AISM (2007) 59:77–93
DOI 10.1007/s10463-006-0099-8

The geometry of proper scoring rules

A. P. Dawid

Received: 25 April 2006 / Revised: 18 August 2006 /
Published online: 16 December 2006

Abstract A decision problem is defined in terms of an outcome space, an action space and a loss function. Starting from these simple ingredients, we can construct: Proper Scoring Rule; Entropy Function; Divergence Function; Riemannian Metric; and Unbiased Estimating Equation. From an abstract viewpoint, the loss function defines a duality between the outcome and action spaces, while the correspondence between a distribution and its Bayes act induces a self-duality. Together these determine a "decision geometry" for the family of distributions on outcome space. This allows generalisation of many standard statistical concepts and properties. In particular we define and study generalised exponential families. Several examples are analysed, including a general Bregman geometry.

Keywords Bregman geometry · Decision geometry · Generalised exponential family · Information geometry · Proper scoring rule · Unbiased estimating equation

1 Introduction

Consider a statistical decision problem $(\mathcal{X}, \mathcal{A}, L)$, defined in terms of an *outcome space* $\mathcal{X}$, *action space* $\mathcal{A}$, and real-valued loss function L. Letting $\mathcal{P}$ be a suitable class of distributions over $\mathcal{X}$ such that $L(P, a) := \mathrm{E}_{X \sim P} L(X, a)$ exists for all $a \in \mathcal{A}$, $P \in \mathcal{P}$, we introduce, for $P, Q \in \mathcal{P}$, $x \in \mathcal{X}$:

Bayes act $a_P := \arg\inf_{a \in \mathcal{A}} L(P, a)$

A. P. Dawid (✉)
Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK
e-mail: dawid@stats.ucl.ac.uk

Proper scoring rule $S(x, Q) := L(x, a_Q)$
Entropy function $H(P) := S(P, P)$
Divergence function $d(P, Q) := S(P, Q) - H(P)$

These quantities have special properties inherited from their construction (Dawid, 1998). In particular:

- $H(P)$ is concave in P
- $S(P, Q)$ is affine in P
- $S(P, Q)$ is minimised in Q at $Q = P$
- $d(P, Q) - d(P, Q_0)$ is affine in P
- $d(P, Q) \geq 0$, with equality if $Q = P$

Conversely, these properties essentially characterise entropy functions, scoring rules and divergence functions that can arise from a decision problem in this way. In Dawid (1998),Dawid and Sebastiani (1999) they are illustrated for a number of important cases, and used to determine the optimal choice of an experimental design.

1.1 Equivalence

Suppose that a new loss function is defined by $L^*(x, a) = c\,L(x, a) + k(x)$ $(c > 0)$. Then the Bayes acts are the same for L and L^*, and, defining $k(P) = \mathrm{E}_P\{k(X)\}$ (assumed to exist), we have

$$\begin{aligned} S^*(x, Q) &= c\,S(x, Q) + k(x) \\ H^*(P) &= c\,H(P) + k(P) \\ d^*(P, Q) &= c\,d(P, Q). \end{aligned}$$

In this case we call the two decision problems *equivalent*, and *strongly equivalent* if moreover $c = 1$. For most purposes we do not need to distinguish between equivalent problems.

2 Parameter estimation

Let $\mathcal{Q} = \{Q_\theta\} \subseteq \mathcal{P}$ be a smooth one-parameter family of distributions. Given data $\mathbf{x} = (x_1, \ldots, x_n)$, with empirical distribution $\widehat{P}_n \in \mathcal{P}$, a popular method of "estimating" θ is by the *minimum divergence* criterion:

$$\hat{\theta} := \arg\min_\theta d(\widehat{P}_n, Q_\theta). \tag{1}$$

When d derives from a decision problem as above, this is equivalent to *optimum score estimation* (Gneiting and Raftery, 2005), which operates by minimising the *cumulative empirical score*:

$$\hat{\theta} := \arg\min_{\theta} \sum_{i=1}^{n} S(x_i, \theta) \tag{2}$$

(where we abuse notation by writing $S(x, \theta)$ for $S(x, Q_\theta)$) – in which form it remains meaningful even when $\widehat{P}_n \notin \mathcal{P}$.

Defining now $s(x, \theta) := (\partial/\partial\theta)S(x, \theta)$, we see that $\hat{\theta}$ will satisfy the estimating equation

$$\sum_{i=1}^{n} s(x_i, \theta) = 0. \tag{3}$$

Theorem 1 $\mathrm{E}_\theta\, s(X, \theta) = 0$.

Proof The quantity $\mathrm{E}_{\theta_0}\, S(X, \theta)$ is minimised in θ at θ_0. Thus at $\theta = \theta_0$,

$$\begin{aligned} 0 &= (\mathrm{d}/\mathrm{d}\theta)\, \mathrm{E}_{\theta_0}\, S(X, \theta) \\ &= \mathrm{E}_{\theta_0}\, s(X, \theta). \end{aligned}$$

□

Corollary 2 *The estimating Eq. (3) is unbiased.*

We can thus apply standard results on unbiased estimating equations to describe the properties of the *minimum empirical score* estimator $\hat{\theta}$: in particular, it will typically be consistent (though not necessarily efficient).

The above results generalise readily to multi-dimensional parameter spaces.

3 Decision geometry

We now introduce a concrete framework within which we can naturally define and manipulate geometric properties associated with a decision problem. The theory outlined below can be made rigorous for the case of a finite outcome space $\mathcal{X}$, but is also indicative of properties that (under appropriate technical conditions) should hold more generally.

