Proper Scoring Rules

Cheng-Han Yu

Department of Applied Math and Statistics
University of California, Santa Cruz

May 26, 2016
1 Review of Scoring Rules

2 Scoring Rules for Categorical Variables

3 Scoring Rules for Continuous Variables

4 R Package ensembleBMA
Probabilistic forecasting and scoring rules

- One purpose of statistics is to make forecasts and provide suitable measures of the uncertainty associated with them. Consequently, forecasts should be **probabilistic in nature**, taking the form of probability distributions over future quantities or events (Dawid 1984).
Probabilistic forecasting and scoring rules

• Scoring rules assess the **quality of probabilistic forecasts**, by assigning a numerical score based on the **predictive distribution** and on the **observed event or outcome**.

• **Elicitation**: encourages the forecaster to make careful assessments and to be honest.

• **Evaluation**: measures the quality of the probabilistic forecasts, reward probability assessors for forecasting jobs, and rank competing forecast procedures (**forecast verification**).
Notations of Scoring Rules

- Gneiting and Raftery (2007) views scoring rules as utility functions of the predictive distribution and event, and so an expected scoring rule is a positively oriented reward that a forecaster wishes to maximize.

- If a forecaster quotes the predictive distribution $P$ and the event $x$ realizes, then his reward is $S(P, x)$.

- $S(P, Q)$ denotes the expected value of $S(P, \cdot)$ under distribution $Q$.

- Suppose that the forecaster’s best judgment is $Q$. The forecaster is encouraged to quote his true belief, $P = Q$, if $S(Q, Q) \geq S(P, Q)$ for any $P \neq Q$ with equality if and only if $P = Q$. It is said to be strictly proper. If $S(Q, Q) \geq S(P, Q)$ for all $P$ and $Q$, then the scoring rule is said to be proper.
Probabilistic forecasts of a categorical variable

- **Sample space:** $\Omega = \{1, \ldots, m\}$.

- **Probabilistic forecast:** $p = (p_1, \ldots, p_m) \in \mathcal{P}$, where $\mathcal{P} = \{p = (p_1, \ldots, p_m) : p_1, \ldots, p_m \geq 0, p_1 + \cdots + p_m = 1\}$.

- **Scoring rule:** The reward is $S(p, i)$ if the forecaster quotes $p$ and the value $i$ occurs.

- **Expected scoring rule under $q$:** $S(p, q) = \sum_{i=1}^{m} S(p, i)q_i$, $q \in \mathcal{P}$.

- **Entropy under $q$:** $H(q) := -\sup_{p \in \mathcal{P}} S(p, q) = -S(q, q)$ if proper.

- **Divergence function:** $d(p, q) := S(q, q) - S(p, q)$

- **Bayes act:** $a_p := \arg \sup_{a \in \mathcal{A}} E[U(\theta, a)]$ for $\theta \sim p \in \mathcal{P}$

- **$S(p, i)$ is a proper rule, because**

  $$S(q, q) = \sum_{i} U(i, a_q)q_i \geq \sum_{i} U(i, a_q)p_i = S(p, q).$$
Proper scoring rules for categorical variables

**Definition (Regular)**

A scoring rule $S$ for categorical forecasts is **regular** if $S(\cdot, i)$ is real-valued for $i = 1, \ldots, m$, except possibly that $S(p, i) = -\infty$ if $p_i = 0$.

**Definition (Subgradient)**

Let $G : \mathcal{P} \rightarrow \mathbb{R}$ be a **convex** function. A vector $G'(p) = (G'_1(p), \ldots, G'_m(p))$ is a **subgradient** of $G$ at the point $p \in \mathcal{P}$ if

$$G(q) \geq G(p) + \langle G'(p), q - p \rangle$$

for all $q \in \mathcal{P}$. If $G$ is differentiable at the interior point $p \in \mathcal{P}$, then $G'(p)$ is unique and equals the **gradient** of $G$ at $p$.

Assume $G'_j(p) \in \mathbb{R}$ and permit $G'_i(p) = -\infty$ if $p_i = 0$. 
**Theorem (Proper scoring rules)**

A regular scoring rule \( S \) for categorical forecasts is (strictly) proper if and only if

\[
S(p, i) = G(p) - \langle G'(p), p \rangle + G'_i(p), \quad i = 1, \ldots, m, \tag{2}
\]

where \( G: \mathcal{P} \rightarrow \mathbb{R} \) is a (strictly) convex function and \( G'(p) \) is a subgradient of \( G \) at the point \( p \), for all \( p \in \mathcal{P} \).

