

Proper Scoring Rules in Statistical Inference

Regole di Punteggio Proprie in Statistica Inferenziale

A. P. Dawid

Abstract Proper scoring rules can be used to motivate You to assess Your true uncertainty honestly, as well as to measure the quality of Your past probability forecasts in the light of the actual outcomes. They also have many other statistical applications. In this overview paper I discuss characterisations, properties and specialisations of proper scoring rules, and describe some of their uses, including robust estimation and Bayesian model selection.

Abstract *Regole di punteggio proprie possono essere utilizzate per motivarci a valutare la nostra incertezza onestamente, così come per misurare la qualità della nostra previsione probabilistica, alla luce dei risultati effettivamente ottenuti. Esse hanno anche molte altre applicazioni statistiche. In questo lavoro presenterò caratterizzazioni, proprietà e specificità delle regole di punteggio proprie, e mi soffermerò su alcune delle loro applicazioni, far cui problemi di stima robusta e selezione del modello Bayesiano.*

Key words: probability forecasting, log score, Brier score, Bregman score, Hyvärinen score, robust estimation, Bayesian model selection

1 Introduction

The theory of proper scoring rules, originally developed to assist meteorologists in issuing weather forecasts, can be considered as belonging to various statistical sub-disciplines: subjective Bayesianism, decision theory, prequential analysis,... Recently it has become apparent that it has many uses that transcend all these, and that it supplies a valuable general tool in statistical theory, where we can often, with advantage, use a proper scoring rule as an alternative to likelihood. The aim of this

A. P. Dawid
Statistical Laboratory, University of Cambridge
e-mail: apd@statslab.cam.ac.uk

paper is to present a brief overview of the theory and applications of proper scoring rules. For more details, see *e.g.* [5, 12, 7].

2 Proper scoring rules

Suppose You have to assess Your uncertainty about an observable (but as yet unobserved) uncertain quantity X , taking values in $\mathcal{X}$. You will quote a distribution Q over $\mathcal{X}$. A *proper scoring rule* can be thought of as a device to encourage You to be honest in this task: that is, to ensure that, if Your “true” uncertainty is in fact represented by distribution P , then You will choose to announce $Q = P$ as Your quoted distribution.

Formally, a scoring rule is a function $S : \mathcal{X} \times \mathcal{P} \rightarrow \mathbb{R}$, where $\mathcal{P}$ is a class of probability distributions over $\mathcal{X}$. The interpretation is that, if You quote distribution Q , and Nature later reveals the value of X to be x , then You will suffer a penalty $S(x, Q)$. Thus S is a loss function in a certain game between You and Nature.

The principles of decision theory say that You should act so as to minimise Your expected loss. If You believe $X \sim P$ but quote Q , Your expected loss will be $S(P, Q) := \mathbb{E}_{X \sim P} S(X, Q)$. We say that the scoring rule S is *proper* with respect to $\mathcal{P}$ if, for $P, Q \in \mathcal{P}$, Your expected score $S(P, Q)$ is minimised in Q at $Q = P$; and S is *strictly proper* if this is the unique minimum: $S(P, Q) > S(P, P)$ for $P, Q \in \mathcal{P}$, $Q \neq P$.

3 Examples

3.1 Quadratic score

For binary X , with $\mathcal{X} = \{0, 1\}$, the *quadratic* or *Brier* score is: $S(x, Q) = (x - q)^2$ with $q := Q(X = 1)$. Then $S(P, Q) = (p - q)^2 + p(1 - p)$, where $p := P(X = 1)$. For any p this is uniquely minimised at $q = p$, showing strict propriety.

Brier [3] developed this scoring rule to assist meteorologists in forecasting rain. de Finetti [8] used it to argue that subjective probabilities ought to obey the classical axioms. For example, suppose You quote probability q_1 for an event A , and q_0 for its complement $\bar{A}$, and that both of these are assessed by means of the quadratic score. Your total penalty will be $(1 - q_1)^2 + q_0^2$ if A occurs, $q_1^2 + (1 - q_0)^2$ if not. It is easy to show that, if (and only if) $q_0 + q_1 \neq 1$, You could have made alternative assessments (for example, $q_1^* = (q_1 + 1 - q_0)/2$, $q_0^* = 1 - q_1^*$) that would have produced a smaller total penalty, for either outcome — so Your initial assessment was “incoherent”.

A similar argument can be based on any other proper scoring rules, and further used to be show that Your assessments should be the same no matter which scoring rule is used [15].

