**Local Scoring Rules: A Versatile Tool for Inference**

Monica Musio[1,2] and Philip Dawid[3]
[1] Università degli Studi di Cagliari, Sardinia, Italy
[3] University of Cambridge, UK
[2] Corresponding author: Monica Musio, e-mail: mmusio@unica.it

**Abstract**

In many applications of highly structured statistical models the likelihood function is intractable; in particular, finding the normalisation constant of the distribution can be demanding. One way to sidestep this problem is to to adopt composite likelihood methods, such as the pseudo-likelihood approach. In this paper we display composite likelihood as a special case of a general estimation technique based on *proper scoring rules*, which supply an unbiased estimating equation for any statistical model. The important class of *key local* scoring rules avoids the need to compute normalising constants. Another application arises in Bayesian model selection. The log Bayes factor measures by how much the predictive log score for one model improves on that for another. However, Bayes factors are not well-defined when improper prior distributions are used. If we replace the log score by a suitable local proper scoring rule, these problems are avoided.

Keywords: Bayesian model selection, composite likelihood, homogeneous scoring rule, Hyvärinen score, proper scoring rule, unbiased estimating equation

# 1 Proper scoring rules

Suppose You are required to quote a probability distribution $Q$ for a quantity $X$, after which Nature will reveal the value $x$ of $X$. You will then suffer a loss $S(x,Q)$. Such a function $S$, contrasting a probabilistic forecast with an observed outcome, is termed a *scoring rule*. If You actually consider $X$ to have distribution $P$, then Your expected loss if You quote $Q$ is $S(P,Q) := \mathrm{E}_{X\sim P} S(X,Q)$. The scoring rule is *proper* when, for each $P$, $\inf_Q S(P,Q)$ is attained at $Q = P$. In this case "honesty is the best policy": You are motivated to quote Your truly held belief.

There is a very wide variety of proper scoring rules. Examples include the *log score* (Good 1952), $S(x,Q) = -\log q(x)$, where $q(\cdot)$ is the density of $A$ with respect to some underlying measure $\mu$; and the *Brier* or *quadratic score* (Brier 1950), $S(x,Q) = \int q(y)^2 d\mu(y) - 2q(x)$.

# 2 Estimation

Armed with a proper scoring rule $S$, for any parametric family $\{P_\theta\}$ we can estimate its parameter $\theta$ from a sample $(x_1,\ldots,x_N)$ by minimising the total empirical score $\sum_{t=1}^N S(x_t,P_\theta)$. This is typically equivalent to solving

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

where $s(x,\theta) := \partial S(x,P_\theta)/\partial\theta$. This supplies an unbiased estimating equation (Dawid 2007), thereby leading to an $M$-estimator. When we use the log score, (1) is just the likelihood equation, and we obtain the maximum likelihood estimator. Estimators based on other proper scoring rules will typically be consistent but not efficient; however, they may have compensating virtues of robustness and/or tractability.

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

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

In this case to solve (1) we generally need to be able to compute and differentiate the normalising factor $Z(\theta) = \int f(x \mid \theta)\,dx$. This can be problematic. However an escape route is possible by using a suitable *local* proper scoring rule.

# 3 Locality

To evaluate the log score we only need to know the value of Your forecast density function, $q(\cdot)$, at the value $x$ of $X$ that Nature in fact produces. It is thus termed a *strictly local* proper scoring rule. It can be shown that this property essentially characterises the log score.

However, we can slightly weaken the locality requirement to admit further proper scoring rules. For the case of a sample space that is a real interval,[1] we ask that $S(x,Q)$ should depend on $q(\cdot)$ only through its value and the value of a finite number of its derivatives at $x$. Parry *et al*. (2012) have characterised all such local proper scoring rules as a linear combination of the log score and what they term a *key local* scoring rule. The simplest key local scoring rule is based on the proposal by Hyvärinen (2005):

$$S_H(x,Q) = 2\Delta \ln q(x) + |\nabla \ln q(x)|^2 \tag{2}$$

where $\nabla := (\partial/\partial x)$, $\Delta := \partial^2/(\partial x)^2$. The same formula (2) can be applied to the case of a multivariate observation $\mathbf{X} = (X_1,\ldots,X_k)$, on interpreting $\nabla := (\partial/\partial x_j)$, $\Delta := \sum_{j=1}^k \partial^2/(\partial x_j)^2$.