**Proof.**

(\( \Leftarrow \)) Suppose \( S(p, i) = G(p) - \langle G'(p), p \rangle + G'_i(p), i = 1, \ldots, m \). Note that

\[
S(q, q) = \sum_{i=1}^{m} S(q, i) q_i = G(q) - \langle G'(q), q \rangle + \sum_{i=1}^{m} G'_i(q) q_i = G(q)
\]

\[
S(p, q) = \sum_{i=1}^{m} S(p, i) q_i = G(p) - \langle G'(p), p \rangle + \sum_{i=1}^{m} G'_i(p) q_i = G(p) + \langle G'(p), q - p \rangle.
\]

Using the fact (1), we have \( S(q, q) \geq S(p, q) \), and so \( S \) is proper.
Theorem (Proper scoring rules)

A regular scoring rule $S$ for categorical forecasts is (strictly) proper if and only if

$$S(p, i) = G(p) - \langle G'(p), p \rangle + G'_i(p), \quad i = 1, \ldots, m,$$

where $G : \mathcal{P} \to \mathbb{R}$ is a (strictly) convex function and $G'(p)$ is a subgradient of $G$ at the point $p$, for all $p \in \mathcal{P}$.

Proof.

($\Rightarrow$) Suppose $S$ is proper. Define $G : \mathcal{P} \to \mathbb{R}$ by

$$G(q) = S(q, q) = \sup_{p \in \mathcal{P}} S(p, q) \text{ (why convex?)}$$

Then (1) holds with $G'_i(p) = S(p, i)$ because

$$G(p) + \langle G'(p), q \rangle - \langle G'(p), p \rangle = S(p, p) + \sum S(p, i)q_i - \sum S(p, i)p_i = S(p, q),$$

and $G(q) \geq S(p, q) = G(p) + \langle G'(p), q - p \rangle$. Hence $G'(p)$ is a subgradient of $G$.

In addition,

$$G(p) - \langle G'(p), p \rangle + G'_i(p) = G(p) - \sum S(p, i)p_i + G'_i(p) = S(p, i) \qed$$
A regular scoring rule $S$ is proper if and only if the expected score function $G(p) = S(p, p)$ is convex on $\mathcal{P}$, and the vector with components $S(p, i)$ for $i = 1, \ldots, m$ is a subgradient of $G$ at the point $p$, for all $p \in \mathcal{P}$. 
Examples of scoring rules of a categorical variable

- Suppose $G(p) = \sum_{j=1}^{m} \log(p_j) p_j$. Then
  
  $$G_i'(p) = (1/p_i)(p_i) + \log(p_i) = \log(p_i) + 1$$

  $$\langle G'(p), p \rangle = \sum_i (\log(p_i) + 1)p_i = \sum_i \log(p_i) p_i + \sum_i p_i$$

  $$= \sum_i \log(p_i) p_i + 1$$

  $$S(p, i) = G(p) - \langle G'(p), p \rangle + G_i'(p)$$

  $$= \sum_{j=1}^{m} \log(p_j) p_j - \left( \sum_i \log(p_i) p_i + 1 \right) + \log(p_i) + 1 = \log(p_i)$$

  $$d(p, q) = S(q, q) - S(p, q) = \sum_{j=1}^{m} \log(q_j) q_j - \sum_{j=1}^{m} \log(p_j) q_j$$

  $$= \sum_{j=1}^{m} \log \left( \frac{q_j}{p_j} \right) q_j$$
### Examples of scoring rules of a categorical variable

<table>
<thead>
<tr>
<th></th>
<th>Quadratic</th>
<th>Spherical</th>
<th>Logarithmic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G(p)$</td>
<td>$\sum_{j=1}^{m} p_j^2 - 1$</td>
<td>$(\sum_{j=1}^{m} p_j^2)^{1/2}$</td>
<td>$\sum_{j=1}^{m} (\log p_j)p_j$</td>
</tr>
<tr>
<td>$S(p, i)$</td>
<td>$2p_i - \sum_{j=1}^{m} p_j^2 - 1$</td>
<td>$(\sum_{j=1}^{m} p_j^2)^{1/2}$</td>
<td>$\log p_i$</td>
</tr>
<tr>
<td>$d(p, q)$</td>
<td>$\sum_{j=1}^{m} (p_j - q_j)^2$</td>
<td>$(\sum_{j=1}^{m} q_j^2)^{1/2} - \frac{\sum_{j=1}^{m} p_j q_j}{(\sum_{j=1}^{m} p_j^2)^{1/2}}$</td>
<td>$\sum_{j=1}^{m} (\log(q_j/p_j))q_j$</td>
</tr>
</tbody>
</table>

- $\sum_{j=1}^{m} (\log p_j)p_j = E[-\log(p)] = -H(p)$ is negative Shannon entropy with base $e$.
- $\sum_{j=1}^{m} (\log(q_j/p_j))q_j$ is Kullback-Leibler divergence $KL(q\|p)$
- Quadratic $d(p, q)$ is symmetric
Probabilistic forecasts of a **binary** variable

