2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo
Pages 000–000

THE GEOMETRY OF DECISION THEORY

A. PHILIP DAWID AND STEFFEN L. LAURITZEN

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. We illustrate these for the case of a Riemannian outcome space.

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.

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

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 [5]. 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 in this way. In [5, 6] they are illustrated for a number of important cases, and used to determine the optimal choice of an experimental design.

2. Minimum divergence estimation

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

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

Research Report No. 261, Department of Statistical Science, University College London
Date: December 2005.

When d derives from a decision problem as above, this is equivalent to minimising the total *empirical score*:

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

— in which form it remains meaningful even when $\widehat{P}_n \notin \mathcal{P}$.

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

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

Theorem 2.1. *The estimating equation* (3) *is unbiased.*

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

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

□

This result generalises readily to multi-dimensional parameter spaces.

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.

2.1. **Quasi-likelihood** [13, 11, § 9.2.1]. Suppose $\mathcal{X}$ and $\mathcal{A}$ are both Euclidean space $\mathbb{R}^n$. Let h be a differentiable strictly concave function on $\mathcal{X}$, and consider the decision problem with loss function

$$L(x, a) = h(a) + \frac{\partial h(a)}{\partial a^j}(x^j - a^j)$$

(where we use Einstein's summation convention for repeated indices).

Then

$$L(P, a) = h(a) + \frac{\partial h(a)}{\partial a^j}(\mu_P^j - a^j),$$

where μ_P is the mean vector of P. This is minimised for $a = \mu_P$, so that the corresponding proper scoring rule is

$$S(x, Q) = h(\mu_Q) + \frac{\partial h(\mu_Q)}{\partial \mu_Q^j}(x^j - \mu_Q^j), \tag{4}$$

the entropy function is $H(P) = h(\mu_P)$, and the divergence is

$$d(P, Q) = h(\mu_Q) - h(\mu_P) + \frac{\partial h(\mu_Q)}{\partial \mu_Q^j}(\mu_P^j - \mu_Q^j).$$

Because of their dependence only on mean vectors we will also write $S(x, \mu)$ and $d(\nu, \mu)$ $(\nu, \mu \in \mathcal{X})$.

It is readily checked that

$$\frac{\partial S(x, \mu)}{\partial \mu^i} = -k_{ij}(\mu)(x^j - \mu^j), \tag{5}$$

where

$$k_{ij}(\mu) := -\frac{\partial^2 h(\mu)}{\partial \mu^i \partial \mu^j} \tag{6}$$

(so that $K(\mu) := (k_{ij}(\mu))$ is a positive definite matrix).

Property (5) establishes $-S(x,\mu)$ as a *quasi-likelihood* function for the model in which the dispersion matrix is specified as the function $V(\mu) = K(\mu)^{-1}$ of the mean μ. Conversely, if (5) holds then $K(\mu)$ must have the form of (6) for some concave function h, so every quasi-likelihood can be obtained from a decision problem in this way. In particular, quasi-likelihood estimation under a constraint $\mu \in E$ can be conducted by minimising the empirical score $S(x,\mu)$ — or, equivalently, the divergence $d(x,\mu)$ — subject to the constraint. It follows from (5) that this will always yield an unbiased estimating equation, in illustration of Theorem 2.1.

3. Estimating a density on a Riemannian space

Hyvärinen [10] has proposed a method for estimating a density over a Euclidean space that can be performed without knowledge of the normalisation constant. Here we generalise this to to the case of an outcome space $\mathcal{X}$ that is a Riemannian manifold, with metric defined by an inner product $\langle\,\cdot\,,\,\cdot\,\rangle$ on the tangent space at each point $x \in \mathcal{X}$ (we do not mention the base point x explicitly). We write $\|v\|^2$ for $\langle v, v\rangle$. Let μ denote the corresponding volume measure on $\mathcal{X}$.

As action space $\mathcal{A}$, we take the set of scalar fields f such that $\|\nabla f\|^2 \to 0$ as x approaches the boundary of $\mathcal{X}$. Here ∇f denotes the *natural gradient* of f, which is a vector field.