An important property of every key local scoring rule is *homogeneity*: it is unchanged if $q(\cdot)$ is scaled by a positive constant. In particular, $S(x,Q)$ can be computed without knowledge of the normalising constant of the distribution $Q$.

# 4 Composite likelihood and beyond

Consider a model for a multidimensional variable $\mathbf{X}$. Let $\{\mathbf{X}_k\}$ be a collection of marginal and/or conditional variables, and let $S_k$ be a proper scoring rule for $\mathbf{X}_k$. Then we can construct a proper scoring rule for $\mathbf{X}$ as

$$S(\mathbf{x},Q) = \sum_k S_k(\mathbf{x}_k, Q_k) \tag{3}$$

where $\mathbf{X}_k \sim Q_k$ when $\mathbf{X} \sim Q$. The form (3) localises the problem to the $\{\mathbf{X}_k\}$, which can simplify computation. When each $S_k$ is the log score, (3) becomes a (negative log) *composite likelihood*. We can thus treat composite likelihood in its own right, as supplying a proper scoring rule, rather than as an approximation to true likelihood. Most of the extensive theory and many applications of composite likelihood (see *e.g.* Statistica Sinica (2011)) extend virtually unchanged to the more general case (3): see Dawid and Musio (2013) for an application to estimation for a spatial process.

[1] There are parallel results for the case of a discrete sample space (Dawid *et al*. 2012).

### 4.1 Alternative approach

When the motivation for a composite likelihood approach is the problem of dealing with the normalising constant of the joint distribution, an alternative is to apply a homogeneous scoring rule to that distribution.

**Example 1** Consider the multivariate normal model: $\mathbf{Y} = (Y_1, \ldots, Y_N) \sim \mathcal{N}(\mu, \Phi^{-1})$. In this case the multivariate Hyvärinen score becomes

$$-2\mathrm{tr}(\Phi) + |\Phi(\mathbf{y} - \mu)|^2 . \tag{4}$$

Now restrict to $\mu = \mathbf{0}$ and $\Phi$ of the form $\Phi_N = \beta A_N + \alpha I_N$, where $A_N$ has entries 1 immediately above and below the diagonal, and 0 elsewhere. This defines a (not quite stationary) AR(1) process. The normalising constant of the joint density involves the nasty term $\det \Phi_N$. However, applying instead the multivariate Hyvärinen score (4) eliminates this problem, giving a simple quadratic:

$$-2N\alpha + \sum_{r=1}^{N} (\alpha y_r + \beta z_r)^2, \tag{5}$$

where $z_r := y_{r-1} + y_{r+1}$ (taking $y_{-1} = y_{N+1} = 0$). Defining $\lambda = -\beta/\alpha$ (and assuming (5) is minimised at an interior point of the parameter space) we get estimates $\widehat{\lambda} = s_{yz}/s_{zz}$, $\widehat{\alpha} = N/s_{yy.z}$, where $s_{yz} := \sum_{r=1}^{N} y_r z_r$ *etc.*, and $s_{yy.z} := s_{yy} - s_{yz}^2/s_{zz}$.

Thee above estimates in fact agree with those that would be given by pseudo-likelihood. However if we have $\nu > N-1$ multiple, independent and identically distributed, series generated by the above model, we might wish to base inference on the associated sum-of-squares-and-product matrix $S$, which is a sufficient statistic, having the Wishart distribution $W_N(\nu; \Phi^{-1})$. It is not obvious how to apply pseudo-likelihood to this, but we can use the multivariate Hyvärinen score, based on data $\{s_{ij} : 1 \le i \le j \le N\}$. This delivers the unbiased estimates

$$\begin{aligned} \widehat{\alpha} &= \frac{\nu - N - 1}{N} \sum_{i=1}^{N} s^{ii} \\ \widehat{\beta} &= \frac{\nu - N - 1}{N-1} \sum_{i=1}^{N-1} s^{i,i+1} \end{aligned}$$

where $s^{ij}$ denotes the $(i,j)$ entry of $S^{-1}$.

