LOCAL SCORING RULES AND STATISTICAL INFERENCE IN UNNORMALIZED MODELS

Matthew Parry[1]

[1]Dept of Mathematics & Statistics, University of Otago,
P.O.Box 56, Dunedin 9054, NEW ZEALAND, mparry@maths.otago.ac.nz

ABSTRACT

A scoring rule is a principled way to assess probabilistic forecasts. Associated with every scoring rule is a divergence and a concave entropy. Conversely, every scoring rule can be generated by a concave entropy. Scoring rules can also be straightforwardly adapted to statistical inference. The resulting estimating equations are unbiased but typically entail some loss of efficiency. Local scoring rules are a class of scoring rules with the remarkable property that they do not depend on the normalization of the quoted probability distribution. Consequently, they allow inference in unnormalized statistical models, i.e. models in which the normalization is either difficult or impossible to compute. Local scoring rules provide a unifying framework in which to understand existing approaches to such intractable problems, for example pseudolikelihood, score matching and ratio matching.

1. INTRODUCTION

Scoring rules have long been used to evaluate probabilistic forecasts [1, 2]. If Q stands for the forecaster's distribution – Q is used to denote *quote* – then $S(x, Q)$ is the score given to the forecaster when outcome x is observed. Scoring rules fit into standard statistical decision theory: a scoring rule is a loss function where the action is to quote a probability distribution [3].

A key feature of scoring rules is that they can be crafted to elicit a forecaster's honestly held belief. Such a scoring rule is said to be *proper*. We can usefully overload the definition of a scoring rule by defining a forecaster's *expected score* when $X \sim P$ as $S(P, Q)$. A proper scoring rule has the property that $S(P, Q) \geq S(P, P)$ for $Q \neq P$. In other words, if a forecaster thinks $X \sim P$, they will minimize their expected score by quoting P.

Scoring rules connect naturally to information theory. Every proper scoring rules gives rise to a *divergence*

$$d(P, Q) := S(P, Q) - H(P), \tag{1}$$

where $H(P) := S(P, P)$ is the (concave) *entropy* associated with the scoring rule. As we will see, we can profitably reverse this connection and use a concave entropy function to generate a scoring rule.

2. STATISTICAL INFERENCE

Although scoring rules are formulated to evaluate predictions, they can be easily turned to the task of estimation. Given a parametric model Q_θ, we have the obvious estimator

$$\widehat{\theta}(x) = \arg\min_\theta S(x, Q_\theta). \tag{2}$$

Typically, this amounts to solving the estimating equation $\partial S(x, Q_\theta)/\partial\theta = 0$. When the scoring rule is proper, the estimating equation is unbiased [4] since at $\theta = \theta_0$, $0 = \partial S(Q_{\theta_0}, Q_\theta)/\partial\theta = \mathbb{E}_{\theta_0}\partial S(X, Q_\theta)/\partial\theta$. As a result, inference via scoring rules fits into the established theory of unbiased estimating equations. An important consequence is that typically there will be a loss of efficiency. If we define $D := \mathbb{E}_\theta\, \partial^2 S/\partial\theta^2$ and $J := \mathbb{E}_\theta\, (\partial S/\partial\theta)^2$, then the *Godambe* or *sandwich* information G cannot exceed the Fisher information F [5]:

$$G := DJ^{-1}D \leq F. \tag{3}$$

3. LOCAL SCORING RULES

The *logarithmic scoring rule* or *log score* is the simplest example of a scoring rule:

$$S(x, Q) = -\log q(x). \tag{4}$$

It is straightforward to see that using this for statistical inference amounts to maximum likelihood estimation. Furthermore, the divergence and entropy associated with the log score are simply the Kullback-Leibler divergence and Shannon entropy, respectively. It is also possible to show that the log score is the only scoring rule that depends on the value of the quoted probability distribution at the observed outcome x and no other (counterfactual) outcome.

For this reason, we call the log score *strongly local*. The main idea of this paper is the concept of a *local* scoring rule. A local scoring rule is a rule that depends on the quoted distribution at x and on a "neighbourhood" of x. It turns out that local scoring rules are of the form [6]

$$S(x, Q) = -\lambda \log q(x) + S_0(x, Q), \tag{5}$$

where $\lambda \geq 0$ and $S_0(x, Q)$ is a 0-homogeneous function of Q. Consequently, when $\lambda = 0$, we obtain a scoring rule that does not depend on the normalization of the quoted probability distribution.

4. ENTROPY AND SCORING RULES

Under mild regularity conditions, it can be shown [2, 7, 8] that $S(x, Q)$ is a scoring rule if and only if there exists a concave function $H(Q)$ such that

$$S(x, Q) = H(Q) + H^\star(x, Q) - H^\star(Q, Q), \quad (6)$$

where $H^\star(\cdot, Q)$ is a *subgradient* of H at x. Furthermore, $H(Q)$ is the entropy associated with the scoring rule. In practice, $H^\star(\cdot, Q)$ is often the gradient, in which case the scoring rule is uniquely defined by $H(Q)$.