Let W be the vector space of all signed measures over $\mathcal{X}$, and V the vector space of all real functions on $\mathcal{X}$. These spaces are in natural duality with respect to the bilinear product $\langle m, f \rangle = \int f(x)\, \mathrm{d}m(x)$. In particular $\langle P, f \rangle = \mathrm{E}_{X \sim P}\{f(X)\}$ for P a probability distributions on $\mathcal{X}$. The set $\mathcal{P}$ of all distributions on $\mathcal{X}$ is a convex subset of W of codimension 1, and its relative interior $\mathcal{P}^\circ$ – the set of all everywhere positive probability distributions on $\mathcal{X}$ – is a differentiable manifold. At any $P \in \mathcal{P}^\circ$, the tangent space to $\mathcal{P}^\circ$ can be naturally represented as the subspace $W^+ := \{m \in W : m(\mathcal{X}) = 0\}$. This allows us to identify tangent vectors in different tangent spaces, so defining a flat *affine connexion* on $\mathcal{P}^\circ$ – the *mixture connexion* (Dawid, 1975) – which we denote by ∇.

The dual of W^+ is the quotient space $V^+ := V/\mathbf{1}$, where $\mathbf{1}$ denotes the one-dimensional space of constant functions on $\mathcal{X}$. We denote by π^+ the natural projection from V to V^+. For $v, v' \in V$ we write $v \sim^+ v'$ if $\pi^+(v) = \pi^+(v')$, i.e. the functions v and v' differ by a constant.

Consider now a decision problem $(\mathcal{X}, \mathcal{A}, L)$. With L understood, we henceforth identify $a \in \mathcal{A}$ with its loss function $L(\,\cdot\,, a)$, thus converting the action space into a subset $\mathcal{L}$ of V, which we shall assume closed and bounded from below. Allowing randomised acts, $\mathcal{L}$ is convex. Let $\mathcal{L}^*$ denote its lower boundary, consisting of the admissible acts. Without any essential effect, we henceforth replace $\mathcal{L}$ by the convex set $\{v \in V : v \geq a$ for some $a \in \mathcal{L}\}$, which has the same lower boundary $\mathcal{L}^*$. Then $T_P := \{v \in V : \langle P, v\rangle = H(P)\}$ is a supporting hyperplane to $\mathcal{L}$, and we can characterise $\mathcal{L}$ dually as $\{v \in V : \langle P, v\rangle \geq H(P)$, all $P \in \mathcal{P}\}$.

For present purposes we make the following *basic assumptions*:

(i) For any $P \in \mathcal{P}$ there is exactly one Bayes act $\mathbf{p} \in \mathcal{L}^*$. (We use corresponding upper case and boldface lower case symbols for a distribution in $\mathcal{P}$ and its Bayes act in $\mathcal{L}$.)

(ii) Distinct distributions in $\mathcal{P}$ have distinct Bayes acts in $\mathcal{L}^*$. Equivalently, the scoring rule S is *strictly* proper: $S(P, Q) > S(P, P)$ for $Q \neq P$.

(iii) Every $a \in \mathcal{L}^*$ is a Bayes act for some $P \in \mathcal{P}$.

The function $\lambda : \mathcal{P} \to \mathcal{L}^*$ taking each P to its Bayes act $\mathbf{p}$ is then a $(1, 1)$ correspondence. The supporting hyperplane T_P now becomes the tangent plane to $\mathcal{L}$ at $\mathbf{p}$, intersecting $\mathcal{L}$ at the single point $\mathbf{p}$.

We note the following identifications:

- The expected loss $L(P, a)$ is $\langle P\,,\, a\rangle$
- The Bayes act is the score function: $\mathbf{p}(\,\cdot\,) \equiv S(\,\cdot\,, P)$
- $S(P, Q)$ is $\langle P\,,\, \mathbf{q}\rangle$
- $H(P)$ is $\langle P\,,\, \mathbf{p}\rangle$
- $d(P, Q)$ is $\langle P\,,\, \mathbf{q} - \mathbf{p}\rangle$.

Now let $\mathcal{L}^+ := \pi^+(\mathcal{L}^*) \subseteq V^+$. Note that at most one member of a ray $v^+ := \{v + k : k \in \mathbb{R}\} \in V^+$ can be in $\mathcal{L}^*$, so that $\pi^+ : \mathcal{L}^* \to \mathcal{L}^+$ is a $(1, 1)$ correspondence.

Lemma 3 *$\mathcal{L}^+$ is convex.*

Proof We have to show that, for $P, Q \in \mathcal{P}$ and $0 \leq \omega \leq 1$, there exist $R \in \mathcal{P}$, $k \in \mathbb{R}$ such that $\mathbf{r}(x) \equiv \omega\,\mathbf{p}(x) + (1 - \omega)\;\mathbf{q}(x) - k$.

For $\Pi \in \mathcal{P}$, let $k(\Pi) := \omega\, S(\Pi, P) + (1 - \omega)\; S(\Pi, Q) - H(\Pi) = \omega\, d(\Pi, P) + (1 - \omega)\; d(\Pi, Q)$. This is a non-negative convex function on $\mathcal{P}$. Let $k := \inf_{\Pi \in \mathcal{P}} k(\Pi)$, and suppose that this infimum is attained at $R \in \mathcal{P}$. Also let $v := \omega\,\mathbf{p} + (1 - \omega)\,\mathbf{q} - k$.

For any $\Pi \in \mathcal{P}$, $\langle \Pi, v\rangle = \omega\,\langle \Pi, \mathbf{p}\rangle + (1 - \omega)\,\langle \Pi, \mathbf{q}\rangle - k = k(\Pi) + H(\Pi) - k \geq H(\Pi)$, whence $v \in \mathcal{L}$. Moreover $\langle R, v\rangle = H(R) = \inf_{a \in \mathcal{L}} \langle R, a\rangle$. Thus $v = \mathbf{r}$, the Bayes act for R, and the required property is demonstrated. □

We have thus shown that the map $\lambda^+ := \pi^+ \circ \lambda$ provides a (1,1) correspondence between the convex sets $\mathcal{P} \subseteq W$ and $\mathcal{L}^+ \subseteq V^+$. (Since the orientation of the tangent plane T_P in V to $\mathcal{L}$ at $\mathbf{p} = \lambda(P)$ is determined by P, we further see that, knowing λ^+, we can recover $\mathcal{L}^*$ and λ up to an unimportant translation by a constant.) This correspondence determines the *decision geometry* on $\mathcal{P}$ induced by the given decision problem. In particular, it represents $\mathcal{P}^\circ$ as an open convex subset of V^+, which in turn allows us to represent the tangent space to $\mathcal{P}^\circ$ at any of its points by V^+. This then allows us to identify tangent vectors at different points, thereby defining a flat affine connexion on $\mathcal{P}^\circ$, which we denote by ∇^* and term the *decision connexion*. Since this depends on the specific decision problem from which it is constructed, there is no general relationship between ∇^* and the mixture connexion ∇.