3.2 Log score [9]

Let P, Q have density functions $p(\cdot), q(\cdot)$ with respect to a fixed measure μ on $\mathcal{X}$. The *log score* is given by: $S(x, Q) = -\log q(x)$. Then

$$\begin{aligned} S(P, Q) &= -\int p(x) \log q(x) d\mu(x) \\ &= \int \{q(x) - p(x) \log q(x)\} d\mu(x) - 1. \end{aligned}$$

Since $q - p \log q$ is uniquely minimised in q at $q = p$, the log score is strictly proper.

An important property is *locality*: to compute it we only need to know the value of the function $q(\cdot)$ at the realised outcome x . So long as $\#(\mathcal{X}) > 2$ the log score is essentially the only local proper scoring rule [1].

Many fundamental statistical concepts — most obviously, log likelihood — are intimately related to the log score. The minimised expected score $H(P) := S(P, P)$ is the *Shannon entropy*. The difference between prior and expected posterior Shannon entropy, $C(X, U) := H(P_X) - E_U\{H(P_{X|U})\}$, is the *mutual information* between X and an additional variable U . The excess expected score over the minimum, $D(P, Q) := S(P, Q) - H(P)$, is the *Kullback-Leibler discrepancy*. Its local form, $D(P_\theta, P_{\theta+d\theta})$ defines a Riemannian metric structure $g(\theta)d\theta^2$, where $g(\theta)$ is the *Fisher information*; likewise other concepts of *information geometry* arise naturally. All the above definitions can be applied starting from an arbitrary proper scoring rule, and lead to generalisations of these concepts which retain many of their familiar properties.

3.3 Bregman score [2]

Define

$$S(x, Q) = -\psi'\{q(x)\} - \int d\mu(y) \cdot [\psi\{q(y)\} - q(y) \psi'\{q(y)\}], \quad (1)$$

where $\psi : \mathbb{R}^+ \rightarrow \mathbb{R}$ is strictly convex and differentiable. We find (writing $p = p(y), q = q(y)$):

$$S(P, Q) - S(P, P) = \int d\mu(y) \cdot [\psi(p) - \{\psi(q) + \psi'(q)(p - q)\}]. \quad (2)$$

The integrand in (2) is positive for $q \neq p$, so S is strictly proper. This class includes the Brier score, for $\psi(q) = (2q^2 - 1)/4$, and the log score, for $\psi(q) = q \ln q$.

3.4 Hyvärinen score [10]

Take $\mathcal{X} = \mathbb{R}^k$, and

$$S(\mathbf{x}, Q) = \Delta \ln q(\mathbf{x}) + \frac{1}{2} |\nabla \ln q(\mathbf{x})|^2 = \frac{\Delta \sqrt{q(\mathbf{x})}}{\sqrt{q(\mathbf{x})}} \quad (3)$$

where ∇ denotes gradient, and Δ the Laplacian operator $\sum_{i=1}^k \partial^2 / (\partial x_i)^2$, on $\mathcal{X}$. Assuming that boundary terms arising on integrating by parts vanish, it can be shown that

$$S(P, Q) - S(P, P) = \frac{1}{2} \int d\mu(\mathbf{y}) \cdot |\nabla \ln p(\mathbf{y}) - \nabla \ln q(\mathbf{y})|^2$$

which is positive unless $Q = P$, demonstrating strict propriety.

This scoring rule depends on $q(\cdot)$ through its value and its first two derivatives at the realized value x — it is *local of order 2*. It is also *homogeneous*: it can be computed even if the normalising constant of $q(\cdot)$ is unknown. These properties can be generalized and characterized [13].

3.5 General construction

Consider a general decision problem, with state space $\mathcal{X}$, action space $\mathcal{A}$, and loss function $L(x, a)$. For $P \in \mathcal{P}$, the *Bayes act* is $a_P := \arg \min_{a \in \mathcal{A}} L(P, a)$ where $L(P, a) := \mathbb{E}_{X \sim P} L(X, a)$. Define a scoring rule S by: $S(x, Q) = L(x, a_Q)$ ($x \in \mathcal{X}, Q \in \mathcal{P}$). Then $S(P, Q) = L(P, a_Q) \geq L(P, a_P) = S(P, P)$. Thus S is a proper scoring rule with respect to $\mathcal{P}$. This construction converts any statistical decision problem into one involving a proper scoring rule.

