The Annals of Statistics
1994, Vol. 22, No. 3, 1587–1602

PREFERRED POINT GEOMETRY AND THE LOCAL DIFFERENTIAL GEOMETRY OF THE KULLBACK–LEIBLER DIVERGENCE[1]

BY FRANK CRITCHLEY, PAUL MARRIOTT AND MARK SALMON

University of Birmingham, University of Surrey and European University Institute

A new preferred point geometric structure for statistical analysis, closely related to Amari's α-geometries, is introduced. The added preferred point structure is seen to resolve the problem that divergence measures do not obey the intuitively natural axioms for a distance function as commonly used in geometry. Using this tool, two key results of Amari which connect geodesics and divergence functions are developed. The embedding properties of the Kullback–Leibler divergence are considered and a strong curvature condition is produced under which it agrees with a statistically natural (squared) preferred point geodesic distance. When this condition fails the choice of divergence may be crucial. Further, Amari's Pythagorean result is shown to generalise in the preferred point context.

1. Introduction. Ideas of distance in geometry have mostly been developments of the Euclidean axiom that the shortest path between two points is a straight line. The distance between these points is then defined as the length of this line. Following the developments which enable us to define what is meant by a straight line in spaces more complex than Euclid's plane, we find that we pass through most of the history of geometry itself. The following are the familiar axioms for $\delta(m,n)$, the distance from m to n:

1. *positivity*, $\delta(m,n) \geq 0$ with equality if and only if $m = n$;
2. *symmetry*, $\delta(m,n) = \delta(n,m)$;
3. *the triangle inequality* $\delta(m,n) \leq \delta(m,p) + \delta(p,n)$.

These reflect the intuitive idea of distances being minimum pathlengths. Condition 1 is self-evident. Condition 2 follows from the intuitive idea that the length of a path is independent of the direction travelled. Condition 3 derives from the idea that if we take the shortest path from m to p followed by the shortest path from p to n, then we have taken a path from m to n and since pathlengths are assumed additive we have gone at least as far as the shortest path joining them.

There has also been a natural interest in statistics in how to measure the separation of two density functions [see, e.g., Rao (1945, 1987), Burbea and Rao (1982), Jeffreys (1948), Bhattacharyya (1943) and Kullback and Leibler (1951)].

Received October 1991; revised November 1993.

[1]Supported by ESRC Grant R000232270, "Geodesic inference, encompassing and preferred point geometries in econometrics."

AMS 1991 *subject classifications.* Primary 53B99; secondary 62F05, 62F12.

Key words and phrases. Amari α-geometry, differential geometry, distance, divergence, geodesic, Kullback-Leibler divergence, preferred point, Pythagoras' theorem, statistical manifold.

The fundamental role played by Fisher's information matrix in Rao's notion of distance may be contrasted with the Kullback–Leibler divergence function

$$d_{kl}(\theta, \theta') = E_{p(x,\theta)}\big[\ln p(x,\theta) - \ln p(x,\theta')\big].$$

This function and many other proposed divergence or discrimination functions are apparently quite different from the more geometric ideas of distance. For example, they do not necessarily satisfy conditions 2 and 3 above. These functions do, however, reflect the asymmetry which is fundamental to statistical inference given the singling out of some particular distribution as representing, for example, either the true data generation process or the maintained hypothesis.

The past 15 years have seen substantial developments in the relationship between differential geometry and statistics. See, for example, the review papers by Barndorff–Nielsen, Cox and Reid (1986) and by Kass (1989). In particular we note Amari's (1990) construction of an expected geometry on a parametric family of density functions. Using this geometry, Amari was able to forge links between the differential geometric concept of a geodesic and some common divergence functions, including the Kullback–Leibler measure. There has also been work on the "Euclidean" geometry of the Kullback–Leibler divergence [see, e.g., Čencov (1972), Csiszar (1975) and Loh (1983)]. Barndorff–Nielsen (1989) and Blaesild (1990, 1991) use the concept of a yoke, one of which is minus the Kullback–Leibler measure, to generate very general geometric structures, and these have strong links with the contrast functions of Eguchi (1983, 1984). Marriott (1989) defined and introduced a new differential geometric construction called a preferred point geometry. This has been further developed in Critchley, Marriott and Salmon (1993), where it is shown how preferred point geometry relates to Amari's expected α-geometries. As Amari, Kurata and Nagaoka (1990) affirm, the projection theorem and the generalised Pythagorean theorem are the highlights of the theory of dually flat manifolds, such as the expected α-geometries.

In this paper we show how preferred point geometry gives rise to an asymmetric geometric structure which is particularly relevant to statistics. In particular, we develop both key theorems of Amari mentioned above. The added structure of a preferred point geometry gives a clearer picture of the parallel between metric-based geometric distances, such as proposed by Rao, and statistically based divergence functions. Further, a number of model selection procedures such as Akaike's information criterion [Akaike (1973)] are based on the Kullback–Leibler notion of distance and hence our development clarifies, to some extent, how the use of these discrimination measures may be related to formal statistical hypothesis tests and decision theory.