3.1 Estimating function

The image $\mathbf{p}^+ \in \mathcal{L}^+$ of $P \in \mathcal{P}$ is the set of functions on $\mathcal{X}$ of the form $S(\cdot, P) +$ constant. Hence in order to determine the decision geometry, we only need to know $S(\cdot, P)$ up to a constant – which could however depend on P.

Suppose then we are given a function $Z : \mathcal{X} \times \mathcal{P} \rightarrow \mathbb{R}$ such that we know only that $S(x, P) \sim^+ Z(x, P)$, i.e. $S(x, P)$ has the form $Z(x, P) - k(P)$. For a parametric family $\mathcal{Q} = \{Q_\theta\}$ we then have $S(x, \theta) = Z(x, \theta) - k(\theta)$. Defining $z(x, \theta) := \partial\, Z(x, \theta)/\partial\, \theta$, we thus know that $s(x, \theta)$ is of the form $z(x, \theta) - \dot{k}(\theta)$ (with a dot denoting differentiation with respect to θ). Applying Corollary 2, we see that

$$\dot{k}(\theta) = \mathrm{E}_\theta\{z(X, \theta)\}. \tag{4}$$

We can integrate this with respect to θ to obtain $k(\theta)$ – and thence the function $k : \mathcal{P} \rightarrow \mathbb{R}$ (up to an additive scalar).

We term $z(x, \theta)$ the *estimating function*. Using (4), the estimating equation (3) is equivalent to equating $\sum_{i=1}^n z(x_i, \theta)$ to its expectation under P_θ.

3.2 Differential geometry

The tangent space to the manifold $\mathcal{P}^\circ \subseteq W$ at any point is concretely represented by W^+, and that to $\mathcal{L}^+$ at any point by V^+. Under our basic assumptions, the function λ^+ is differentiable at $P \in \mathcal{P}^\circ$, and its derivative supplies an isomorphism between W^+ and V^+, which links the two representations of any tangent vector at P. Also through this isomorphism, the negative of the natural bilinear product is converted into a inner product on W^+, so defining a metric g – the *decision metric* – on $\mathcal{P}^\circ$: multiplying this by 1/2 yields the local form of the divergence d. These constructions and properties, which are special cases of the general theory of Lauritzen (1987a), make $(\mathcal{P}^\circ, g, \nabla, \nabla^*)$ a *dually flat statistical manifold* (Amari and Nagaoka, 1982; Lauritzen, 1987b; Amari and Nagaoka, 2000). We again remark that ∇ is always the mixture connexion, whereas the dual connexion ∇^* and the metric g will depend on the specific decision problem.

However, much of the geometric framework can be fruitfully applied at a global level, without invoking the differentiable structure. We illustrate this below.

4 Generalised exponential family

Let $\mathcal{F}$ be the intersection of some affine subspace of V^+ with $\mathcal{L}^+$, and $\mathcal{E} = (\lambda^+)^{-1}(\mathcal{F})$ the corresponding subfamily of $\mathcal{P}$. We call such $\mathcal{E}$ a *linear generalised exponential family* (LGEF). A 1-dimensional LGEF is a ∇^*-geodesic. Through its identification with the convex subset $\mathcal{F}$ of V^+, a LGEF $\mathcal{E}$ inherits an affine parametrisation.

Since $\mathbf{q}(\cdot) \equiv S(\cdot, Q)$, a LGEF $\mathcal{E} = \{Q_\beta : \beta \in \mathcal{B} \subseteq \mathbb{R}^k\}$, with an affine parametrisation, is thus defined by the *linear loss*, or equivalently *linear score*, property (Grünwald and Dawid, 2004, Sect. 7.2):

$$S(x, Q_\beta) \equiv \beta_0 + m(x) + \sum_{i=1}^{k} \beta_i\, t_i(x), \tag{5}$$

for some m, $t_i \in V$, with β_0 then a uniquely determined function of β. Applying Corollary 2 we find $d\beta_0/d\beta_i = -\mathrm{E}_{Q_\beta}\{t_i(X)\}$ $(\beta \in \mathcal{B}^\circ)$.

Note that the property of being a LGEF is unaffected if we replace our underlying decision problem by an equivalent one, as described in Sect. 1.1.

Let $t := (t_1, \ldots, t_k)$, and define, for $\tau \in \mathbb{R}^k$: $\Gamma_\tau := \{P \in \mathcal{P} : \mathrm{E}_P\{t(X)\} = \tau\}$. Suppose that there exists $P_\tau \in \Gamma_\tau \cap \mathcal{E}$ (note that this need not hold in general: see below.) Since $S(P, Q) = \langle P\, ,\, \mathbf{q}\rangle$, an easy calculation, using (5), yields:

$$\langle P - P_\tau\, ,\, \mathbf{p}_\tau - \mathbf{q}\rangle = 0 \qquad (P \in \Gamma_\tau, Q \in \mathcal{E}).$$

This in turn implies the "Pythagorean equality":

$$d(P, P_\tau) + d(P_\tau, Q) = d(P, Q) \qquad (P \in \Gamma_\tau, Q \in \mathcal{E}). \tag{6}$$

It readily follows that, for any $P \in \Gamma_\tau$,

$$P_\tau = \arg\min_{Q \in \mathcal{E}} d(P, Q). \tag{7}$$

When P is the empirical distribution $\widehat{P}_n$ of data $(x_1, \ldots, x_n)$ from $\mathcal{X}$, if there exists $P_{\bar{t}} \in \mathcal{E}$ satisfying $E_{P_{\bar{t}}}\{t(X)\} = \bar{t} := n^{-1}\sum_{i=1}^{n} t(x_i)$, then this will minimise the empirical score $\sum_{i=1}^{n} S(x_i, Q)$ over $Q \in \mathcal{E}$.

