ELSEVIER

Journal of Statistical Planning and Inference 102 (2002) 229–245

journal of statistical planning and inference

www.elsevier.com/locate/jspi

On preferred point geometry in statistics

Frank Critchley[a, *], Paul Marriott[b], Mark Salmon[c]

[a] *Department of Statistics, The Open University, UK*

[b] *Department of Statistics and Applied Probability, National University of Singapore, Singapore*

[c] *Financial Econometrics Research Group, City University Business School, UK*

Accepted 15 March 2001

Abstract

A brief synopsis of progress in differential geometry in statistics is followed by a note of some points of tension in the developing relationship between these disciplines. The preferred point nature of much of statistics is described and suggests the adoption of a corresponding geometry which reduces these tensions. Applications of preferred point geometry in statistics are then reviewed. These include extensions of statistical manifolds, a statistical interpretation of duality in Amari's expected geometry, and removal of the apparent incompatibility between (Kullback–Leibler) divergence and geodesic distance. Equivalences between a number of new expected preferred point geometries are established and a new characterisation of total flatness shown. A preferred point geometry of influence analysis is briefly indicated. Technical details are kept to a minimum throughout to improve accessibility.

Keywords: Differential geometry; Divergence; Geodesic distance; Influence analysis; Kullback–Leibler divergence; Statistical manifold; Parametric statistical modelling; Preferred point geometry; Rao distance; Riemannian geometry; Yoke geometry

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. This journey takes us via Pythagoras' theorem, Newton's calculus, Gauss's differential geometry and Euler's calculus of variations to Einstein's use of geometry in physics. Throughout this long history runs the central theme that we measure the separation of two points by finding the shortest path between them. In particular, this use of

* Corresponding author.

E-mail address: f.critchley@open.ac.uk (F. Critchley).

0378-3758/02/$ - see front matter
PII: S0378-3758(01)00115-X

minimum path lengths provided the intuitive basis for the now familiar metric axioms:

(M1) *Nonnegativity*: The only paths of zero length are the trivial ones from a point to itself, whence $d(a,b) \geqslant 0$ with equality if and only if $a = b$.

(M2) *Symmetry*: The reverse of every path from a to b is a path from b to a of the same length, whence $d(a,b) = d(b,a)$.

(M3) *The triangle inequality*: Every path from a to b of length $l(a,b)$ followed by a path from b to c of length $l(b,c)$ produces a path from a to c of length $l(a,b)+l(b,c)$, whence $d(a,b) + d(b,c) \geqslant d(a,c)$.

Metrics in the above sense are related to, but distinct from, metrics—more fully, metric tensors—on manifolds. Such tensors define geodesic distances. They lie at the heart of Riemannian geometry and formed the basis for the first work on differential geometry in statistics (Rao, 1945). This field has undergone rapid expansion in recent years and forms the natural backdrop to the present paper.

A little geometric background is given in Section 2, which may be referred back to at any point as required. Here, as throughout the paper, technical details are kept to a minimum so as to improve accessibility and focus on the key ideas involved. In particular, technical results from our earlier papers, which we draw on as appropriate, are cited without proof and woven into an overall discussion.

The plan (and arrangement) of the rest of the paper are as follows. A succinct account of main ideas and key references in differential geometry in statistics is offered (Section 3) and some points of tension between these disciplines noted (Section 4). The preferred point nature of much of statistics is described and suggests the adoption of a corresponding geometry which reduces these tensions (Section 5). Applications of preferred point geometry in statistics are then reviewed (Sections 6 and 7). These geometries bring attractive benefits. For example, the nonmetric connections, so important for statistical calculations, are given a conceptually simple metric connection interpretation. This interpretation also provides more insight into the statistical relevance of the key duality results associated with statistical manifolds. Further, the properties of statistically natural divergence functions, which at first sight appear unappealing geometrically (asymmetry and lack of a triangle inequality), turn out to be entirely natural geometrically in a preferred point geometric context. Equivalences between a number of new expected preferred point geometries are also established and a new characterisation of total flatness shown. Section 8 briefly indicates future research in the direction of establishing a preferred point geometry of influence analysis.

2. A little geometrical background

Manifold: Under standard regularity conditions (Amari, 1990, p. 16), a finite dimensional parametric statistical model $\{p(x,\theta)\colon \theta \in \Theta\}$ has the form of a *manifold*, M say.

Metric tensor and curve (*path*) *length*: Adding a *metric* (*tensor*) to M enables us to define lengths and angles in the tangent space to M at each point θ and hence, by integration, the length of any *curve* (*path*) in the manifold and, directly, the angle between any two intersecting curves. Moreover, all this is done in a way that is automatically invariant to changes of coordinate system (reparameterisations) $\theta \to \theta^*$.