In Section 2 we introduce the necessary geometric and statistical background, including a statement (Theorem 1) of the two key results which we develop. In Section 3 we introduce a new (preferred point) geometric structure and prove (Theorem 2) a general equivalence theorem between divergences and preferred point geodesic distances. For the rest of the paper we concentrate on the widely used Kullback–Leibler divergence and a particular, statistically

natural, preferred point geometry. Theorem 3 characterises a strong flatness condition under which the two measures are compatible, thereby yielding a stronger form of Amari's projection theorem. Theorem 4 gives a partial classification of the parametric families which fulfil this condition. Section 4 uses the preferred point geometry to generalise the Pythagorean result of Theorem 1(ii). It would be of interest to perform a similar study of other divergences [see, e.g., Rao (1987)].

2. Geodesics and divergences. Throughout $\{p(x,\theta)\}$ will denote a p-dimensional parametric family (or manifold) of probability density functions obeying the regularity conditions of Amari [(1990), page 16]. Further, we will denote the coordinate system by $\theta = (\theta^1, \ldots, \theta^p)$.

2.1. *Riemannian geometry.*

DEFINITION. A *Riemannian manifold* $(M, g(\theta))$ is a manifold M and a *metric tensor* $g(\theta)$. See Amari [(1990), page 26].

EXAMPLE 1 [Rao (1945)]. Let $M = \{p(x,\theta)\}$ be a parametric family of density functions and $g(\theta)$ the Fisher information matrix, that is,

$$\big(g(\theta)\big)_{ij} = g_{ij}(\theta) = \mathbf{E}_{p(x,\theta)}\left[\frac{\partial}{\partial\theta^i}\ln p(x,\theta)\frac{\partial}{\partial\theta^j}\ln p(x,\theta)\right].$$

Then $(M, g(\theta))$ is Riemannian manifold [see Amari (1990), page 27].

In a Riemannian manifold we define a *path* γ to be a smooth map,

$$\gamma(t)\colon [a,b] \subset \mathbf{R} \to M,$$
$$t \to \big(\gamma^1(t), \ldots, \gamma^p(t)\big),$$

and its *length* from $t = a$ to b by

$$L(a,b) = \int_a^b \sqrt{\sum_{i,j} g_{ij}\big(\gamma(t)\big)\frac{d}{dt}\gamma^i(t)\frac{d}{dt}\gamma^j(t)}\,dt.$$

This length will be invariant to a change of parametrisation on M due to the tensorial nature of the metric.

DEFINITION. A *geodesic* between two points in a Riemannian manifold is defined to be a curve joining them of minimum pathlength, and this length is called the *geodesic distance* between them.

In this paper we work purely locally, which removes any complications for this simple definition. For a detailed discussion on the general existence and uniqueness problem, see Postnikov [(1967), page 95]. Assuming (as we always

shall) the existence and uniqueness of the geodesic, then the geodesic distance obeys conditions 1, 2 and 3 above.

It will become important to know when a Riemannian manifold has no curvature.

DEFINITION. A Riemannian manifold $(M, g(\theta))$ is defined to have *zero curvature* or to be *flat* if there exists a coordinate system θ in which $g(\theta)$ is independent of θ, that is, a constant G, say, [see Dodson and Poston (1977), page 418]. These coordinates are then called *affine*.

Under these circumstances the geodesic from θ_1 to θ_2 is given by $\gamma(t) = (1 - t)\theta_1 + t\theta_2$ and the squared geodesic distance by $(\theta_1 - \theta_2)^t G(\theta_1 - \theta_2)$, an exact quadratic function of $(\theta_1 - \theta_2)$. For details see Dodson and Poston [(1977), page 374].

2.2. *Divergence functions.* We make the following definition.

DEFINITION. A *divergence* function $d(\theta_1, \theta_2)$ is a smooth function on pairs of points in a parametric family $M = \{p(x, \theta)\}$ for which the following hold:

(i) $d(\theta_1, \theta_2) \geq 0$, with equality when and only when $\theta_1 = \theta_2$;
(ii) $\partial_i d(\theta_1, \theta_2)|_{\theta_1 = \theta_2} = \partial_i' d(\theta_1, \theta_2)|_{\theta_1 = \theta_2} = 0$, where $\partial_i = \partial/\partial\theta_1^i$; and $\partial_i' = \partial/\partial\theta_2^i$;
(iii) $\partial_i \partial_j' d(\theta_1, \theta_2)|_{\theta_1 = \theta_2} = g_{ij}(\theta_1)$, the Fisher information matrix.

This definition is closely related to that of a *yoke* [see Barndorff–Nielsen (1989)], except we insist that the Hessian of the divergence is the Fisher information, while the more general yoke allows any nonsingular Hessian.

Some well-known examples of divergence functions include the Kullback–Leibler, the Hellinger and the Renyi α-information [see Amari (1990), page 88]. In particular we denote the Kullback–Leibler divergence by $d_{kl}(\theta_1, \theta_2)$.