Now fix $Q \in \mathcal{P}$, take $m \equiv \mathbf{q}$, and, for given $t_i \in V$, let $\mathcal{E}$ be given by (5): then $\mathcal{E}$ is a LGEF containing Q. Again, if there exists $P_\tau \in \Gamma_\tau \cap \mathcal{E}$ then (6) holds for all $P \in \Gamma_\tau$, whence we readily deduce

$$P_\tau = \arg \min_{P \in \Gamma_\tau} d(P, Q). \tag{8}$$

What happens if, for some τ, $\Gamma_\tau \neq \emptyset$ but $\Gamma_\tau \cap \mathcal{E} = \emptyset$? In this case P_τ can still be defined by (8), but will not now be in $\mathcal{E}$ (P_τ will in fact lie on the boundary of $\mathcal{P}$). However it turns out (Grünwald and Dawid, 2004, Sect. 10) that, under mild conditions, such a case will satisfy the still stronger Pythagorean *inequality*:

$$d(P, P_\tau) + d(P_\tau, Q) \leq d(P, Q) \qquad (P \in \Gamma_\tau). \tag{9}$$

The family $\mathcal{E}^m \supseteq \mathcal{E}$ of all P_τ given by (8), for all τ such that $\Gamma_\tau \neq \emptyset$, constitutes a *full* generalised exponential family. In general this will not be flat – indeed, even in simple problems it need not correspond to a smooth submanifold of V^+ (Grünwald and Dawid, 2004, Example 7.1).

One might conjecture that, for any $P \in \Gamma_\tau$, (7) will continue to hold, in the form $P_\tau = \arg\min_{Q \in \mathcal{E}^m} d(P, Q)$ – but this need not be so (Grünwald and Dawid, 2004, Sect. 7.6.1).

Grünwald and Dawid (2004) investigate further properties of a statistical decision problem, using convex duality and saddle-points in an associated game against nature. These include but extend beyond properties of generalised exponential families. It is likely that many of these properties can be given interesting geometric interpretations within the framework set out above. However in order to incorporate the full generality of this game-theoretic approach into our geometrical framework it would be necessary to find ways of relaxing the basic assumptions (i) and (ii).

5 Examples

Because any loss function determines a decision geometry, the class of these is very wide. We consider a few special examples of some interest. Other examples can be based on the various scoring rules presented in Dawid (1998),Dawid and Sebastiani (1999),Gneiting and Raftery (2005).

For the following examples we take $\mathcal{P} = \mathcal{A} =$ the set $\mathcal{M}$ of all distributions on $\mathcal{X}$ that are absolutely continuous with respect to a given σ-finite measure μ. For $Q \in \mathcal{P}$ we denote the density of Q with respect to μ by $q(\cdot)$, etc.

5.1 Logarithmic score and information geometry

Consider the loss function:

$$S(x, Q) = -\log q(x). \tag{10}$$

Then $S(P,Q) = -\int_{\mathcal{X}} p(x)\log q(x)\,\mathrm{d}\mu(x)$ is well-defined, and, as is well-known (Cover and Thomas, 1991), is uniquely minimised in Q for $Q = P$, so that S is a strictly proper scoring rule.

Correspondingly we have entropy function

$$H(P) = -\int p(t)\log p(t)\,\mathrm{d}\mu(t), \tag{11}$$

the *Shannon entropy* of P with respect to μ; and divergence function

$$d(P,Q) = \int_{\mathcal{X}} p(t)\log\{p(t)/q(t)\}\ \mathrm{d}\mu(t), \tag{12}$$

the *Kullback–Leibler divergence* between P and Q. The total empirical score (2) is the negative log-likelihood function, and the minimum divergence estimator $\hat{\theta}$ given by (1) is the maximum likelihood estimator. The function $s(x,\theta)$ becomes Fisher's (*efficient*) *score* function (not to be confused with *scoring rule*), and the unbiased estimating Eq. (3) becomes the usual *likelihood equation*.

The Bayes act $\mathbf{p} \in \mathcal{L}^*$ corresponding to $P \in \mathcal{P}$ is the negative log-density function, $\mathbf{p}(x) \equiv -\log p(x)$; and its image $\pi^+(P)$ in $\mathcal{L}^+$ is the set of negative log-densities of positive multiples of P. Given any such function $\mathbf{p}'$ we can recover $\mathbf{p} \sim^+ \mathbf{p}'$ from the normalisation condition $\int \exp(-\mathbf{p})\,\mathrm{d}\mu = 1$.

In this case the decision geometry reduces to the familiar information geometry: the decision connexion ∇^* is the exponential connexion (introduced as the "Efron connexion" in Dawid, 1975), the decision metric is the Fisher metric, and a "generalised' exponential family" is just an ordinary exponential family.

5.2 Bregman geometry

We now investigate the general structure of a decision problem whose score function $S(x,Q)$ is determined, up to a constant possibly dependent on Q, by the value of the density function $q(\cdot)$ at x:

$$S(x,Q) \sim^+ Z(x,Q) := -\xi\{q(x)\} \tag{13}$$

for some smooth function $\xi : \mathbb{R} \to \mathbb{R}$. For ease of further analysis we also express ξ as ψ', where $'$ denotes derivative.

