A construction principle for proper scoring rules

Jonas R. Brehmer
Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

July 29, 2021

Abstract

Proper scoring rules enable decision-theoretically principled comparisons of probabilistic forecasts. New scoring rules can be constructed by identifying the predictive distribution with an element of a parametric family and then applying a known scoring rule. We introduce a condition which ensures propriety in this construction and thereby obtain novel proper scoring rules.

Keywords and phrases: Proper scoring rule, Dawid-Sebastiani score, forecast evaluation, exponential family.

2010 MSC: Primary 62C99

1 Introduction

In order to account for the inherent uncertainty of future quantities or events, it is preferable to issue forecasts in the form of probability distributions (Gneiting and Katzfuss, 2014). One way to measure the predictive ability of such probabilistic forecasts is to assign a score, or loss, $S(F, y)$ to each pair of forecast distribution $F \in \mathcal{F}$ and observation $y \in \mathsf{O}$. In this setting O is a topological space with Borel σ-algebra $\mathcal{O}$ and $\mathcal{F}$ is a class of probability distributions on $(\mathsf{O}, \mathcal{O})$. A *scoring rule* is a function $S : \mathcal{F} \times \mathsf{O} \to \bar{\mathbb{R}}$ such that for all $F, G \in \mathcal{F}$ the expectation

$$\mathbb{E}_G S(F, Y) = \int_{\mathsf{O}} S(F, y) \, \mathrm{d}G(y)$$

is well defined. Here we let $\bar{\mathbb{R}} := [-\infty, \infty]$ be the extended real line. The scoring rule S is *proper* relative to $\mathcal{F}$ if

$$\mathbb{E}_G S(G, Y) \leq \mathbb{E}_G S(F, Y)$$

for all $F, G \in \mathcal{F}$. It is *strictly proper* if equality holds if and only if $F = G$. If a forecaster believes that the quantity y is drawn from the distribution G and receives a penalty $S(F, y)$ for reporting F, then propriety ensures that reporting her true belief $F = G$ is an optimal strategy in expectation. For recent reviews of the theory and application of proper scoring rules we refer to Dawid (2007), Gneiting and Raftery (2007), and Dawid and Musio (2014).

Various proper scoring rules have been proposed in the literature, in particular for the special situation where each member of $\mathcal{F}$ admits a density with respect to some σ-finite measure on $(\mathsf{O}, \mathcal{O})$. The *logarithmic score* (Good, 1952) is defined via

$$\mathrm{LogS}(f, y) := -\log f(y),$$

where f denotes the probability density function of F. It is the most popular strictly proper scoring rule for densities since it connects to various fundamental statistical concepts, such as maximum-likelihood estimation, information criteria, and Bayes factors (Gneiting and Raftery, 2007). For $\mathsf{O} = \mathbb{R}^d$ a popular scoring rule which depends on the first two moments only is the *Dawid-Sebastiani (DS) score* (Dawid and Sebastiani, 1999). If $\mathcal{F}$ is a class of distributions with finite second moments, then it is given by

$$\mathrm{DSS}(F, y) := \log\det\Sigma_F + (y - \mu_F)^\top \Sigma_F^{-1}(y - \mu_F),$$

where μ_F and Σ_F denote the mean and the covariance matrix of the predictive distribution F. The DS score is proper, but not strictly proper, as distributions with the same first and second moments attain the same score.

This work is motivated by the fact that, up to unimportant constants, the DS score of F equals the logarithmic score of a multivariate normal distribution with the same mean and covariance matrix as F. More precisely,

$$\begin{aligned}\mathrm{DSS}(F, y) &= -2\log\left(\varphi(y \mid \mu_F, \Sigma_F)\right) - d\log(2\pi) \\ &= 2\,\mathrm{LogS}\left(\varphi(\cdot \mid \mu_F, \Sigma_F), y\right) - d\log(2\pi),\end{aligned}$$

where $\varphi(\cdot \mid \mu, \Sigma)$ denotes the density of the multivariate normal distribution with mean μ and covariance matrix Σ. This connection raises the question, under which conditions we obtain a proper scoring rule by identifying the predictive distribution F with an element of a parametric family (e.g. the normal distributions) and then applying another proper scoring rule (e.g. the logarithmic score). The ensuing section gives a simple condition which ensures propriety in this construction and is neither restricted to the normal family nor to the logarithmic score. The manuscript concludes with several examples which yield new proper scoring rules and recover existing ones.