2.3. *Amari's geometry.* Amari's geometric structure can be seen as a development of Rao's Riemannian geometry in Example 1. Amari's geometric structure on $M = \{p(x, \theta)\}$ is formally defined by a pair of tensors, $g_{ij}(\theta)$ and $T_{ijk}(\theta)$, where g is the Fisher information and T is defined by

$$T_{ijk}(\theta) = \mathbf{E}_{p(x,\theta)}\left[\frac{\partial}{\partial\theta^i}\ln p(x,\theta)\frac{\partial}{\partial\theta^j}\ln p(x,\theta)\frac{\partial}{\partial\theta^k}\ln p(x,\theta)\right].$$

The new geometric concept used in Amari's geometry is that of a *connection*. For a good basic introduction to the concept of a connection and the related Christoffel symbols, see Dodson and Poston [(1977), Chapter 8]. In fact Amari uses a whole family of connections parametrised by the real number α. Each one is then called an α-connection.

DEFINITION. The α-connections for Amari's geometry are defined by their *Christoffel symbols*

$$\Gamma^{\alpha}_{ijk} = \Gamma^{0}_{ijk} - \frac{\alpha}{2} T_{ijk},$$

where Γ^{0}_{ijk} is derived from the metric tensor g_{ij} by

$$\Gamma^{0}_{ijk} = \frac{1}{2}\left[\frac{\partial g_{jk}}{\partial \theta^{i}} + \frac{\partial g_{ik}}{\partial \theta^{j}} - \frac{\partial g_{ij}}{\partial \theta^{k}}\right],$$

the Christoffel symbol for the *Levi–Civita* or *metric* connection.

Similarly to the Riemannian case, each α-connection defines a set of *geodesics*. In general, a geodesic is a solution of a set of differential equations determined by a connection [see Dodson and Poston (1977), page 347]. Intuitively geodesics are the "straight lines" of the geometry. The Reimannian geodesics are produced in this way from the 0-connection. However, *only* the geodesics of the 0-connection have the property of length minimisation. The other connections do not have an associated notion of geodesic distance as they can, indeed, be defined independently of the metric.

For each α-connection there is a definition of (local) flatness which generalises the Riemannian one.

DEFINITION. The manifold M is α-flat if there exists a coordinate system θ such that, for all θ_1 and θ_2, the α-geodesic joining them is of the form $\gamma(t) = (1-t)\theta_1 + t\theta_2$.

As an example Amari proves that any full exponential family is both +1-flat and, as he shows, consequently -1-flat.

2.4. *Geodesics and divergence functions.* Using this more general definition of geodesic, Amari was able to derive a relationship between the Kullback–Leibler divergence and the -1-geodesic in the full exponential family case. In fact his result was more general relating α-divergences to α-geodesics for an α-family; however, for brevity and simplicity we concentrate on the $\alpha = -1$ case.

THEOREM 1.

(i) [*Amari* (1990), *page* 90]. *If M is a full exponential family and N a submanifold of M, then, for any point θ in M, the point θ' in N which minimises $d_{kl}(\theta, \theta')$ is joined to θ via a -1-geodesic which cuts N orthogonally in the Fisher metric at θ'.*

(ii) [*Amari* (1990), *page* 86]. *Consider three points, θ_1, θ_2 and θ_3 in a full exponential family. Let c be the -1-geodesic connecting θ_1 and θ_2 and c' the +1-geodesic joining θ_2 and θ_3. If the angle between c and c' at θ_2 is $\pi/2$, measured in the Fisher metric, then*

$$d_{kl}(\theta_1, \theta_3) = d_{kl}(\theta_1, \theta_2) + d_{kl}(\theta_2, \theta_3).$$

There is a strong analogy between the first part of this theorem and the result in Riemannian geometry which states that the point on a submanifold of (M,g) which is closest to a fixed point in M is found by dropping a geodesic which cuts the submanifold orthogonally [see Postnikov (1967), page 108], while the second part is of course a direct generalisation of Pythagoras' theorem. Thus there is a parallel between squared geodesic distances and divergences. However, it is important to notice that since the α-geodesics are nonmetric, whenever $\alpha \neq 0$, there is no concept of α-geodesic distance involved in the above results. Thus there is no connection between the divergence function and a squared geodesic distance. To get this we must use preferred point geometry, as we demonstrate.

3. Preferred point geometry. In this section we investigate the relationship between divergence functions and geodesics using *preferred point geometry*. We first look at the general question of symmetry for distance functions, and then, concentrating on the Kullback–Leibler divergence, extend Theorem 1(i).

DEFINITION. Let M be a finite-dimensional manifold with a parametrisation θ. A *preferred point tensor* $T^{\phi}(\theta)$ is a smooth function of $(\phi, \theta) \in M \times M$ such that, for each given ϕ, $T^{\phi}(\theta)$ is a tensor on M. We call ϕ the *preferred point*.

Preferred point tensors are particularly suited to the analysis of parametric families of distributions where the preferred point ϕ could be the true, or the hypothesised, distribution. This can be considered fixed, whereas θ denotes a general member of the family. The geometry of this family, defined by the tensor structure, will then be conditional on which distribution is chosen as the preferred one.

DEFINITION. Let M be a manifold with a parametrisation θ. A preferred point tensor $g_{ij}^{\phi}(\theta)$ is a *preferred point metric* if, for each ϕ and all θ in an open neighbourhood of ϕ, $g_{ij}^{\phi}(\theta)$ is metric tensor as a function of θ.