Let $\{Q_\theta : \theta \in R\}$ be a smooth parametric family in $\mathcal{P}$. Then $z(x,\theta) = -\psi''\{q_\theta(x)\}\,\dot{q}_\theta(x)$. Hence from (4)

$$\dot{k}(\theta) = -\int q_\theta(t)\,\psi''\{q_\theta(t)\}\,\dot{q}_\theta(t)\,\mathrm{d}\mu(t), \tag{14}$$

whence we can take $k(\theta) = \kappa(\theta) - \kappa(0)$, with

$$\kappa(\theta) := \int \left[\psi\{q_\theta(t)\} - q_\theta(t)\,\psi'\{q_\theta(t)\}\right] \mathrm{d}\mu(t).$$

Since $\kappa(0)$ depends only on Q_0 and not otherwise on the family $\{Q_\theta\}$, we deduce that, for all $Q \in \mathcal{P}$, up to a constant

$$k(Q) = \int \left[\psi\{q(t)\} - q(t)\,\psi'\{q(t)\}\right] \mathrm{d}\mu(t) \tag{15}$$

and so

$$S(x, Q) = -\psi'\{q(x)\} - \int \left[\psi\{q(t)\} - q(t)\,\psi'\{q(t)\}\right] \mathrm{d}\mu(t). \tag{16}$$

So long as ξ is increasing, ψ is convex, and then this recovers the general form of a (*separable*) *Bregman score* (Grünwald and Dawid, 2004, Eq. 34). The corresponding *Bregman divergence* (Bregman, 1967; Csiszár, 1991) is

$$d(P, Q) = \int \Delta\{p(t), q(t)\}\,\mathrm{d}\mu(t) \tag{17}$$

with

$$\Delta(a, b) := \psi(a) - \psi(b) - \psi'(b)\,(a - b). \tag{18}$$

This is non-negative for ψ convex, so that then S is indeed a proper scoring rule. The associated *Bregman entropy* is

$$H(P) = -\int \psi\,\{p(t)\}\;\mathrm{d}\mu(t). \tag{19}$$

By construction, the image $\mathbf{p}^+ \in \mathcal{L}^+$ of $P \in \mathcal{P}$ is the set of translates of the function $-\xi\{p(x)\}$. With respect to the decision connexion ∇^*, the convex combination P_ω of P_0 with weight $1 - \omega$ and P_1 with weight ω satisfies

$$\xi\{p_\omega(x)\} = (1 - \omega)\,\xi\{p_0(x)\} + \omega\,\xi\{p_1(x)\} - c_\omega, \tag{20}$$

where the "normalising constant" c_ω is chosen to make P_ω a probability distribution. As ω varies, P_ω traces a ∇^*-geodesic. Correspondingly a LGEF $\{P_\beta\}$ has the form

$$\xi\{p_\beta(x)\} = \beta_0 + m(x) + \sum_{i=1}^{k} \beta_i\,t_i(x).$$

Now let $M, N \in W^+$, absolutely continuous with respect to μ with densities $m(\cdot)$, $n(\cdot)$, be tangent vectors to $P \in \mathcal{P} \subseteq W$. The counterpart $\mathbf{n}$ to N in the tangent space V^+ at $\mathbf{p}^+ \in \mathcal{L}^+$ is the set of translates of the function

$$\frac{\partial}{\partial\theta}\left[-\xi\{p(x)+\theta n(x)\}\right]\Big|_{\theta=0} = -\xi'\{p(x)\}\, n(x).$$

Thus the *Bregman metric* is the inner product $(\,\cdot\,,\,\cdot\,)_P$ on W^+ given by

$$\begin{aligned}(M,N)_P = -\langle M,\mathbf{n}\rangle &= \int \xi'\{p(x)\}\, n(x)\,\mathrm{d}M(x)\\ &= \int \xi'\{p(x)\}\, m(x)\, n(x)\,\mathrm{d}\mu(x).\end{aligned} \tag{21}$$

In particular, for a parametric family $\{P_\theta\}$, the divergence between P_θ and $P_{\theta+\mathrm{d}\theta}$ is $\frac{1}{2}g(\theta)\,\mathrm{d}\theta^2$, with

$$g(\theta) = \int \xi'\{p_\theta(x)\}\,\{\dot{p}_\theta(x)\}^2\;\mathrm{d}\mu(x).$$

For properties and applications of Bregman geometry see Eguchi (2005),Murata et al. (2004). However most of these do not depend on the Bregman form, and continue to hold for a general decision geometry.

5.2.1 An extension

Much of the above analysis still goes through if we relax the requirement (13) to:

$$S(x,Q) \sim^+ Z(x,Q) := -\xi\{x,q(x)\}. \tag{22}$$

Again we introduce ψ such that $\xi(x,q) = \psi'(x,q)$, where $'$ now denotes differentiation with respect to the second argument, q. We require that, for each x, ψ be convex in q (equivalently, ξ be non-decreasing in q). We obtain

$$k(Q) = \int \left[\psi\{t,q(t)\} - q(t)\,\psi'\{t,q(t)\}\right]\,\mathrm{d}\mu(t) \tag{23}$$

$$S(x,Q) = -\psi'\{x,q(x)\} - \int \left[\psi\{t,q(t)\} - q(t)\,\psi'\{t,q(t)\}\right]\,\mathrm{d}\mu(t) \tag{24}$$

$$H(P) = -\int \psi\,\{t,p(t)\}\;\mathrm{d}\mu(t) \tag{25}$$

$$d(P,Q) = \int \Delta\{t,p(t),q(t)\}\,\mathrm{d}\mu(t) \tag{26}$$

with

$$\Delta(t,a,b) := \psi(t,a) - \psi(t,b) - \psi'(t,b)\,(a-b). \tag{27}$$

We may describe the constructions above as *extended Bregman* score etc.

In particular, taking $\psi(x,q)$ of the form $a(x)\,\psi(q)$ for a non-negative function a effectively allows us to use a different integrating measure in (16), (17) and (19) from that used for defining probability densities.

5.3 Special cases

5.3.1 Logarithmic score

The logarithmic score of Sect. 5.1 is the special case of the Bregman score for $\xi(q) \equiv \log q$ (equivalently, $\psi(q) \equiv q\log q - q$). In this case, though not in general, the geometry is independent of the choice of base measure μ.