Our loss function is given by:

$$L(x,\kappa) := \kappa^{-1}\Delta\kappa \text{ (evaluated at } x). \tag{7}$$

Here Δ denotes the *Laplace-Beltrami operator*, with coordinate expression, for a scalar field f:

$$\Delta f := \partial_i \partial^i f + \frac{1}{2}(\partial^i f)\, \partial_i(\log\det g),$$

where g is the metric tensor, and $\partial^i f := g^{ij}\partial_j f$ is the coordinate expression of ∇f. Then Δf is a scalar field.

For $\mathcal{P}$, we take the the set of probability distributions P over $\mathcal{X}$ that are absolutely continuous with respect to μ, and such that the *natural density* $p := dP/d\mu$ (a scalar field) satisfies $\log p \in \mathcal{A}$. We assume existence of all the required expectations below.

The expected loss function is

$$\begin{aligned} L(P,\kappa) &= \int (\Delta\kappa)(p/\kappa)\, d\mu \\ &= -\int \langle \nabla\kappa, \nabla(p/\kappa)\rangle\, d\mu, \end{aligned}$$

on applying Stokes's theorem.

Introducing $\pi := p^{\frac{1}{2}}$, we find:

$$L(P,\pi) = -\int \|\nabla\pi\|^2\, d\mu$$

and

$$L(P,\kappa) - L(P,\pi) = \int \kappa^2\, \|\nabla(\pi/\kappa)\|^2\, d\mu$$

which is always non-negative, and vanishes if and only if $\kappa \propto \pi$.

We thus have the following expressions:

Bayes act: $\kappa_P := p^{\frac{1}{2}}$

Entropy: $H(P) = -\int \|\nabla p^{\frac{1}{2}}\|^2\, d\mu$

Scoring rule: $S(x,Q) = q^{-\frac{1}{2}}\Delta q^{\frac{1}{2}}$ (evaluated at x)

Expected score: $S(P,Q) = -\int \langle \nabla q^{\frac{1}{2}}, \nabla(p/q^{\frac{1}{2}})\rangle\, d\mu$

Divergence: $d(P,Q) = \int q \, \|\nabla(p/q)^{\frac{1}{2}}\|^2 \, d\mu$

We remark that $L(x,\kappa)$, $S(x,Q)$, $d(P,Q)$ can be calculated even if we only know κ or q up to a scale factor. In particular, use of the estimating equation (3) does not require knowledge of the normalising constant for distributions in $\mathcal{Q}$, which is often hard to obtain. We further remark that, since $d(P,Q)$ and $H(P)$ are defined only for distributions $P \in \mathcal{P}$, whereas $\widehat{P}_n \notin \mathcal{P}$, the expression of the estimation rule in terms of minimum discrepancy, as in (1), is no longer meaningful; however, there is no such difficulty in minimising the empirical score, as in (2).

We can re-express the above quantities in terms of $l_P := \log p$ and l_Q. We obtain:

$$\begin{aligned}
S(x,Q) &= \frac{1}{2}\Delta l_Q + \frac{1}{4}\|\nabla l_Q\|^2 \\
S(P,Q) &= \frac{1}{4}E_P\langle \nabla l_Q - 2\nabla l_P, \nabla l_Q\rangle \\
H(P) &= -\frac{1}{4}E_P\|\nabla l_P\|^2 \\
d(P,Q) &= \frac{1}{4}E_P\|\nabla l_P - \nabla l_Q\|^2
\end{aligned}$$

We do not need to know μ to calculate the divergence: $d(P,Q)$ is unchanged if we interpret l_P and l_Q as log-densities with respect to any fixed underlying measure.

When $\mathcal{X}$ is Euclidean, $2\,d(P,Q)$ becomes the criterion proposed in [10, equation (2)]; while $2\,S(x,Q)$ is the expression whose expectation (first theoretical, then empirical) appears in [10, equations (3) and (4)].

We remark that the above proper scoring rule $S(x,Q)$ is determined by the values of l_Q in an arbitrarily small neighborhood of x. Contrast this with the result of [3] that the only proper scoring rule depending only on the value of l_Q at x is essentially identical with $-l_Q$.