If the preferred point ϕ is fixed, $(M, g^{\phi}(\theta))$ is (locally to ϕ) a Riemannian manifold. We can therefore define the *preferred point distance* between ϕ and θ to be the geodesic distance between them for the g^{ϕ} metric. As before this will be well defined for all points θ in an open neighbourhood of ϕ. We denote the squared preferred point distance from ϕ to θ in the g^{ϕ}-metric by $D(\phi, \theta)$.

Squared preferred point distances share the same basic characteristics as divergence functions. They do not have to be symmetric. [Consider $D(\phi, \theta)$ and $D(\theta, \phi)$. The first is the squared geodesic distance from ϕ to θ as defined by the metric g^{ϕ}, whereas the second is defined in the g^{θ} geometry. In general geodesic paths and pathlengths will be different due to the different metrics.] Further, the triangle inequality does not have to hold for preferred point distances. Also they obey conditions (i) and (ii) of the definition in Section 2.2 and they can be shown to be yokes. In fact we now show that any divergence function can be interpreted locally as a squared preferred point geodesic distance. This is

not, however, a one-to-one correspondence since there are a great number of compatible preferred point metrics for each divergence function.

THEOREM 2. *Let M be a p-dimensional manifold with a parametrisation θ. Let $d(\ ,\ )$ be a divergence function. There exists a preferred point metric $g^{\phi}(\theta)$ such that if $D(\phi, \theta)$ is the squared g^{ϕ}-geodesic distance from ϕ to θ, then $D(\phi, \theta) = d(\phi, \theta)$ for all θ in some open neighbourhood of ϕ.*

The proof is given in the Appendix.

This existence result holds for any divergence function. However, for the rest of this paper we concentrate on the relationship between the widely used Kullback–Leibler divergence and a particular set of preferred point tensors which have a direct statistical interpretation.

There is a basic property of the Kullback–Leibler divergence which is not reflected in the axioms of Section 2. These define the properties of a divergence on a (finite-dimensional) parametric family, but the Kullback–Leibler divergence is in fact well defined on a much larger family.

DEFINITION. Let $S = \{p(x)\}$ be the set of all mutually absolutely continuous regular density functions on a sample space X with respect to a dominating measure P.

On S it is well known [see Loh (1983) or Kullback (1968)] that the Kullback–Leibler divergence is well defined, nonnegative and, for all $p, q \in S$,

$$d_{kl}(p, q) = 0 \quad \text{iff } p = q \text{ a.e.}$$

We now compare the Kullback–Leibler divergence with a preferred point geometry. Consider the triple $(M, \mu^{\phi}(\theta), g^{\phi}(\theta))$, where μ^{ϕ} and g^{ϕ} are a preferred point 1-tensor and metric, respectively, given by

$$\mu_i^{\phi}(\theta) = \mathbf{E}_{p(x,\,\phi)}\left[\frac{\partial}{\partial\theta^i}\ln p(x, \theta)\right],$$

$$g_{ij}^{\phi}(\theta) = \mathbf{Cov}_{p(x,\,\phi)}\left[\frac{\partial}{\partial\theta^i}\ln p(x,\ \theta), \frac{\partial}{\partial\theta^j}\ln p(x, \theta)\right].$$

In Critchley, Marriott and Salmon (1993) these tensors are shown to relate closely to Amari's geometric structure. Statistically they are nothing more than the mean and the covariance of the score at θ taking expectations with respect to $p(x, \phi)$.

The parametric family M is embedded in the infinite-dimensional space S. This simple observation highlights a clear distinction between the Kullback–Leibler divergence and the geodesic distance induced by the preferred point metric g^{ϕ} defined above. The geodesic distance between two densities is a function of the particular, finite-dimensional, family in which they are considered to lie. However, the Kullback–Leibler divergence is purely a function of the two

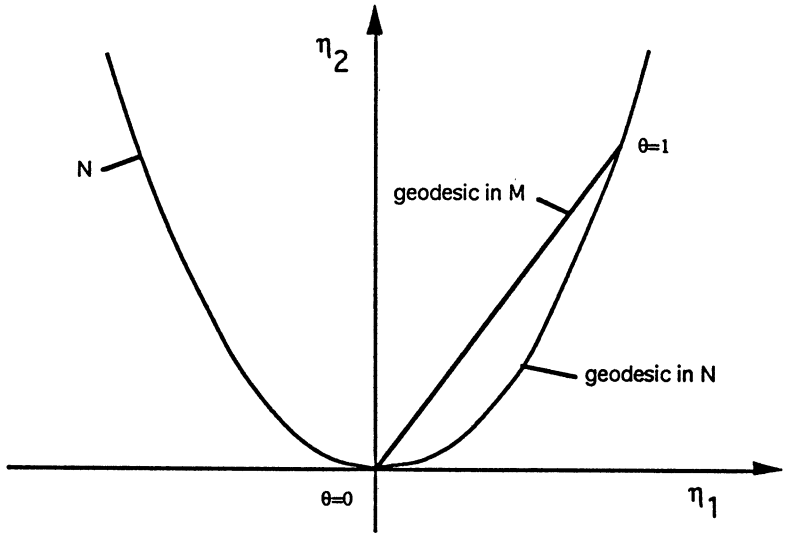

FIG. 1.

densities and is independent of the choice of family. This distinction is demonstrated in the following example.