□

## 5 Bayesian model selection

In Bayesian model selection, we assume that our observations $\mathbf{X}$ are generated from one of a discrete collection $\mathscr{M}$ of parametric models, of possibly varying parameter dimension, and wish to identify the correct model. For a model $M \in \mathscr{M}$, let its parameter be $\theta_M \in \mathbb{R}^{d_M}$, its density at $\mathbf{X} = \mathbf{x}$ be $p_M(\mathbf{x} \,|\, \theta_M)$, and the prior density function for $\theta_M$ be $\pi_M(\theta_M)$. The marginal ("prior predictive") distribution $P_M$ of $\mathbf{X}$ under $M$ then has density

$$p_M(\mathbf{x}) = \int p_M(\mathbf{x} \,|\, \theta_M)\, \pi_M(\theta_M)\, d\theta_M. \tag{6}$$

Given data $\mathbf{X} = \mathbf{x}_0$, the various models can be compared by means of the *marginal likelihood* function, $L(M) \propto p_M(\mathbf{x}_0)$. In particular, the posterior odds in favour of model $M$, as against

model $M'$, are obtained on multiplying the corresponding prior odds by the *Bayes factor*, $\mathrm{BF}^M_{M'} = L_M/L_{M'}$.

Expression (6) is somewhat sensitive to the choice of prior density $\pi_M$. However, unlike the case for within-model estimation, we can not simply replace that by an "objective" improper prior density, *e.g.* uniform on $\mathbb{R}^{d_M}$, since there is no natural scale for such a density — and the formal marginal likelihood will be highly sensitive to the choices made of the scale factors for the various models.

## 5.1 Use of scoring rules

We note that $-\log \mathrm{BF}^M_{M'}$ is just the difference between the log scores for the two predictive distributions, $P_M$ and $P_{M'}$, at the outcome $\mathbf{X} = \mathbf{x}_0$. If we replace log score by a different proper scoring rule, $S(\mathbf{x}, Q)$, we can use that instead to compare the two models. That is, we replace $-\log \mathrm{BF}^M_{M'}$ by the "score factor"

$$\mathrm{SF}^M_{M'} := S(\mathbf{x}, P_M) - S(\mathbf{x}, P'_M). \tag{7}$$

In particular, if $S$ is homogeneous, SF will be entirely insensitive to the arbitrary choice of scale factors in the prior, and will typically deliver a well-defined value, so long only as $p_M$, given by (6) (and similarly $p_{M'}$), is finite at each $\mathbf{x}$—but $p_M$ need not be integrable, as indeed it will not be if $\pi_M$ is not. In this way we evade the above difficulties and obtain an "objective" Bayesian model comparison.

## 5.2 Consistency

Suppose our models relate to a potentially infinite sequence $X_1, X_2, \ldots$ (not necessarily independent and identically distributed). We would like our method to exhibit *model consistency*, whereby, as $n \to \infty$, $\mathrm{SF}^M_{M'} \to -\infty$ with probability 1 under any $P_\theta \in M$, for any alternative model $M'$. In order even to make sense of this requirement, we need to relate the scoring rules used for different sample sizes. We can do this by associating a fixed scoring rule $S_i$ with observation $X_i$ (typically $S_i$ will not vary with $i$), and, for the case of $n$ observations $\mathbf{X}^n = (X_1, \ldots, X_n)$, using the *prequential score*:

$$S^n(\mathbf{x}^n, Q) := \sum_{i=1}^{n} S_i(x_i, Q_i), \tag{8}$$

where $Q_i$ is the conditional distribution, under $Q$, of $X_i$, given $(X_j = x_j : j = 1, \ldots, i-1)$. When $S$ is the log score this is just the overall multivariate log score, and is insensitive to the ordering of the data; for other scores there may be some sensitivity to ordering.