It is well known (Bernardo 1979) that the logarithmic score is essentially the only proper scoring rule having the form $S(x,Q) \equiv -\xi\{q(x)\}$, i.e. with $k(Q)$ in (15) independent of Q. In fact more is true: it is essentially the only proper scoring rule having the form $S(x,Q) = -\xi\{x,q(x)\}$, i.e. with $k(Q)$ in (23) independent of Q. To see this, note that for an extended Bregman score the analogue of (14) is

$$\dot{k}(\theta) = -\int q_\theta(t)\,\psi''\{t,q_\theta(t)\}\,\dot{q}_\theta(t)\,\mathrm{d}\mu(t), \tag{28}$$

which has to vanish in such a case. Because the family $\{Q_\theta\}$ is arbitrary, we must thus have, for all $Q \in \mathcal{P}$:

$$\int q(t)\,\psi''\{t,q(t)\}\,\mathrm{d}m(t) = 0 \tag{29}$$

for all $m \in W^+$ with $m \ll \mu$. This can only hold if $q(t)\,\psi''\{t,q(t)\}$ is constant in t [a.e. μ], and thus of the form $\kappa(Q)$, which in turn yields

$$\psi'\{x,q(x)\} = k(Q)\log q(x) + a(x). \tag{30}$$

Moreover, since the distribution Q does not explicitly enter the left-hand side of (30), we must in fact have

$$\xi(x,q) = k\log q + a(x), \tag{31}$$

equivalent to the logarithmic score.

5.3.2 Brier geometry

If we take $\xi(q) \equiv q$, so that $\psi(q) \equiv \frac{1}{2}q^2$, we obtain

$$S(x,Q) = \frac{1}{2}\int q(t)^2\,\mathrm{d}\mu(t) - q(x). \tag{32}$$

This defines the *Brier score*. The corresponding *Brier entropy* function is

$$H(P) = -\frac{1}{2}\int p(t)^2 \, \mathrm{d}\mu(x) \tag{33}$$

and the *Brier divergence* function is

$$d(P,Q) = \frac{1}{2}\int \{p(t) - q(t)\}^2 \, \mathrm{d}\mu(t). \tag{34}$$

The image $\mathbf{p}^+ \in \mathcal{L}^*$ corresponding to $P \in \mathcal{P}$ is now the set of translates of the negative density $-p(\,\cdot\,)$, and the decision connexion $\nabla^* = \nabla$, the mixture connexion. In particular a LGEF is just a mixture family.

The *Brier metric* is determined by the function

$$g(\theta) = \int \{\dot{p}_\theta(x)\}^2 \, \mathrm{d}\mu(x).$$

5.3.3 Tsallis geometry

Now for $\gamma \in \mathbb{R}$, $\gamma \neq 0, 1$, take $\psi(q) \equiv -\sigma q^\gamma$ with (for convexity) $\sigma = 1$ for $0 < \gamma < 1, \sigma = -1$ otherwise. We obtain the *Tsallis score*:

$$S(x,Q) = \sigma\left\{\gamma\, q(x)^{\gamma-1} - (\gamma - 1)\int q(t)^\gamma \, \mathrm{d}\mu(t)\right\}. \tag{35}$$

The corresponding *Tsallis entropy*, equivalent (when μ is Lebesgue measure) to that introduced by Tsallis (1988) in a physical context, is

$$H(P) = \sigma \int p(t)^\gamma \, \mathrm{d}\mu(t), \tag{36}$$

and the *Tsallis divergence* function is

$$d(P,Q) = \sigma\,\gamma \int p(t)\, q(t)^{\gamma-1} \, \mathrm{d}\mu(t) - (\gamma - 1)\, H(Q) - H(P). \tag{37}$$

The image $\mathbf{p}^+ \in \mathcal{L}^*$ of $P \in \mathcal{P}$ is the set of translates of $-\sigma\gamma\, p(x)^{\gamma-1}$. With respect to the decision connexion ∇^*, the convex combination P_ω of P_0 with weight $1 - \omega$ and P_1 with weight ω satisfies

$$p_\omega(x)^{\gamma-1} = (1-\omega)\, p_0(x)^{\gamma-1} + \omega\, p_1(x)^{\gamma-1} - c_\omega, \tag{38}$$

and a LGEF has the form

$$p_\beta(x) = \left\{\beta_0 + m(x) + \sum_{i=1}^{k} \beta_i\, t_i(x)\right\}^{1-\gamma}.$$

The *Tsallis metric* is defined by

$$g(\theta) = |\gamma\,(1-\gamma)| \int \{p_\theta(x)\}^{\gamma-2}\{\dot{p}_\theta(x)\}^2\, \mathrm{d}\mu(x).$$

Formula (38) may be contrasted with the α-*mixture*, or convex combination based on the α-connexion (Amari, 2005), which, for $\gamma = \frac{1}{2}(3-\alpha)$, can be expressed as

$$p_\omega(x)^{\gamma-1} = c_\omega \times \left\{(1-\omega)\, p_0(x)^{\gamma-1} + \omega\, p_1(x)^{\gamma-1}\right\}. \tag{39}$$

The α-connexion does not appear to be directly representable as a decision connexion.

6 Further examples

Here we consider some examples that are not of Bregman type.