EXAMPLE 2. Consider the curved parametric family defined by Efron (1975). Let M be the two-dimensional parametric family of bivariate normal distributions with covariance matrix I, the identity, and parametrised by their mean values (η_1, η_2). Let N be the one-dimensional subfamily parametrised by θ such that the mean vector of a point of N is given by $(\theta, (\gamma_0/2)\theta^2)$. See Figure 1. Let the point $\theta = 0$ be the preferred point. In M the preferred point metric g^ϕ and the Fisher information are easily shown to be I. Both metrics are a constant in the parametrisation (η_1, η_2). The geodesic in M between $\theta = 0$ and 1 is a straight line. On the other hand the geodesic in N between the same two points is the relevant chord of the curve N itself. The geodesic distance in N reduces to the relevant arc length of N inside M. This is clearly a different distance. In contrast the Kullback–Leibler divergence between the two points is independent of the manifold in which they are considered to lie. In fact an easy calculation shows that it agrees with half the squared geodesic distance as measured in M. This result will be generalised in Theorem 3.

For this example there is the intuitive remark that the distance measured in M will equal that measured in N if and only if N is straight (has no embedding curvature) in M.

Consider now a general parametric family M embedded in the function space S. Due to its dependence on the choice of manifold it is clear that in general the g^ϕ-geodesic distance will not equal the Kullback–Leibler divergence on M. It is the aim of this section to produce a flatness condition analogous to that in the example above under which the two measures will agree.

Consider Theorem 1(i) in the light of this intuition. In S we have the concept of minimising the Kullback–Leibler divergence, while in M we have that of projecting along geodesics. We can now interpret Theorem 1(i) as stating that the two concepts coincide because of a particular flatness (or zero curvature) property of the full exponential family.

In Critchley, Marriott and Salmon (1993) it is shown that for a full exponential family the preferred point g^ϕ-geodesics are the same as Amari's +1-geodesics. We describe a strong flatness condition, related to +1-flatness, under which the Kullback–Leibler divergence will equal half the squared preferred point geodesic distance. This exactly mirrors the intuitive observation for our finite-dimensional Euclidean example. Further, minimising geodesic distances is well known to be equivalent to geodesic projection [see Postnikov (1967)]. Hence the strengthened flatness condition produces a strengthened version of Amari's projection theorem.

Recall that a Riemannian manifold (M,g) is called *flat* if there exists a coordinate system θ such that the metric $g(\theta)$ is constant for all θ. We now generalise this condition for the preferred point geometry $(M,\mu^\phi(\theta),g^\phi(\theta))$.

DEFINITION. For each ϕ, the preferred point geometry $(M,\mu^\phi(\theta),g^\phi(\theta))$ is g^ϕ-flat if there exists a coordinate system for which $g^\phi(\theta)$ is constant for all θ. The θ-coordinates are called g^ϕ-affine. Further, M is *totally flat* if there exists a coordinate system θ for which $g^\phi(\theta)$ is a constant and also $\mu^\phi(\theta)$ is a linear function of $\theta - \phi$.

EXAMPLE 3. Consider the family of p-variate nonsingular normal distributions with constant covariance matrix Σ which are parametrised by their mean values $\eta = (\eta^1,\ldots,\eta^p)$. Apart from a constant, the log-likelihood is given by

$$-\tfrac{1}{2}(x-\eta)^t\Sigma^{-1}(x-\eta).$$

Simple calculations then show

$$\mu^\phi(\eta) = -\Sigma^{-1}(\eta-\phi) \quad \text{and} \quad g^\phi(\eta) = \Sigma^{-1}.$$

Thus the space is totally flat.

This flatness condition gives a simple correspondence between the Kullback–Leibler divergence and the g^ϕ-geodesic distance.

THEOREM 3. *Let the preferred point geometry $(M,\mu^\phi(\theta), g^\phi(\theta))$ be g^ϕ-flat, and let the θ-coordinates be g^ϕ-affine. Then the following three statements are equivalent:*

(i) *The manifold M is totally flat (locally to ϕ).*

(ii) *The Kullback–Leibler divergence $d_{kl}(\phi,\theta)$ (locally to ϕ) equals half the squared g^ϕ-geodesic distance.*

(iii) *The Kullback–Leibler divergence is (locally to ϕ) an exact quadratic function of the θ-coordinates given by*

$$d_{kl}(\phi, \theta) = \tfrac{1}{2}(\theta - \phi)^t g(\phi)(\theta - \phi),$$

where g is the Fisher information metric.

The proof is given in the Appendix.

It is natural to ask if a more direct characterisation of total flatness can be found. Below we give a solution to this problem for full exponential families.

Let P_θ be a p-dimensional full exponential family with canonical parameter θ and with representation

$$\frac{dP_\theta}{dP}(x) = B(x)\exp\left[\sum_{i=1}^{p}\left(\theta^i t_i(x)\right) - \psi(\theta)\right],$$

relative to some dominating measure P.

THEOREM 4. *Consider the above full exponential family P_θ. The following three statements are equivalent*:

(i) *The family P_θ is totally flat.*

(ii) *The covariance of the canonical statistic $t(x)$ does not depend on the canonical parameter θ.*

(iii) *The log-likelihood is a quadratic function of the canonical parameter.*

The proof is given in the Appendix.