2 Construction principle

Let $\mathcal{E} := \{F_\theta \mid \theta \in \Theta\} \subseteq \mathcal{F}$ be a parametric family of distributions with parameter space Θ. Let $\phi : \mathcal{F} \to \mathcal{E}$, $F \mapsto F_\theta$ be a mapping onto $\mathcal{E}$ and write θ_F for the parameter θ in $\phi(F) = F_\theta$.

Theorem 2.1. *Let $S : \mathcal{F} \times \mathsf{O} \to \bar{\mathbb{R}}$ be a proper scoring rule and $\phi : \mathcal{F} \to \mathcal{E}$. If there is a function $H : \mathsf{O} \to \mathbb{R}$ which is integrable with respect to all $F \in \mathcal{F}$*

and such that for all $F, G \in \mathcal{F}$

$$\mathbb{E}_G S(\phi(F), Y) + \mathbb{E}_G H(Y) = \mathbb{E}_{\phi(G)} S(\phi(F), Y) + \mathbb{E}_{\phi(G)} H(Y), \tag{1}$$

then the scoring rule

$$S^*(F, y) = S(\phi(F), y) = S(F_{\theta_F}, y)$$

is proper.

Proof. For $F, G \in \mathcal{F}$ invoke Equation (1) two times to obtain

$$\begin{aligned}
\mathbb{E}_G S^*(F, Y) = \mathbb{E}_G S(\phi(F), Y) &= \mathbb{E}_{\phi(G)} S(\phi(F), Y) + \mathbb{E}_{\phi(G)} H(Y) - \mathbb{E}_G H(Y) \\
&\geq \mathbb{E}_{\phi(G)} S(\phi(G), Y) + \mathbb{E}_{\phi(G)} H(Y) - \mathbb{E}_G H(Y) \\
&= \mathbb{E}_G S(\phi(G), Y) = \mathbb{E}_G S^*(G, Y),
\end{aligned}$$

where the inequality stems from the propriety of S. □

Strict propriety in Theorem 2.1 is only possible for special choices of $\mathcal{E}$ and ϕ, which render the mapping ϕ a bijection, since otherwise two different distributions can attain the same score.

Exponential families are natural and flexible candidates for distributional classes in statistics. We call a set of densities $\{f(\cdot \mid \theta) \mid \theta \in \Theta\}$ on O an *exponential family* if any member can be represented via

$$f(y \mid \theta) = h(y) \exp\left(\eta(\theta)^\top t(y) - A(\theta)\right)$$

for measurable functions $h : \mathsf{O} \to (0, \infty)$, $t : \mathsf{O} \to \mathbb{R}^m$, $\eta : \Theta \to \mathbb{R}^m$, and $A : \Theta \to \mathbb{R}$, where $m \in \mathbb{N}$. The mapping A is often called *log-partition* function and t is a sufficient statistic for the parameter θ, see Barndorff-Nielsen (2014) for details.

When the scoring rule S in Theorem 2.1 is the logarithmic score, exponential families are convenient candidates for the class $\mathcal{E}$. In detail, let $\mathcal{E}$ be an exponential family on O and set $H(y) := \log h(y)$, then Equation (1) holds if

$$\mathbb{E}_G t(Y) = \mathbb{E}_{\phi(G)} t(Y), \tag{2}$$

i.e. if the expectations of t agree. Since $\phi(G)$ is a member of the exponential family $\mathcal{E}$, the right hand side of (2) can be calculated and expressed in terms of $\theta \in \Theta$ via the partial derivatives of the log-partition function A. If a closed-form expression exists, this yields sufficient conditions on the mapping $\phi : \mathcal{F} \to \mathcal{E}$ for (1) to hold, see Section 3 for concrete examples.