6.1 Pseudospherical score

As a variation on Sect. 5.3.3, suppose we seek a proper scoring rule having the form:

$$S(x, Q) = c(Q)\, q(x)^{\gamma-1}.$$

From Theorem 1 we must have

$$\dot{c}(\theta) \int q_\theta(x)^\gamma\, \mathrm{d}\mu(x) + (\gamma-1) \int c(\theta)\, q_\theta(x)^{\gamma-1}\, \dot{q}_\theta(x)\, \mathrm{d}\mu(x) = 0,$$

whence

$$\frac{\dot{c}(\theta)}{c(\theta)} = -\frac{\gamma-1}{\gamma}\, \frac{\int \gamma\, q_\theta(x)^{\gamma-1}\, \dot{q}_\theta(x)\, \mathrm{d}\mu(x)}{\int q_\theta(x)^\gamma\, \mathrm{d}\mu}.$$

Integrating we obtain

$$c(\theta) \propto \|q_\theta\|_\gamma^{-(\gamma-1)},$$

where

$$\|q\|_\gamma := \left\{\int q(x)^\gamma \, \mathrm{d}\mu(x)\right\}^{\frac{1}{\gamma}}$$

is the L_γ norm of $q(\cdot)$. Thus, up to a scalar multiple, we must have

$$S(x, Q) = -\left\{\frac{q(x)}{\|q\|_\gamma}\right\}^{\gamma-1}. \tag{40}$$

For $\gamma > 1$ this is indeed a proper scoring rule, the *pseudospherical* (spherical for $\gamma = 2$) *score* (Good 1971; Gneiting and Raftery 2005). Correspondingly we have

$$\begin{aligned}
H(P) &= -\|p\|_\gamma \\
d(P, Q) &= \|p\|_\gamma - \frac{\int p(x)\, q(x)^{\gamma-1}\, \mathrm{d}\mu(x)}{\|q\|_\gamma^{\gamma-1}} \\
&= \|q\|_\gamma^{1-\gamma} \left[\left\{\int p(x)^\gamma \, \mathrm{d}\mu(x)\right\}^{\frac{1}{\gamma}} \left\{\int q(x)^\gamma \, \mathrm{d}\mu(x)\right\}^{1-\frac{1}{\gamma}}\right. \\
&\qquad \left. - \int \{p(x)^\gamma\}^{\frac{1}{\gamma}} \{q(x)^\gamma\}^{1-\frac{1}{\gamma}} \, \mathrm{d}\mu(x)\right] \\
g(\theta) &= (\gamma - 1)\, \|p_\theta\|_\gamma^{1-2\gamma} \left[\int p_\theta(x)^{\gamma-2}\, \dot{p}_\theta(x)^2 \, \mathrm{d}\mu(x) \int p_\theta(x)^\gamma \, \mathrm{d}\mu(x)\right. \\
&\qquad \left. - \left\{\int p_\theta^{\gamma-1}\, \dot{p}_\theta(x)\, \mathrm{d}\mu(x)\right\}^2\right].
\end{aligned}$$

The ∇^*-convex combination satisfies:

$$\frac{p_\omega(x)^{\gamma-1}}{\|p_\omega\|_\gamma} = (1-\omega)\, \frac{p_0(x)^{\gamma-1}}{\|p_0\|_\gamma} + \omega\, \frac{p_1(x)^{\gamma-1}}{\|p_1\|_\gamma} - c_\omega. \tag{41}$$

For the spherical case $\gamma = 2$ these reduce to

$$\begin{aligned}
H(P) &= -\left\{\int p_\theta(x)^2 \, \mathrm{d}\mu(x)\right\}^{\frac{1}{2}} \\
d(P, Q) &= \frac{1}{\left\{\int q(x)^2 \, \mathrm{d}\mu(x)\right\}^{\frac{1}{2}}} \left[\left\{\int p(x)^2 \, \mathrm{d}\mu(x)\right\}^{\frac{1}{2}} \left\{\int q(x)^2 \, \mathrm{d}\mu(x)\right\}^{\frac{1}{2}}\right. \\
&\qquad \left. - \int p(x)\, q(x)\, \mathrm{d}\mu(x)\right]
\end{aligned}$$

$$g(\theta) = \|p_\theta\|^{-3} \left[\int p_\theta(x)^2 \, \mathrm{d}\mu(x) \int \dot{p}_\theta(x)^2 \, \mathrm{d}\mu(x) - \left\{ \int p_\theta(x) \dot{p}_\theta(x) \, \mathrm{d}\mu(x) \right\}^2 \right]$$

$$p_\omega(x) = 1 + k_\omega \left\{ (1-\omega) \frac{p_0(x) - 1}{\|p_0\|_2} + \omega \frac{p_1(x) - 1}{\|p_1\|_2} \right\},$$

with k_ω chosen so that $\|p_\omega\|_2 = k_\omega$.

6.2 Ranked probability score

Consider the (non-strict) proper scoring rule:

$$S(x, Q) = \{Q(A) - 1_A(x)\}^2 \tag{42}$$

where $A \subseteq \mathcal{X}$ and 1_A is the indicator function of A. This is equivalent to the Brier score for predicting whether or not $X \in A$. We have

$$H(P) = P(A)\{1 - P(A)\} \tag{43}$$

$$d(P, Q) = \{P(A) - Q(A)\}^2. \tag{44}$$

The associated (semi)-metric gives the divergence between P_θ and $P_{\theta+\mathrm{d}\theta}$ as $\frac{1}{2}g(\theta)\,\mathrm{d}\theta^2$ with

$$g(\theta) = 2\{\dot{P}_\theta(A)\}^2.$$

If now we have a collection of such events, $\{A_t\}$, indexed by $t \in \mathcal{T}$, we can mix the above quantities using a measure μ over $\mathcal{T}$, thus obtaining:

$$S(x, Q) = \int \{Q(A_t) - 1_{A_t}(x)\}^2 \, \mathrm{d}\mu(t) \tag{45}$$

$$H(P) = \int P(A_t)\{1 - P(A_t)\} \, \mathrm{d}\mu(t) \tag{46}$$

$$d(P, Q) = \int \{P(A_t) - Q(A_t)\}^2 \, \mathrm{d}\mu(t) \tag{47}$$

$$g(\theta) = 2 \int \{\dot{P}_\theta(A_t)\}^2 \, \mathrm{d}\mu(t). \tag{48}$$

In the case $\mathcal{T} = \mathcal{X}$ (discrete) and $A_t = \{t\}$, this becomes equivalent to the Brier problem of Sect. 5.3.2. If instead we take $\mathcal{T} = \mathcal{X} = \mathbb{R}$ and $A_t = (-\infty, t]$, so that $Q(A_t) = F_Q(t)$, the cumulative distribution function of Q evaluated at t, we obtain the *ranked probability score* (Epstein, 1969; Gneiting and Raftery, 2005). In either case the scoring rule is strictly proper so long as μ has full support, and the decision connexion is the mixture connexion.