4 Motivating and assessing probability forecasts

Just like the penal code, a proper scoring rule S has a dual function: both to deter and to punish dishonesty. Suppose You know You will be penalised according to S . The deterrent function applies when, ahead of observation of X , You are contemplating Your announcement of Q : then, as described above, You should announce Your true distribution P . The punishment function applies when, after You have announced Q , Nature reveals the value x of X , and You receive the associated penalty, $S(x, Q)$. An observer could use this penalty, accumulated over a sequence of forecasting tasks, to compare the performances of several forecasters.

4.1 Motivation

A good way to motivate You to be careful in assessing Your probabilities is to replace the numerical details of Your Q by a table of its consequences. Suppose You wish to assess Your probability for an event A . You are presented with Table 1 (based on the Brier score — other proper scores could be used), but with the last column covered up. Each row specifies a pair of penalties: one that You would suffer if A occurs, the other if not. Your task is to select Your most favoured row. Intuitively, if You believe A is very improbable You will be willing to suffer a large penalty if it occurs in exchange for a small penalty if not, and so gravitate towards the top of the table. When You have chosen Your favoured row, the value in the third column is revealed: it is Your implicit probability value, consistent with Your displayed preferences.

A	$\bar{A}$	q
100.0	0.0	.00
81.0	1.0	.10
64.0	4.0	.20
49.0	9.0	.30
36.0	16.0	.40
25.0	25.0	.50
16.0	36.0	.60
9.0	49.0	.70
4.0	64.0	.80
1.0	81.0	.90
0.0	100.0	1.00

Table 1 Brier score disguised

4.2 Assessment

Suppose, on each day $i - 1$, a weather forecaster issues probability forecasts q_i for the event A_i of rain on the following day, i , and the actual outcome is $a_i \in \{0, 1\}$. Over n such days his cumulative Brier penalty score is

$$S_+ := \sum_{i=1}^n (q_i - a_i)^2. \quad (4)$$

We can partition S_+ into components that measure two effectively orthogonal dimensions: *substantive ability*, representing knowledge and skill in the forecasting task; and *normative ability*: accuracy in assessing empirical frequencies.

Suppose that the forecasts are confined to a set $(\pi_j : j = 1, 2, \dots, k)$ of values, and that forecast π_j is issued on n_j occasions, forming a group G_j say, in which a

proportion ρ_j result in the event occurring. Then we find

$$\begin{aligned} S_+ &= \sum_{j=1}^k n_j \{ \rho_j(1 - \pi_j)^2 + (1 - \rho_j)\pi_j^2 \} \\ &= S_1 + S_2 \\ &= S_0 + S_1 - S_3 \end{aligned}$$

where

$$\begin{aligned} S_1 &= \sum_{j=1}^k n_j (\pi_j - \rho_j)^2 \\ S_2 &= \sum_{j=1}^k n_j \rho_j (1 - \rho_j) \\ S_3 &= \sum_{j=1}^k n_j (\rho_j - \bar{\rho})^2 \\ S_0 &= \bar{\rho}(1 - \bar{\rho}). \end{aligned}$$

and $\bar{\rho} = (\sum_{i=1}^n a_i)/n = (\sum_{j=1}^k n_j \rho_j)/n$ is the overall proportion of rainy days. Similar decompositions can be based on other proper scoring rules.

The criterion of *calibration* requires that $\pi_j \approx \rho_j$; that is, in the group G_j , the issued forecast π_j should approximate the actual proportion ρ_j of rainy days. We can interpret the non-negative component S_1 as measuring overall departure from calibration: it vanishes for a “well-calibrated” forecaster who has $\pi_j = \rho_j$, all j , and so has normative ability, in being able to assign realistic probability values. This component can be eliminated (without affecting the other components) by *recalibration*, whereby on each day in G_j the forecast value is changed to ρ_j .

The term S_2 measures *refinement*, or substantive ability. It is small when all the ρ_j ’s are close to 0 or 1: this will be so for a forecaster who is able to sort the days effectively into rainy and dry. We can also express $S_2 = S_0 - S_3$. Here S_0 is fixed by Nature, while S_3 is large (so reducing the overall penalty) when the sample variance of the proportions ρ_j ’s is large; S_3 attains its minimum (corresponding to highest penalty) when the ρ_j ’s do not vary across the groups $\{G_j\}$, meaning that the forecaster is totally unable to discriminate between rainy and dry days.