4. Decision geometry: General framework

We now return to the general decision problem of § 1, and introduce a concrete framework within which we can naturally define and manipulate geometric properties associated with the problem. The theory outlined below can be made rigorous for the case of a finite outcome space $\mathcal{X}$, and is 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 functions on $\mathcal{X}$. These spaces are in duality with respect to the bilinear product $\langle m\,,\,f\rangle = \int f(x)\,dm(x)$. 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, and thereby inherits its natural parallel displacement ∇. At any $P \in \mathcal{P}$, the tangent space to $\mathcal{P}$ is naturally represented as the subspace $W^+ := \{m \in W : m(\mathcal{X}) = 0\}$, which identification defines an affine connexion, also denoted by ∇, on $\mathcal{P}$.

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^+, and by ∇^* the natural parallel displacement on V^+.

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 $\{v \in V : v \geq a \text{ for some } a \in \mathcal{L}\}$, 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 $p \in \mathcal{L}^*$.[1]
(ii) Distinct distributions in $\mathcal{P}$ have distinct Bayes acts in $\mathcal{L}^*$.[2]
(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 p is then a $(1, 1)$ correspondence. The supporting hyperplane T_P now becomes the tangent plane to $\mathcal{L}$ at p, intersecting $\mathcal{L}$ at the single point p.

We note the following identifications:

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

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

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

For $\Pi \in \mathcal{P}$, let $k(\Pi) := \alpha\, S(\Pi, P) + (1 - \alpha)\; S(\Pi, Q) - H(\Pi) = \alpha\, d(\Pi, P) + (1 - \alpha)\, d(\Pi, Q)$. 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 $v := \alpha\, p + (1-\alpha)\, q - k$.

For any $\Pi \in \mathcal{P}$, $\langle \Pi, v\rangle = \alpha\,\langle \Pi, p\rangle + (1-\alpha)\,\langle \Pi, q\rangle - k = k(\Pi) + H(\Pi) - k \geq H(\Pi)$, whence $v \in \mathcal{L}$. Moreover $\langle R, v\rangle = H(R) = \inf_{a\in\mathcal{L}}\langle R, a\rangle$. Thus $v = r$, 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 $p = \lambda(P)$ 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, in addition to the parallel displacement ∇ inherited by $\mathcal{P}$ directly as a convex subset of W, it also inherits a parallel displacement ∇^* through its correspondence with the convex subset $\mathcal{L}^+$ of V^+.

Differential geometry. The tangent space to the manifold $\mathcal{P}$ at any point P is naturally represented by W^+, and that to $\mathcal{L}^+$ at any point p^+ by V^+. Under our basic assumptions, the function λ^+ is differentiable at $P \in \mathcal{P}^\circ$ (the interior of $\mathcal{P}$), its derivative thus supplying an isomorphism between W^+ and V^+. Through this, ∇^* is converted into an affine connexion (also denoted by ∇^*) on $\mathcal{P}^\circ$; and the natural bilinear product is converted into an inner product on W^+, so defining a metric g on $\mathcal{P}^\circ$. These constructions and properties, which are special cases of the general theory of [13], make $(\mathcal{P}^\circ, g, \nabla, \nabla^*)$ a dually flat statistical manifold [1, 14, 2]. We remark that ∇ is always the mixture connexion, whereas ∇^* and the metric will depend on the specific decision problem. For the special case that the loss is defined by the logarithmic score, $-l_Q(x)$, we recover the information geometry, which was introduced in terms similar to the above in [4]. Eguchi [8] has studied the decision

[1] We use corresponding upper case and lower case symbols for a distribution in $\mathcal{P}$ and its Bayes act in $\mathcal{L}$.

[2] This is equivalent to the scoring rule S being *strictly* proper.

geometry associated with a separable Bregman score [12, 9].

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

4.1. **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). As a convex subset of V^+, $\mathcal{F}$ has a natural affine parametrisation and parallel displacement ∇^*, which thus transfer to $\mathcal{E}$. A 1-dimensional LGEF is a ∇^*-geodesic.