This last characterisation is of particular interest. Taking P to be Lebesgue measure $\lambda(x)$, then the totally flat exponential families are those whose canonical statistic has a normal distribution $N_p(A\theta + b, A)$ for some A and b independent of the canonical parameter θ. Further denoting the density of $t(x)$ by $p(t)$, then the totally flat full exponential families themselves have densities

$$\widetilde{B}(x)p\big(t(x)\big),$$

where $\widetilde{B}(x)$ has the same support as $B(x)$. In the particular case $\widetilde{B}(x) = 1$ and $t(x) = x$ we recover Example 3, the p-variate nonsingular normals with constant covariance matrix parametrised by their mean values.

An important corollary is that when a manifold is not totally flat minimising the Kullback–Leibler divergence will not in general be equivalent to minimising the g^ϕ- (or, equivalently, the Fisher-) geodesic distance. The choice between these measures will then matter. Clearly the former enjoys certain robustness properties being independent of the parametric family considered, while the latter might be expected to be more efficient when the data generation process lies in the manifold.

4. The preferred point Pythagoras theorem. We turn now to the generalised version of Pythagoras' theorem, Theorem 1(ii), proved in Amari (1990) and Čencov (1972) for full exponential families.

We have already calculated the preferred point structure for such a family. For any full exponential family, $p(x, \theta) = \exp[\sum_{i=1}^{p} \theta^i t_i(x) - \psi(\theta)]$ in +1-affine, or canonical, coordinates θ,

$$g_{ij}^{\phi}(\theta) = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\phi) = g_{ij}(\phi),$$

where g is the Fisher metric. Thus $g^{\phi}(\theta)$ is independent of θ. Hence the +1-affine coordinates are also g^{ϕ}-affine. Further,

$$\mu_i^{\phi}(\theta) = \frac{\partial \psi}{\partial \theta^i}(\phi) - \frac{\partial \psi}{\partial \theta^i}(\theta) = \eta^i(\phi) - \eta^i(\theta),$$

where η are the -1-affine, or expected, coordinates. Thus there is clearly a strong relationship with the ± 1-dual structure of Amari.

We have defined preferred point distances to be measured from the preferred point; however, the Pythagorean result involves the distances $d_{kl}(\theta_1, \theta_2), d_{kl}(\theta_1, \theta_3)$ and $d_{kl}(\theta_2, \theta_3)$. Thus two different preferred points are used. Hence, to understand the preferred point geometry of such a triple, it is necessary to see how the preferred point tensors vary with alternative preferred points. For the case μ^{ϕ} we consider the tensor

$$\Delta(\theta) = \Delta^{(\theta_1, \theta_2)}(\theta) = \mu^{\theta_1}(\theta) - \mu^{\theta_2}(\theta).$$

Before we can use this tensor we need the following definition.

DEFINITION. If Δ is defined as above we define a *null path*, γ: $[0, 1] \to M$, of Δ to be such that

$$\sum_{i=1}^{p} \Delta_i \frac{d\gamma^i}{dt} = 0.$$

By standard results on the existence and uniqueness of solutions to differential equations a null path will pass through θ if $\Delta(\theta)$ is a smooth nonsingular tensor field in a neighbourhood of θ.

Using the concept of a null path we can define a Pythagorean result for our preferred point geometry.

THEOREM 5. *Let θ_1, θ_2 and θ_3 be three points in a finite-dimensional parametric family M. If $\Delta^{(\theta_1, \theta_2)}$ is nonsingular at θ_2, there is a Pythagorean relationship for the Kullback–Leibler divergence*:

$$d_{kl}(\theta_1, \theta_2) + d_{kl}(\theta_2, \theta_3) = d_{kl}(\theta_1, \theta_3),$$

if θ_3 lies on a null path of $\Delta^{(\theta_1, \theta_2)}$ through θ_2.

The proof is given in the Appendix.

For the full exponential family,

$$\Delta^{(\theta_1,\theta_2)}(\theta) = \eta(\theta_1) - \eta(\theta_2),$$

which is a constant, independent of θ and a linear function of the -1-affine coordinates of θ_1 and θ_2. Then for a full exponential family the null paths can be easily calculated.

LEMMA 1. *A null path of $\Delta^{(\theta_1,\theta_2)}$ through θ_2 is a $+1$-geodesic through θ_2 which is Fisher orthogonal to the -1-geodesic joining θ_1 and θ_2.*

Combining Theorem 5 and Lemma 1 we get the following corollary, which is equivalent to Theorem 1(ii).

COROLLARY. *In the case of the full exponential family,*

$$d_{kl}(\theta_1,\theta_2) + d_{kl}(\theta_2,\theta_3) = d_{kl}(\theta_1,\theta_3),$$

if θ_3 lies on a $+1$-geodesic through θ_2 which is Fisher orthogonal to the -1-geodesic joining θ_1 and θ_2.

APPENDIX

PROOF OF THEOREM 2. Let us fix ϕ and treat the divergence function as a function of θ. By the positivity of the Hessian of the divergence function at ϕ we can apply Morse's lemma [see Poston and Stewart (1976), page 15]. We choose coordinates $\psi(\theta) = (\psi^1(\theta), \ldots, \psi^p(\theta))$ such that, locally,

$$d(\phi,\theta) = \tfrac{1}{2}\sum_{i,j} g_{ij}(\phi)\psi^i(\theta)\psi^j(\theta),$$

where $g_{ij}(\phi)$ is the Fisher information at ϕ.

