_i\}}{k_i^2 R_i^2} \quad (i > p)
  \]
  where \( R_i = \sum_{r=p}^{i} Z_r^2 \) [resp., \( \nu_i \)] is the residual sum-of-squares [resp., degrees of freedom] based on \( y^i \)
  - consistent model choice
Conclusion

- Proper scoring rules form a versatile alternative to likelihood for assessing model performance
  - Can be tailored to actual decision problem at hand

- Prequential application typically gives a consistent model selection criterion

- Homogeneous scoring rules are not fazed by the indeterminate normalization of improper distributions

- Further work required to explore theory, applications, computation, ...

- Extend to discrete random variables.
THANK YOU!
References