Another possible choice for S in Theorem 2.1 which fits well with exponential families is the *Hyvärinen score* (Hyvärinen, 2005). Let $\mathsf{O} = \mathbb{R}^d$ and let ∇ denote the gradient and Δ the Laplace operator. Define $\mathcal{L}^*$ as the class

of densities on O which are twice differentiable, positive almost everywhere, and such that $\nabla \log(f(y))g(y) \to 0$ as $\|y\| \to \infty$ for all $f, g \in \mathcal{L}^*$. Then the Hyvärinen score is given by

$$\mathrm{HyvS}(f, y) := \Delta \log f(y) + \frac{1}{2}\|\nabla \log f(y)\|^2$$

and it is strictly proper relative to $\mathcal{L}^*$ if its expectation is finite. The Hyvärinen score has the remarkable property that it is 0-homogeneous, i.e. to compute $\mathrm{HyvS}(f, y)$ the predictive density f needs to be specified up to the normalization constant only, see Hyvärinen (2005), Parry et al. (2012), and Ehm and Gneiting (2012) for details.

To connect to Theorem 2.1 assume for simplicity that $\mathcal{E}$ is an exponential family of distributions on $\mathsf{O} = \mathbb{R}^d$ where the function h is constant and all densities satisfy the regularity conditions of the class $\mathcal{L}^*$. If we define $W_\theta(y) := \eta(\theta)^\top t(y)$, then the Hyvärinen score on $\mathcal{E}$ is completely determined by

$$\Delta W_\theta(y) = \sum_{i=1}^m \eta_i(\theta)\Delta t_i(y) \quad \text{and} \quad \nabla W_\theta(y) = \sum_{i=1}^m \eta_i(\theta)\nabla t_i(y),$$

where the index i denotes the i-th component of a vector in $\mathbb{R}^m$. As a consequence, we can set $H = 0$ and Equation (1) holds if the derivatives of t satisfy

$$\mathbb{E}_G \Delta t_i(Y) = \mathbb{E}_{\phi(G)} \Delta t_i(Y), \tag{3}$$

$$\mathbb{E}_G \left[\nabla t_i(Y)^\top \nabla t_j(Y)\right] = \mathbb{E}_{\phi(G)} \left[\nabla t_i(Y)^\top \nabla t_j(Y)\right], \tag{4}$$

for $i, j = 1, \ldots, m$, giving $m + m(m+1)/2$ identities. Similar to (2) these equations provide sufficient conditions for (1) to hold, which can be used to define a suitable mapping $\phi : \mathcal{F} \to \mathcal{E}$ in Theorem 2.1, see Example 3.4.

3 Examples

Example 3.1 (Normal family)**.** Let $\mathcal{E}$ consist of the multivariate normal distributions with parameter $\theta = (\mu, \Sigma)$, where μ is the mean and Σ the covariance matrix. The exponential family representation of $\mathcal{E}$ implies $t(y) = (y, yy^\top)$. If S is the logarithmic score, then a mapping ϕ can be determined via (2). This yields

$$(\mathbb{E}_G Y, \mathbb{E}_G YY^\top) = \mathbb{E}_G t(Y) = \mathbb{E}_{\phi(G)} t(Y) = (\mu_G, \Sigma_G + \mu_G \mu_G^\top),$$

such that $\theta_F = (\mathbb{E}_F Y, \mathrm{Cov}_F(Y))$ has to be computed from a predictive distribution $F \in \mathcal{F}$. The resulting scoring function

$$S^*(F, y) = \frac{1}{2}\left(\log\det \Sigma_F + d\log(2\pi) + (y - \mu_F)^\top \Sigma_F^{-1}(y - \mu_F)\right)$$

is proper by Theorem 2.1 and an affine transformation of the DS score, as discussed in the introduction.

Example 3.2 (Laplace family)**.** Let $\mathcal{E}$ be the class of centered Laplace distributions with scale parameter $\nu > 0$. Its members have densities $f(y \mid \nu) = (2\nu)^{-1} \exp(-|y|/\nu)$, thus it forms an exponential family with $t(y) = |y|$. In this situation, (2) becomes

$$\mathbb{E}_G|Y| = \mathbb{E}_G t(Y) = \mathbb{E}_{\phi(G)} t(Y) = \mathbb{E}_{\phi(G)}|Y| = \nu_G,$$

such that $\theta_F = \mathbb{E}_F|Y|$ is computed from the predictive distribution. Theorem 2.1 implies that the scoring rule