Thus we can use this coordinate change to define a map from Θ to $\mathbf{R}^p$ by

$$\Psi\colon \theta \to \psi(\theta).$$

Let us put the constant metric $\frac{1}{2}g(\phi) = G$, say (remember ϕ is fixed), on $\mathbf{R}^p$ in its standard coordinates. We now define a metric on M by pulling back this metric on $\mathbf{R}^p$ via the map Ψ; that is to say we define g^ϕ by

$$g^\phi(v_1,v_2) = \sum_{i,j} g^\phi_{ij} v_1^i v_2^j = \sum_{i,j} G_{ij}(\Psi^* v_1)^i(\Psi^* v_2)^j,$$

where $\Psi^*(\theta)\colon TM_\theta \to \mathbf{R}^p$ is the lift by Ψ to the relevant tangent spaces. By construction, the squared geodesic distance in (M, g^ϕ) from ϕ to θ will equal

that in $(\mathbf{R}^p, G)$ from 0 to $\psi(\theta)$. Since $\mathbf{R}^p$ has a constant metric this squared distance will be

$$\sum_{i,j} G_{ij}\psi^i(\theta)\psi^j(\theta),$$

which equals the divergence. □

PROOF OF THEOREM 3. First we see that:

$$\begin{aligned}
\text{statement (i)} &\Leftrightarrow \mu^\phi(\theta) \text{ is linear in } (\theta - \phi) \\
&\Leftrightarrow \frac{\partial}{\partial\theta} d_{kl}(\phi, \theta) \text{ is linear in } (\theta - \phi) \\
&\Leftrightarrow d_{kl}(\phi, \theta) \text{ is quadratic in } (\theta - \phi) \\
&\Leftrightarrow d_{kl}(\phi, \theta) = \sum_{ij} \frac{1}{2} A_{ij}(\theta - \phi)^i(\theta - \phi)^j + \sum_i B_i(\theta - \phi)^i + C
\end{aligned}$$

where A, B and C are independent of θ.

From axioms (i)–(iii) of a divergence we have $A = g$ and $B = C = 0$. Hence in the theorem we have statement (i) $\Leftrightarrow$ statement (iii).

Since M is g^ϕ-flat, and using the result on Riemannian distances in Section 2.1, we immediately have statement (ii) $\Leftrightarrow$ statement (iii) in the theorem. □

PROOF OF THEOREM 4. From Theorem 3 we have that statement (i) $\Rightarrow$ $\mu_i^\phi(\theta) = \sum_j - g_{ij}(\phi)(\theta - \phi)^j$. Thus for any totally flat family we have

$$\frac{\partial \mu_i^\phi}{\partial \theta^j}(\theta) = -g_{ij}(\phi),$$

where $g(\phi)$ is the Fisher information at ϕ.

In the full exponential family example a direct calculation shows that

$$\frac{\partial \mu_i^\phi}{\partial \theta^j}(\theta) = -\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\theta).$$

However, for such a family $g_{ij}(\theta) = -[\partial^2\psi/\partial\theta^i\partial\theta^j](\theta)$. Thus for all θ we have

$$-\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\theta) = -\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\phi).$$

Thus the second derivative of ψ, which equals the covariance of the sufficient statistic for a full exponential family, is a constant. Thus in the theorem we have statement (i) $\Rightarrow$ statement (ii).

To complete the proof, it immediately follows that if

$$-\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\theta) = A_{ij} = \text{a constant},$$

then the log-likelihood is

$$\sum_i t_i(x)\theta^i + \tfrac{1}{2}\sum_{i,j} A_{ij}\theta^i\theta^j + \sum_i B_i\theta^i + C,$$

where A, B and C are independent of θ. This directly implies that μ^ϕ will then be linear. Thus statement (ii) $\Rightarrow$ statement (iii) $\Rightarrow$ statement (i). □

PROOF OF THEOREM 5. If θ_3 equals θ_2, then we clearly have the result. In general, write the above equation in the form

$$d_{kl}(\theta_1, \theta_2) = d_{kl}(\theta_1, \theta_3) - d_{kl}(\theta_2, \theta_3).$$

The left-hand side of this equation is constant with respect to θ_3. If θ_3 moves along a path $\gamma(t)$, then the rate of change of the right-hand side with respect to t is

$$\begin{aligned}\frac{d}{dt}\big(d_{kl}(\theta_1, \theta_3) - d_{kl}(\theta_2, \theta_3)\big) &= \sum_i \frac{d\gamma^i}{dt}\frac{\partial}{\partial\theta_3^i}\big(d_{kl}(\theta_1, \theta_3) - d_{kl}(\theta_2, \theta_3)\big) \\ &= -\sum_i \frac{d\gamma^i}{dt}\Delta_i^{(\theta_1, \theta_2)}\end{aligned}$$

Thus since the required equation holds at $\theta_3 = \theta_2$ it will hold as θ_3 moves along the null path of $\Delta^{(\theta_1, \theta_2)}$. □