6.3 Kernel scores

A further generalisation of the Brier score is as follows (Eaton, 1982; Eaton et al., 1996; Dawid, 1998; Gneiting and Raftery, 2005). Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ be a positive-definite kernel, i.e. $K(x,y) = \overline{K}(y,x)$, and $\|m\|_K^2 := K(m,m) := \int\int K(x,y)\,\mathrm{d}m(x)\,\mathrm{d}m(y) > 0$ for all non-null bounded signed measures m on $\mathcal{X}$. Consider

$$S_0(x,Q) = \|Q - \delta_x\|_K^2, \tag{49}$$

where δ_x denotes the point mass at x. This is a strictly proper scoring rule, strongly equivalent to

$$\|Q\|_K^2 - K(x,Q) - K(Q,x).$$

We have:

$$\begin{aligned} H(P) &= -\|P\|_K^2 + \mathrm{E}_P\|\delta_X\|_K^2 \\ d(P,Q) &= \|P-Q\|_K^2 \\ g(\theta) &= 2\|\dot{P}_\theta\|_K^2. \end{aligned}$$

The decision connexion is in all cases identical with the mixture connexion.

We recover the construction of Sect. 6.2 for $K(x,y) = \int 1_{A_t}(x)\,1_{A_t}(y)\,\mathrm{d}\mu(t) = \mu\{t : A_t \supseteq \{x,y\}\}$. For the case $\mathcal{X} = \mathbb{R}$ another possible kernel is $K(x,y) = \int \mathrm{e}^{\mathrm{i}t(x-y)}\,\mathrm{d}\mu(t)$, where μ is a full-support measure on $\mathbb{R}$. Then $\|m\|_K^2 = \int |\phi_m(t)|^2\,\mathrm{d}\mu(t)$, where $\phi_m(t) \equiv \int \mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}m(x)$ denotes the characteristic function of m. We then have:

$$\begin{aligned} S(x,Q) &= \int \left|\mathrm{e}^{\mathrm{i}tx} - \phi_Q(t)\right|^2\,\mathrm{d}\mu(t) \\ H(P) &= \int \left\{1 - |\phi_P(t)|^2\right\}\,\mathrm{d}\mu(t) \\ d(P,Q) &= \int \left|\phi_P(t) - \phi_Q(t)\right|^2\,\mathrm{d}\mu(t) \\ g(\theta) &= 2\int |\dot{\phi}_\theta(t)|_K^2\,\mathrm{d}\mu(t). \end{aligned}$$

Gneiting and Raftery (2005) give a number of other examples of scoring rules based on (real-valued) kernels.

Acknowledgments This work owes much to stimulating interactions with Steffen Lauritzen. Some of the material was presented in Dawid and Lauritzen (2006).

References

Amari, S. (2005). Integration of stochastic evidences in population coding – theory of α-mixture. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, (pp. 15–21). University of Tokyo, 12–16 December 2005.

Amari, S., Nagaoka, H. (1982). Differential geometry of smooth families of probability distributions. Technical Report METR 82-7, Department of Mathematical Engineering and Instrumentation Physics, University of Tokyo.

Amari, S., Nagaoka, H. (2000). *Methods of Information Geometry.* Translations of Mathematical Monographs, Vol. 191. Providence, Rhode Island: American Mathematical Society and Oxford University Press.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, *7*, 686–690.

Bregman, L. M. (1967). 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.

Cover, T., Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley Interscience.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics*, *19*, 2032–2066.

Dawid, A.P. (1975). Discussion of Efron (1975). *Annals of Statistics*, *3*, 1231–1234.

Dawid, A.P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Department of Statistical Science, University College London. http://www.ucl.ac.uk/Stats/research/abs94.html#139.

Dawid, A. P., Lauritzen, S.L. (2006). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, (pp.22–28). University of Tokyo, 12–16 December 2005.

Dawid, A. P., Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, *27*, 65–81.

Eaton, M.L. (1982). A method for evaluating improper prior distributions. In: S. Gupta, J.O. Berger,(Eds.) *Statistical Decision Theory and Related Topics III*, (pp. 320–352.)New York: Academic Press.

Eaton, M. L., Giovagnoli, A., Sebastiani, P. (1996). A predictive approach to the Bayesian design problem with application to normal regression models. *Biometrika*, *83*, 11–25.

Efron, B.(1975). Defining the curvature of a statistical problem (with applications to second-order efficiency) (with Discussion). *Annals of Statistics*, *3*, 1189–1242.

Eguchi, S. (2005). Information geometry and statistical pattern recognition. *Sugaku Exposition*, American Mathematical Society(to appear).

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*, 985–987.

Gneiting, T., Raftery, A. E. (2005). Strictly proper scoring rules, prediction, and estimation. Technical Report 463R, Department of Statistics, University of Washington.

Good, I.J. (1971). Comment on "Measuring information and uncertainty" by Robert J. Buehler. In V.P. Godambe,D.A. Sprott(Eds.) *Foundations of Statistical Inference*, (pp. 337–339)Toronto: Holt, Rinehart and Winston.

Grünwald, P. D., Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, *32*, 1367–1433.

Lauritzen, S. L. (1987a). Conjugate connections in statistical theory. In C. T. J. Dobson(Ed.) *Geometrization of Statistical Theory: Proceedings of the GST Workshop*, (pp. 33–51)Lancaster: ULDM Publications, Department of Mathematics, University of Lancaster.

Lauritzen, S. L. (1987b). Statistical manifolds. In *Differential Geometry in Statistical Inference*, IMS Monographs, (Vol. X, pp. 165–216) Hayward, California: Institute of Mathematical Statistics.

Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S. (2004). Information geometry of U-boost and Bregman divergence. *Neural Computing*, *16*, 1437–1481.

Tsallis, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, *52*, 479–487.