5 Classical inference

Likelihood inference, while efficient, can be non-robust, as well as requiring the often problematic computation and manipulation of the normalising constant of a distribution. Some of these issues can be addressed if we replace the negative log likelihood by some other proper scoring rule S . For a smooth parametric family $\mathcal{P} = \{P_\theta\}$, define

$$s(x, \theta) = \frac{\partial S(x, P_\theta)}{\partial \theta}. \quad (5)$$

Given a random sample $(x_1, \dots, x_n)$, we can estimate θ by solving the equation

$$\sum_{i=1}^n s(x_i, \theta) = 0. \quad (6)$$

Generalising the likelihood equation, to which it reduces when S is the log score, (6) yields an unbiased estimating equation. Its root will be consistent, though typically inefficient, in repeated sampling, with an asymptotic variance that follows from standard M -estimation theory [7, 14].

Such an estimator will however often be robust. For the case of a location family, with density $p_\theta(x) = f(x - \theta)$, and a Bregman scoring rule (1), the influence function will be bounded if and only if $\psi''\{f(u)\}f'(u)$ is bounded.

Other advantages accrue if we use a homogeneous scoring rule, such as (3). Then we can construct the estimator without having to compute the normalising constant of the distribution [11].

6 Bayesian model selection

Bayesian model selection is based on the predictive density function assuming the validity of a putative model M :

$$p_M(\mathbf{x}) = \int_{\Theta_M} p_M(\mathbf{x} | \theta_M) \pi_M(\theta_M) d\theta_M. \quad (7)$$

As a function of M , $L(M) \propto p_M(\mathbf{x})$ supplies the *marginal likelihood* function over models M , based on data $\mathbf{x}$.

However, use of (7) requires that the within-model prior density π_M be proper. This constraint can be evaded on noting that $-\log p_M(\mathbf{x})$ is the log-score for the predictive distribution, and replacing this by a homogeneous scoring rule S , such as S_H given by (3). If in (7) we formally take $\pi_M(\theta_M) \propto \alpha_M(\theta_M)$ for an unnormalisable function α_M , the unspecified normalising constant will not appear in $S_H(\mathbf{x}, \pi_M)$. Specifically, we find

$$S_H(\mathbf{x}, P_M) = E\{S_H(\mathbf{x}, P_\Theta) | \mathbf{X} = \mathbf{x}\} + \sum_i \text{var} \left\{ \frac{\partial \log p(\mathbf{x} | \Theta)}{\partial x_i} \middle| \mathbf{X} = \mathbf{x} \right\}, \quad (8)$$

which will be well-defined so long as the *posterior* means and variances appearing in it exist. We can then compare models M by means of their scores (8). When this score is computed prequentially [4], this will typically lead to consistent selection of the correct model [6].

References

1. J. M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7:686–690, 1979.
2. L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.
3. Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.
4. A. P. Dawid. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:277–305, 1984.
5. A. P. Dawid. Probability forecasting. In S. Kotz, N. L. Johnson, and C. B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218, New York, 1986. Wiley-Interscience.
6. A. P. Dawid and M. Musio. Bayesian model selection based on proper scoring rules. Manuscript in preparation.
7. A. P. Dawid and M. Musio. Theory and applications of proper scoring rules. *Metron*, 2014. To appear, DOI: 10.1007/s40300-014-0039-y [arXiv:1401.0398](https://arxiv.org/abs/1401.0398).
8. Bruno de Finetti. *Theory of Probability (Volumes 1 and 2)*. John Wiley and Sons, New York, 1975. (Italian original Einaudi, 1970).
9. I. J. Good. Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14:107–114, 1952.
10. A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Machine Learning Research*, 6:695–709, 2005.
11. V. Mameli, M. Musio, and A. P. Dawid. Comparison of approaches to inference in stationary AR(1) models. In *Proceedings of the 47th Scientific Meeting of the Italian Scientific Society, Cagliari*, 2014.
12. M. Musio and A. P. Dawid. Local scoring rules: A versatile tool for inference. In *Proceedings of the 59th ISI World Statistics Congress, Hong Kong*, 2013. <http://www.statistics.gov.hk/wsc/STS019-P3-S.pdf>.
13. Matthew F. Parry, A. Philip Dawid, and Steffen L. Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40:561–92, 2012.
14. E. Ruli, N. Sartori, and L. Ventura. Approximate Bayesian Computation with proper scoring rules. In *Proceedings of the 47th Scientific Meeting of the Italian Scientific Society, Cagliari*, 2014.
15. L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801, 1971.