PROOF OF LEMMA 1. Now Δ is constant in θ-coordinates. Hence the null line is a θ-affine line or a +1-geodesic through θ_2. Its direction is defined by its tangent vector at θ_2, which satisfies

$$\sum_{i=1}^{p} \Delta_i \frac{d\gamma^i}{dt} = 0. \tag{1}$$

The tangent to the -1-geodesic joining θ_1 and θ_2 is most easily calculated by using the change of basis matrix, which is the inverse Fisher information $g^{ij}(\theta_2)$ [see Amari (1990), page 80], on the tangent vector in η-coordinates. This will be $\eta(\theta_2) - \eta(\theta_1)$. Thus in θ-coordinates it will have the tangent vector

$$v_i = \sum_j g^{ij}\big(\eta(\theta_2) - \eta(\theta_1)\big)_j = \sum_j g^{ij}\Delta_j.$$

Combining with (1) shows that the tangent vector of the null line is Fisher orthogonal to the -1-geodesic at θ_2. □

Acknowledgment. We wish to thank the referees and Editor for several helpful ideas which greatly improved the presentation of this paper.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B. N. Petrov and F. Czaki, eds.) 267–281. Akademiai Kiado, Budapest.

AMARI, S.-I. (1990). *Differential-Geometric Methods in Statistics*, 2nd ed. *Lecture Notes in Statist.* **28**. Springer, Berlin.

AMARI, S.-I., KURATA, K. and NAGAOKA, H. (1990). Differential geometry of Boltzmann machines. Technical Report, METR 90-19. Dept. Mathematical Engineering, Univ. Tokyo.

BARNDORFF-NIELSEN, O. E. (1989). *Parametric Statistical Models and Likelihood. Lecture Notes in Statist.* **50**. Springer, Berlin.

BARNDORFF-NIELSEN, O. E., COX, D. R. and REID, N. (1986). The role of differential geometry in statistical theory. *Internat. Statist. Rev.* **54** 83–96.

BHATTACHARRYA, A. (1943). On discrimination and divergence. In *Proceedings of the 29th Indian Scientific Congress Part 3* 13.

BLAESILD, P. (1990). Yokes: orthogonal and extended normal coordinates. Research Report 205, Aarhus Univ.

BLAESILD, P. (1991). Yokes and tensors derived from yokes. *Ann. Inst. Statist. Math.* **43** 95–113.

BURBEA, J. and RAO, C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *J. Multivariate Anal.* **12** 575–596.

ČENCOV, N. N. (1972). *Statistical Decision Rules and Optimal Inference.* Nauka, Moscow. (Translated into English 1982.)

CRITCHLEY, F., MARRIOTT, P. K. and SALMON, M. (1993). Preferred point geometry and statistical manifolds. *Ann. Statist.* **21** 1197–1224.

CSISZAR, I. (1975). I-divergence geometry of probability distributions and minimisation problems. *Ann. Probab.* **3** 146–158.

DODSON, C. T. J. and POSTON, T. (1977). *Tensor Geometry*. Pitman, London.

EFRON, B. (1975). Defining the curvature of a statistical problem (with discussion). *Ann. Statist.* **3** 1189–1242.

EGUCHI, S. (1983). Second order efficiency of minimum contrast estimators. *Ann. Statist.* **11** 793–803.

EGUCHI, S. (1984). A characterisation of second order efficiency in a curved exponential family. *Ann. Inst. Statist. Math.* **36** 199–206.

JEFFREYS, H. (1948). *Theory of Probability*, 2nd ed. Clarendon, Oxford.

KASS, R. E. (1989). The geometry of asymptotic inference (with discussion). *Statist. Sci.* **4** 188–234.

KULLBACK, S. L. (1968). *Information Theory and Statistics*. Peter Smith, Magnolia.

KULLBACK, S. L. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86.

LOH, W.-Y. (1983). A note on the geometry of the Kullback–Leibler information numbers. Technical Report 716, Dept. Statistics, Univ. Wisconsin.

MARRIOTT, P. K. (1989). Applications of differential geometry to statistics. Ph.D. dissertation, Dept. Mathematics, Univ. Warwick.

POSTNIKOV, M. M. (1967). *The Variational Theory of Geodesics*. Saunders, London.

POSTON, T. and STEWART, I. N. (1976). *Taylor Expansions and Catastrophes: Research Notes in Mathematics*. Pitman, London.

RAO, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91.

RAO, C. R. (1987). Differential metrics in probability spaces. In *Differential Geometry in Statistical Inference* 217–240. IMS, Hayward, CA.

FRANK CRITCHLEY
SCHOOL OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF BIRMINGHAM
EDGBASTON
BIRMINGHAM B15 2TT
UNITED KINGDOM

PAUL MARRIOTT
DEPARTMENT OF MATHEMATICAL
AND COMPUTING SCIENCES
UNIVERSITY OF SURREY
GUILDFORD
SURREY GU2 5XH
UNITED KINGDOM

MARK SALMON
DEPARTMENT OF ECONOMICS
EUROPEAN UNIVERSITY
INSTITUTE
BADIA FIESOLANA
I-50016 SAN DOMENICO DI FIESOLE (FI)
ITALY