<table>
<thead>
<tr>
<th></th>
<th>Quadratic</th>
<th>Spherical</th>
<th>Logarithmic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(G(p))</td>
<td>(2p(p - 1))</td>
<td>((2p^2 - 2p + 1)^{1/2})</td>
<td>(p \log \left(\frac{p}{1-p}\right) + \log(1 - p))</td>
</tr>
<tr>
<td>(S(p, 1))</td>
<td>(-2(1 - p)^2)</td>
<td>(p(2p^2 - 2p + 1)^{-1/2})</td>
<td>(\log p)</td>
</tr>
<tr>
<td>(S(p, 0))</td>
<td>(-2p^2)</td>
<td>((1 - p)(2p^2 - 2p + 1)^{-1/2})</td>
<td>(\log(1 - p))</td>
</tr>
</tbody>
</table>
Scoring rules for density forecasts

- Let $\mu$ be a $\sigma$–finite measure on the measurable set $(\Omega, \mathcal{F})$.

- For $\alpha > 1$, let $\mathcal{L}_\alpha$ be the class of probability measures on $(\Omega, \mathcal{F})$ that are absolutely continuous wrt $\mu$ and have $\mu$–density $p$ such that
  \[\|p\|_\alpha = \left( \int p(\omega)^\alpha \mu(d\omega) \right)^{1/\alpha} < \infty\]

- For a probabilistic forecast $P \in \mathcal{L}_\alpha$ with $\mu$–density $p$, call $p$ a predictive density or density forecast.

- For example, think about $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ and $\mu$ is Lebesgue measure. Then $P$ is the distribution of some random variable (r.v.) $X$ with its density $p(x)$.

- Expected score under density $q$: $S(p, q) := \int S(p, x)q(x)dx$
<table>
<thead>
<tr>
<th></th>
<th>Quadratic</th>
<th>Spherical</th>
<th>Logarithmic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G(p)$</td>
<td>$\sum_{j=1}^{m} p_j^2 - 1$</td>
<td>$(\sum_{j=1}^{m} p_j^2)^{1/2}$</td>
<td>$\sum_{j=1}^{m} (\log p_j) p_j$</td>
</tr>
<tr>
<td>$S(p, i)$</td>
<td>$2p_i - \sum_{j=1}^{m} p_j^2 - 1$</td>
<td>$\frac{p_i}{(\sum_{j=1}^{m} p_j^2)^{1/2}}$</td>
<td>$\log p_i$</td>
</tr>
<tr>
<td>$d(p, q)$</td>
<td>$\sum_{j=1}^{m} (p_j - q_j)^2$</td>
<td>$(\sum_{j=1}^{m} q_j^2)^{1/2} - \frac{\sum_{j=1}^{m} p_j q_j}{(\sum_{j=1}^{m} p_j^2)^{1/2}}$</td>
<td>$\sum_{j=1}^{m} (\log q_j / p_j) q_j$</td>
</tr>
</tbody>
</table>

- Quadratic and Spherical scores are strictly proper relative to the class of $\mathcal{L}_2$; Log score is strictly proper relative to the class of $\mathcal{L}_1$.
- For log score, the maximized expected score is the negative Shannon entropy and the divergence function is the Kullback-Leibler divergence.
Continuous ranked probability score (CRPS)

Problems:

- The restriction to absolutely continuous predictive distributions is often impractical.
- For example, probabilistic quantitative precipitation forecasts involve distributions with a point mass at zero.
- \( S(p, x) \) are not sensitive to distance: no credit is given for assigning high probabilities to values near but not identical to the one realizing.
- Log score penalizes low probability events harshly and is highly sensitive to outliers.
- All quadratic, spherical and log scores cannot be used to assess deterministic or discrete ensemble forecasts.
- Continuous ranked probability score (CRPS) is used to address these problems.
Continuous ranked probability score (CRPS)

**Definition (CRPS)**

Let \( \mathcal{P} \) be consist of all probability measures on \((\mathbb{R}, \mathcal{B})\). Let \( F \) be the cumulative distribution function of a probabilistic forecast \( P \in \mathcal{P} \). The **continuous ranked probability score (CRPS)** is defined as

\[
CRPS(F, x) = \int_{-\infty}^{\infty} (F(y) - 1\{y \geq x\})^2 \, dy
\]

where \( x \) is the verifying observation, and \( 1\{y \geq x\} \) is a step function along the real line that attains the value 1 if \( y \geq x \) and the value 0 otherwise.

- \( CRPS(F, x) = E_F|X - x| - \frac{1}{2} E_F|X - X'| \), where \( X \) and \( X' \) are independent copies of a r.v. with cdf \( F \): **analytic expression**.