$$S^*(F, y) = \log(2\nu_F) + \frac{|y|}{\nu_F},$$

where $\nu_F = \mathbb{E}_F|Y|$, is proper. A natural question is whether it is possible to transfer these arguments to the general class of Laplace distributions with parameters (μ, ν), i.e. to the situation of a non-constant location parameter $\mu \in \mathbb{R}$. In this case, (1) reads

$$\mathbb{E}_G\left[\frac{|Y - \mu_F|}{\nu_F} + H(Y)\right] = \mathbb{E}_{\phi(G)}\left[\frac{|Y - \mu_F|}{\nu_F} + H(Y)\right],$$

with $\theta_F = (\mu_F, \nu_F)$. Since the random variable Y and the parameter μ_F cannot be separated, it is not clear how to obtain a mapping ϕ which satisfies this identity for all $F, G \in \mathcal{F}$ if $\mathcal{F}$ is sufficiently large. Consequently, it is not obvious whether Theorem 2.1 can be applied to the logarithmic score in concert with the general Laplace family.

Example 3.3 (Poisson family)**.** Let $\mathsf{O} = \mathbb{N}$ and $\mathcal{E}$ be the class of Poisson distributions with parameter $\lambda > 0$. The exponential family representation implies $t(y) = y$ and (2) becomes

$$\mathbb{E}_G Y = \mathbb{E}_G t(Y) = \mathbb{E}_{\phi(G)} t(Y) = \mathbb{E}_{\phi(G)} Y = \lambda_G,$$

hence $\theta_F = \mathbb{E}_F Y$ gives a suitable mapping ϕ. By Theorem 2.1 the resulting scoring rule

$$S^*(F, y) = -y \log(\lambda_F) + \lambda_F + \log(y!),$$

where λ_F is the expectation of F, is proper.

Example 3.4 (Normal family, continued)**.** Let $\mathcal{E}$ be as in Example 3.1. If S is the Hyvärinen score, the conditions in (3) and (4) simplify to equations which contain the moments $\mathbb{E}_G Y_i$ and mixed moments $\mathbb{E}_G Y_i Y_j$ for $i, j = 1, \ldots, d$, only. Hence, the mapping ϕ of Example 3.1, which is given by the

parameter choice $\theta_F = (\mathbb{E}_F Y, \mathrm{Cov}_F(Y))$, satisfies these conditions. As a result we obtain a Dawid-Sebastiani type scoring rule given by

$$S^*(F,y) = -2\operatorname{tr}\Sigma_F^{-1} + \|\Sigma_F^{-1}(y-\mu_F)\|^2 = -2\operatorname{tr}\Sigma_F^{-1} + (y-\mu_F)^\top \Sigma_F^{-2}(y-\mu_F),$$

which is proper by Theorem 2.1. It already appears in (Hyvärinen, 2005, Section 3.1) in the context of score matching, however, our derivation establishes propriety in wide generality, not only relative to the normal family.

Acknowledgements

I am grateful for funding by the Klaus Tschira Foundation and for infrastructural support provided by the University of Mannheim. I thank Tilmann Gneiting for fruitful comments and discussions.

References

Barndorff-Nielsen, O. E. (2014). *Information and exponential families in statistical theory*. Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester. URL https://doi.org/10.1002/9781118857281.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Ann. Inst. Statist. Math.*, 59, 77–93. URL https://doi.org/10.1007/s10463-006-0099-8.

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72, 169–183. URL https://doi.org/10.1007/s40300-014-0039-y.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.*, 27, 65–81. URL https://doi.org/10.1214/aos/1018031101.

Ehm, W. and Gneiting, T. (2012). Local proper scoring rules of order two. *Ann. Statist.*, 40, 609–637. URL https://doi.org/10.1214/12-AOS973.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Ann. Rev. Stat. Appl.*, 1, 125–151. URL https://doi.org/10.1146/annurev-statistics-062713-085831.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.*, 102, 359–378. URL https://doi.org/10.1198/016214506000001437.

Good, I. J. (1952). Rational decisions. *J. R. Stat. Soc. Ser. B*, 14, 107–114. URL https://www.jstor.org/stable/2984087?

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6, 695–709. URL https://jmlr.org/papers/v6/hyvarinen05a.html.

Parry, M., Dawid, A. P. and Lauritzen, S. (2012). Proper local scoring rules. *Ann. Statist.*, 40, 561–592. URL https://doi.org/10.1214/12-AOS971.