The idea of locality amounts to requiring $H(Q) - H^\star(Q, Q) = \lambda$, where λ is a Q-independent constant. This essentially requires $H(Q)$ to be of the form $H(Q) = -\lambda\, q(x) \log q(x) + H_1(Q)$, where $H_1(Q)$ is a 1-homogeneous function of Q. Eq. (5) then follows, with $\lambda \geq 0$ required for propriety.

4.1. Continuous outcome spaces

Let $q(x)$ be a sufficiently differentiable and strictly positive probability density on a continuous outcome space and let ϕ be a 1-homogeneous concave function of $\{q(x), q'(x), \ldots, q^{(k)}(x)\}$ for all x. Then

$$H(Q) = \int \mathrm{d}x\, \phi\left(x, q(x), q'(x), \ldots, q^{(k)}(x)\right) \quad (7)$$

is 1-homogeneous and generates a local scoring rule of order $2k$ in the derivatives of $q(x)$. In this case, the neighbourhood of x is an infinitesimal neighbourhood about x.

We expect only rules of order 2 and 4 to be of practical use. Second order rules take the form [6, 9]

$$S(x, Q) = \left(-\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial}{\partial q'} + \frac{\partial}{\partial q}\right)\phi[q], \quad (8)$$

where $\phi[q] = \phi(x, q, q')$. The simplest case, which occurs when $\phi[q] = -\frac{1}{2}\frac{q'^2}{q}$, was discovered by Almeida & Gidas [10] and by Hyvärinen [11], who dubbed it *score matching*:

$$S(x, Q) = \frac{q''(x)}{q(x)} - \frac{1}{2}\left(\frac{q'(x)}{q(x)}\right)^2. \quad (9)$$

4.2. Discrete outcome spaces

On a discrete outcome space, the neighbourhood is defined by an undirected graph G, i.e. y is in the neighbourhood of x if y is in the connection set of x. This relationship is also symmetric. Local scoring rules are then generated by the entropies of the form [12]

$$H(Q) = \sum_{K \in \mathcal{M}} \phi^K(Q_K), \quad (10)$$

where $\mathcal{M}$ is the set of maximal cliques, ϕ^K is 1-homogeneous and concave, and Q_K is the quoted distribution restricted to clique K. Specifically,

$$S(x, Q) = \sum_{K \in \mathcal{M}_x} \frac{\partial}{\partial q_x}\phi^K(Q_K), \quad (11)$$

where $\mathcal{M}_x$ are the maximal cliques containing x, and $q_x := q(x)$. Note that decomposition of the entropy into functions over cliques is directly analogous to the Hammersley-Clifford theorem for the factorization of a joint probabilty distribution on a graph [13].

5. EXAMPLES

5.1. Pseudolikelihood

Suppose $x = (x^1, \ldots, x^N)$ is an outcome from a product space. Let $x^{\setminus i} := \{x^k | k \neq i\}$. Then the *pseudolikelihood* [14] is

$$\mathrm{PL}(P; X = x) := \prod_i P(X^i = x^i | X^{\setminus i} = x^{\setminus i}). \quad (12)$$

Defining $y \in \mathrm{nhd}(x)$ if and only if $y^{\setminus i} = x^{\setminus i}$ for some i, then for each i and $y^{\setminus i}$, $K_{i,y^{\setminus i}} = \{x | x^{\setminus i} = y^{\setminus i}\}$ is a clique. Simplifying our notation so that $\phi^K(Q_K) = \phi_i(Q_K)$,

$$S(x, Q) = \sum_i S_i\left(x^i, Q(\cdot | X^{\setminus i} = x^{\setminus i})\right) \quad (13)$$

is a scoring rule, where the S_i are individual scoring rules for a single variable. Using the log score for each S_i justifies the use of the pseudolikelihood for inference:

$$S_{\mathrm{PL}}(x, Q) := -\ln \mathrm{PL}(Q; X = x) = \sum_i \ln \frac{q_{x^{\setminus i}}}{q_x}. \quad (14)$$

In the case of binary data outcomes, using the *Brier score*, $S(x, Q) = (x - Q(X = 1))^2$, for each S_i gives Hyvärinen's *ratio matching* method [15]. See [16] for a spatial modelling application.

5.2. Overdispersion

Overdispersion is the observation that statistical models do not always capture the amount of variation seen in the data. In many cases, overdispersion is due to the fact that there are unknown or unrecorded predictor variables. More subtly, overdispersion may indicate a breakdown of the assumption of independent observations. A phenomenological solution that is sometimes appropriate is to introduce a *dispersion parameter* ϕ to quantify the "anomalous" variation, namely $\mathrm{var}\, Y \to \phi\, \mathrm{var}\, Y$, where we expect $\phi > 1$.

The estimating equation for ϕ is often ad hoc. An obvious approach, however, is to suppose the effective number of observations is n/ϕ, leading to the updated probability model:

$$q(y) \to q(y|\phi) = \frac{q(y)^{1/\phi}}{Z(\phi)}. \quad (15)$$

Since the normalization $Z(\phi)$ will typically be impossible to compute, the usual methods of estimation will come up short. The local scoring rule in eq. (9), however, gives a delightfully simple expression:

$$\widehat{\phi} = -\frac{\ell(y)^2}{\ell'(y)}, \quad (16)$$

where $\ell(y) := q'(y)/q(y)$.