Since $q(\,\cdot\,) \equiv S(\,\cdot\,, Q)$, 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* property [9, § 7.2]:

$$S(x, Q_\beta) \equiv \beta_0 + m(x) + \sum_{i=1}^{k} \beta_i \, t_i(x), \tag{8}$$

for some m, $t_i \in V$, with β_0 then a uniquely determined function of β. Applying Theorem 2.1 we find $d\beta_0/d\beta_i = -\mathrm{E}_{Q_\beta}\{t_i(X)\}$ $(\beta \in \mathcal{B}^\circ)$.

Let $t := (t_1, \dots, t_k)$, and define, for $\tau \in \mathbb{R}^k$: $\Gamma_\tau := \{P \in \mathcal{P} : \mathrm{E}_P\{t(X)\} = \tau\}$. Suppose[3] that there exists $P_\tau \in \Gamma_\tau \cap \mathcal{E}$. Since $S(P, Q) = \langle P\,,\, q\rangle$, an easy calculation, using (8), yields:

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

This in turn implies the "Pythagorean equality":

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

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

$$P_\tau = \arg\min_{Q \in \mathcal{E}} d(P, Q). \tag{10}$$

When P is the empirical distribution $\widehat{P}_n$ of data $(x_1, \dots, x_n)$ 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 $m = q$, and, for given $t_i \in V$, let $\mathcal{E}$ be given by (8): then $\mathcal{E}$ is a LGEF containing Q. Again, if there exists $P_\tau \in \Gamma_\tau \cap \mathcal{E}$ then (9) holds for all $P \in \Gamma_\tau$, whence we readily deduce

$$P_\tau = \arg\min_{P \in \Gamma_\tau} d(P, Q). \tag{11}$$

What happens when $\Gamma_\tau \neq \emptyset$ but $\Gamma_\tau \cap \mathcal{E} = \emptyset$? In this case P_τ can still be defined by (11), but will not now be in $\mathcal{E}$ (P_τ will in fact lie on the boundary of $\mathcal{P}$). The family $\mathcal{E}^m \supseteq \mathcal{E}$ of all P_τ given by (11) 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^+ [9, Example 7.1]. Nonetheless, under mild conditions we can apply minimax theory to a suitable game between Nature and Decision Maker, constructed from the decision problem, to derive the *Pythagorean inequality* (a strengthening of (9) and so *a fortiori* of (11)):

$$d(P, P_\tau) + d(P_\tau, Q) \le d(P, Q) \qquad (P \in \Gamma_\tau). \tag{12}$$

[3] This need not hold in general: see below.

One might conjecture that (10) also continues to hold in the form $P_\tau = \arg\min_{Q \in \mathcal{E}^m} d(P, Q)$ for $P \in \Gamma_\tau$, but this need not be so [9, § 7.6.1].

Grünwald and Dawid [9] investigate further game-theoretic aspects of statistical decision problems related to convex duality and the existence of saddle-points, including but extending beyond properties of generalised exponential families. It is likely that many of these can be given interesting geometric interpretations within the framework set out above. However to incorporate the full generality of the game-theoretic approach within the geometry it would be important to find ways of relaxing our basic assumptions (i) and (ii).

References

[1] Amari, S. and 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.

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

[3] Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, **7**, 686–90.

[4] Dawid, A. P. (1975). Discussion of [7]. *Annals of Statistics*, **3**, 1231–4.

[5] 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`.

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

[7] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second-order efficiency) (with Discussion). *Annals of Statistics*, **3**, 1189–242.

[8] Eguchi, S. (2005). Information geometry and statistical pattern recognition. To appear in *Sugaku Exposition*, Amer. Math. Soc.

[9] Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, **32**, 1367–433.

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

[11] Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference.* John Wiley and Sons.

[12] Lafferty, J. (1999). Additive models, boosting and inference for generalized divergences. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT '99)*, pp. 125–33. University of California at Santa Cruz.

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

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

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
E-mail address: `dawid@stats.ucl.ac.uk`
URL: `http://tinyurl.com/2675a`

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
E-mail address: `steffen@stats.ox.ac.uk`
URL: `http://www.stats.ox.ac.uk/~steffen/`