Subject to some regularity conditions on the $S_i$, it can be shown that use of the prequential score will lead to consistent model selection.

**Example 2** Consider the following normal linear model for a data-vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$:

$$\mathbf{Y} \sim \mathcal{N}(X\theta, \sigma^2 I), \tag{9}$$

where $X$ $(n \times p)$ is a known design matrix of rank $p$, and $\theta \in \mathbb{R}^p$ is an unknown parameter vector. We take $\sigma^2$ as known.

We give $\theta$ a normal prior distribution: $\theta \sim \mathcal{N}(m,V)$. The marginal distribution $Q$ of $\mathbf{Y}$ is then $\mathbf{Y} \sim \mathcal{N}(Xm, XVX' + \sigma^2 I)$, with precision matrix

$$\begin{aligned} \Phi &= (XVX' + \sigma^2 I)^{-1} \\ &= \sigma^{-2}\left\{I - X\left(X'X + \sigma^2 V^{-1}\right)^{-1} X'\right\} \end{aligned}$$

on applying the matrix lemma (equation (10)) of Lindley and Smith (1972).

An improper prior can be generated by allowing $V^{-1} \to 0$, yielding $\Phi = \sigma^{-2}\Pi$, where $\Pi := I - X(X'X)^{-1}X'$ is the projection matrix onto the space of residuals. Although this $\Phi$ is singular, and thus can not arise from any genuine dispersion matrix, there is no problem in using it to evaluate the Hyvärinen score given by (4). We obtain:

$$S_H(\mathbf{y}, Q) = \frac{1}{\sigma^4}\{\mathrm{RSS} - 2\nu\sigma^2\} \tag{10}$$

where RSS is the usual residual sum-of-squares, on $\nu := n - p$ degrees of freedom. This is well-defined so long as $\nu > 0$.

When we are comparing normal linear models all with the same known variance $\sigma^2$, (10) is equivalent to $(\mathrm{RSS}/\sigma^2) + 2p$, Akaike's AIC for this case — which is known not to deliver consistent model selection. If instead we use the prequential score (8), with each term based on the univariate Hyvärinen score, we obtain

$$S_H^n = \sum_{i=p}^{n} \frac{1}{k_i^2 \sigma^4}(Z_i^2 - 2\sigma^2) \quad (n \geq p) \tag{11}$$

where $Z_i \sim \mathcal{N}(0, \sigma^2)$ is the difference between $Y_i$ and its least-squares predictor based on $(Y_1, \ldots, Y_{i-1})$, divided by $k_i$. Without the term $k_i^2$, (11) would be the same as (10), and so inconsistent. With it (even when $k_i \to 1$, which will typically be the case), the difference between the two expressions tends to infinity, and use of $S_H^n$ does indeed deliver consistent model selection. □

# 6 Conclusion

Proper scoring rules, of which there is a very great variety, supply a valuable and versatile extension to standard statistical theory based on the likelihood function. Many of the standard results can be applied, with little modification, in this more general setting. Homogeneous proper scoring rules, which do not make any use of normalising constant of a distribution, prove particularly useful in cases where that constant is computationally intractable, or even non-existent. We have illustrated the application of proper scoring rules for parameter estimation and Bayesian model selection. We believe that there will be many other problems for which they will supply a valuable additional tool in the statistician's kitbag.

## References


Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, **59**, 77–93.

Dawid, A. P., Lauritzen, S., and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *Annals of Statistics*, **40**, 593–608.

Dawid, A. P. and Musio, M. (2013). Estimation of spatial processes using local scoring rules. *AStA Advances in Statistical Analysis*, **97**, 173–9. doi:10.1007/s10182-012-0191-8.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, **14**, 107–14.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning*, **6**, 695–709.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Statistica Sinica (2011). Special issue on composite likelihood. *Statistica Sinica*, **21**, (1). http://www3.stat.sinica.edu.tw/statistica/j21n1/21-1.html.

Parry, M. F., Dawid, A. P., and Lauritzen, S. L. (2012). Proper local scoring rules. *Annals of Statistics*, **40**, 561–92.