5.3. Sequential prediction

We end with a speculative application of local scoring rules. Suppose we observe $(x_1, \ldots, x_n)$ iid outcomes and wish to make a probabilistic prediction for x_{n+1}. Given a parametric model Q_θ, if $\widehat{\theta}_n$ is a consistent estimator for θ, then we would quote $Q_n := Q_{\widehat{\theta}_n}$ for x_{n+1}. When we are in the model, the desired result is $\mathbb{E}_\theta \, \mathrm{KL}(P, Q_n) \to \frac{1}{2} n^{-1}$, where KL is the Kullback-Leibler divergence.

Normalized maximum likelihood gives the optimal sequential prediction [17]:

$$q(x_{n+1}) = \frac{q(x_{n+1}|\widehat{\theta}_{n+1}(x_{1:n}, x_{n+1}))}{\sum_y q(y|\widehat{\theta}_{n+1}(x_{1:n}, y))}, \qquad (17)$$

where $\widehat{\theta}_{n+1}$ is the maximum likelihood estimate. Unfortunately, however, the denominator is often infinite. One might hope that there exists an appropriate local scoring rule that gives rise to a different estimator $\widehat{\theta}$ and a divergence which does not depend on normalization of the prediction. Admittedly, the loss of efficiency detailed in eq. (3) means we can expect only kn^{-1} convergence with $k \geq \frac{1}{2}$, but this may be an acceptable price to pay for tractability.

6. CONCLUSION

In addition to evaluating predictions, scoring rules provide a useful approach to statistical estimation. The requirement that a scoring rule be local gives rise to a surprising class of scoring rules that do not depend on the normalization of the quoted probability distribution. Consequently, inference can be carried out in models for which the normalization is either difficult or impossible to compute. Local scoring rules also appear to provide a unifying framework in which to understand existing approaches to such intractable problems. It would be interesting to see whether local scoring rules could be adapted to so called doubly intractable problems in Bayesian inference.

7. ACKNOWLEDGMENTS

It is a real pleasure to thank the organizers for their kind invitation to speak at WITMSE 2014. The work discussed here is joint work with Philip Dawid and Steffen Lauritzen. I acknowledge the financial support of a University of Otago travel grant.

8. REFERENCES

[1] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, pp. 1–3, 1950.

[2] J. McCarthy, "Measures of the value of information," *Proc. Nat. Acad. Sci.*, vol. 42, pp. 654–655, 1956.

[3] Peter D. Grünwald and Alexander Philip Dawid, "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory," *Annals of Statistics*, vol. 32, pp. 1367–1433, 2004.

[4] A. Philip Dawid and Steffen L. Lauritzen, "The geometry of decision theory," in *Proceedings of the Second International Symposium on Information Geometry and its Applications*. University of Tokyo, 2005, pp. 22–28.

[5] V. P. Godambe, "An optimum property of regular maximum likelihood estimation," *Ann. Math. Statist.*, , no. 4, pp. 1208–1211, 1960.

[6] M. Parry, A. P. Dawid, and S. Lauritzen, "Proper local scoring rules," *Annals of Statistics*, vol. 40, pp. 561–592, 2012.

[7] A. D. Hendrickson and R. J. Buehler, "Proper scores for probability forecasters," *Ann. Math. Statist.*, vol. 42, pp. 1916–1921, 1971.

[8] Tilmann Gneiting and Adrian E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, pp. 359–378, 2007.

[9] Werner Ehm and Tilmann Gneiting, "Local proper scoring rules of order two," *Annals of Statistics*, vol. 40, pp. 609–637, 2012.

[10] M. P. Almeida and B. Gidas, "A variational method for estimating the parameters of mrf from complete or incomplete data," *The Annals of Applied Probability*, vol. 3, pp. 103–136, 1993.

[11] Aapo Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning*, vol. 6, pp. 695–709, 2005.

[12] A. P. Dawid, S. Lauritzen, and M. Parry, "Proper local scoring rules on discrete sample spaces," *Annals of Statistics*, vol. 40, pp. 593–608, 2012.

[13] G. R. Grimmett, "A theorem about random fields," *Bull. Lond. Math. Soc.*, pp. 81–84, 1973.

[14] J. Besag, "Statistical analysis of non-lattice data," *J. Roy. Statist. Soc. Ser. D*, vol. 24, pp. 179–195, 1975.

[15] Aapo Hyvärinen, "Some extensions of score matching," *Computational Statistics and Data Analysis*, vol. 51, pp. 2499–2512, 2007.

[16] A. P. Dawid and M. Musio, "Estimation of spatial processes using local scoring rules," *Advances in Statistical Analysis*, pp. 173–179, 2013.

[17] Peter Grünwald, "A tutorial introduction to the minimum description length principle," in *Advances in Minimum Description Length: Theory and Applications*. 2005, MIT Press.