- \( CRPS(F, x) \) has the same unit as the observations, and it generalizes the absolute error to which it reduces if \( F \) is a deterministic forecast.
Idea of \( CRPS(F, x) = \int_{-\infty}^{\infty} (F(y) - 1\{y \geq x\})^2 \, dy \)
Why CRPS generalizes the absolute error?

- Absolute error is defined as $|f - o|$, where $f$ is the deterministic forecast value and $o$ is the realized value.

- The predicted value is $y = f = 0$ and the realized value is $x = o = 2$, and so the CRPS is

$$ \int_{-\infty}^{\infty} (F(y) - 1\{y \geq x\})^2 \, dy = \int_{0}^{2} 1^2 \, dy = 2 = |0 - 2| = |f - o|$$
More CRPS

- With $CRPS(F, x) = E_F |X - x| - \frac{1}{2} E_F |X - X'|$, when the predictive distribution is normal $N(\mu, \sigma^2)$, the analytic expression of CRPS is

$$CRPS(N(\mu, \sigma^2), x) = \sigma \left( z(2\Phi(z) - 1) + 2\phi(z) - \frac{1}{\sqrt{\pi}} \right),$$

where $\phi$ and $\Phi$ represent the standard Gaussian density and cdf, and $z = \frac{x-\mu}{\sigma}$.

- The CRPS for a Gaussian mixture distribution is

$$CRPS \left( \sum_{m=1}^{M} w_m N(\mu_m, \sigma_m^2), x \right) =$$

$$\sum_{m=1}^{M} w_m A(x - \mu_m, \sigma_m^2) - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} w_m w_n A(\mu_m - \mu_n, \sigma_m^2 + \sigma_n^2),$$

where $A(\mu, \sigma^2) = 2\sigma \phi \left( \frac{\mu}{\sigma} \right) + \mu \left( 2\Phi \left( \frac{\mu}{\sigma} \right) \right)$.
More CRPS

- CRPS is proper relative to the class $\mathcal{P}$, all probability measures on $(\mathbb{R}, \mathcal{B})$, and strictly proper relative to the subclass of $\mathcal{P}_1$ of Borel probability measures that have finite first moment.

- Maximized expected score is

$$G(F) = CRPS(F, F) = \int_{-\infty}^{\infty} F(y)(1 - F(y)) \, dy = \frac{1}{2} E_F |X - X'|$$

- The divergence function is

$$d(F, H) = \int_{-\infty}^{\infty} (F(y) - H(y))^2 \, dy,$$

which is symmetric.
An R package for probabilistic forecasting using ensemble post-processing via Bayesian model averaging.

The modeling functions estimate model parameters from training data via the EM algorithm for normal mixture models (appropriate for temperature or pressure), mixtures of gamma distributions (appropriate for maximum wind speed), and mixtures of gamma distributions with a point mass at 0 (appropriate for quantitative precipitation).
• Use `ensembleBMA()` to fit the model to obtain forecasts:
  
  ```r
  srftBMA290104fit <- ensembleBMA(srftData, dates = "2004012900", model = "normal", trainingDays = 25)
  ```

• Use `quantileForecast()` to get BMA forecasts on grids:
  
  ```r
  gridForc290104 <- quantileForecast(srftBMA290104fit, srftGridData, quantiles = c(.1, .5, .9))
  ```

• Use `cdf()` to estimate the probability of some event at grid locations:
  
  ```r
  probFreeze290104 <- cdf(srftBMA290104fit, srftGridData, date = "2004012900", value = 273.15)
  ```
Review of Scoring Rules

Scoring Rules for Categorical Variables

Scoring Rules for Continuous Variables

R Package ensembleBMA
- CRPS() computes the mean continuous ranked probability scores:
  > CRPS(srftBMA290104fit, srftData)
  ensemble BMA
  1.945544 1.491162

- crps() computes continuous ranked probability score for each observation location:
  > head(crps(srftBMA290104fit, srftData))
  ensemble BMA
  19195  0.6937344  0.6984111
  19196  3.4370938  2.0123193
  19197  1.0063906  0.7673071
  19198  3.0689531  1.4611981
  19199  0.0847500  0.6717240
  19200  2.4739688  1.4290152
Thank you!
• Ensemble forecasting: a numerical weather prediction method that is used to attempt to generate a representative sample of the possible future states of a dynamic system.

• If $F = F_{ens}$ is a discrete predictive distribution from a forecast ensemble of size $M$, the evaluation of the CRPS is straightforward. The predictive cumulative distribution function $F_{ens}$ has jumps of size $1/M$ at the ensemble member values $x_1, \ldots, x_M$, and

$$CRPS(F_{ens}, x) = \frac{1}{M} \sum_{m=1}^{M} |x_m - x| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{n=1}^{M} |x_m - x_n|$$