Matrix representation of a metric: A metric (tensor) ρ is specified by the inner product which it places on each tangent space. That is, by associating to each point θ of M a positive definite symmetric *matrix* $\rho(\theta)$ which transforms appropriately (as a covariant 2-tensor) under $\theta \to \theta^*$.

Connection and geodesic (*straight path*): Adding a *connection* to M enables us to define geodesics (straight paths) in a way that is similarly invariant. There is a natural Riemannian or metric connection induced by a metric tensor, whose geodesics are paths of minimum length. There is no notion of minimum path length associated with the geodesics of any other, nonmetric, connection.

Flat manifold and affine coordinate system: A manifold M with a connection is called *flat* if there is a coordinate system (parameterisation) θ on M in which its geodesics are line segments. That is, in which the geodesic joining two points $\theta^{(1)}$ and $\theta^{(2)}$ is just the set of all their convex combinations $(1-\lambda)\theta^{(1)} + \lambda\theta^{(2)}$, $(0 \leqslant \lambda \leqslant 1)$. Such a coordinate system is called *affine* and is unique up to nonsingular affine transformation.

Example: For example, the usual 3-D Euclidean geometry is flat with Cartesian coordinates as affine, whereas cylindrical or spherical coordinates are not affine. Again, the surface of a Euclidean sphere is curved (not flat).

Riemannian geometry: *Riemannian geometries* (M, ρ) are natural generalisations of Euclidean geometry to curved spaces in which the metric tensor ρ determines the whole geometry when, as we assume, the induced metric connection is used.

Flat metric: It can be shown that an affine coordinate system on a flat Riemannian manifold is one in which the metric is constant at all points of M. A metric which admits such a coordinate system is called a *flat* metric.

Example: For example, the usual Euclidean metric is flat since, in Cartesian coordinates, it is represented by the same matrix (the identity) at each point of E^n.

3. An overview of differential geometry in statistics

3.1. A core question

A core question in the geometrisation of statistics is:

Given a manifold M identified with a parametric statistical model $\{p(x,\theta): \theta \in \Theta\}$, and given a particular statistical purpose,

precisely what extra structure is it appropriate to add to M?

It will be helpful to keep this core question in mind throughout the paper.

3.2. Rao's Riemannian geometry based on Fisher information

The origins of recent research at the interface between differential geometry and statistics can be traced back some 55 years. As Amari (1990, p. 3) writes,

"It was Rao (1945), in his early 20s, who first noticed the importance of the differential-geometric approach. He introduced the Riemannian metric in a statistical manifold by using the Fisher information matrix and calculated the geodesics between two distributions for various statistical models."

That is, Rao proposed adding the metric ρ defined by $\rho(\theta) = I(\theta)$, the Fisher information matrix, to produce a *Riemannian geometry* (M, ρ) in which to answer the natural question:

"How far apart are two distributions?"

His answers, *Rao distances*, are the geodesic distances induced by this natural choice of metric. They are appropriate to any two distributions in M.

3.3. Kullback and Leibler's divergence geometry

This natural question was also addressed, at about the same time, by Bhattacharrya (1943, 1946), Jeffreys (1948) and Kullback and Leibler (1951) from a variety of directions. In particular, Kullback and Leibler (1951) placed their divergence on M resulting in the divergence geometry (M, d_{KL}).

3.4. Statistical manifolds

Independent work by Chentsov (1972) (translated into English from the Russian in 1982) and Efron (1975, 1978) extended Riemannian geometries for statistics by introducing a whole family of connections, rather than just the Riemannian or metric connection used earlier. Efron also introduced the central idea of statistical curvature.

Since then there has been an explosion of research activity. In particular, these advances inspired the development of new geometries for statistics, including the minimum contrast geometry of Eguchi (1983), the expected geometry of Amari (1990, 1st Edition, 1985), and the observed geometry of Barndorff-Nielsen (1987a, 1988).

These three geometries are all instances of a single elegant structure. A *statistical manifold* is a triple (M, ρ, T) in which the extra ingredient, the skewness tensor T, transforms appropriately (as a covariant 3-tensor) under reparameterisation. Early mathematical accounts of this unifying structure were provided by Amari (1990, 1st Edition, 1985) and Lauritzen (1987).

There is a one parameter family of connections, called the α-connections ($\alpha \in R$), associated with a statistical manifold. The 0-connection is the metric connection. There is an important formal duality between the $+\alpha$ and $-\alpha$ connections: for every $\alpha \in R$, a statistical manifold is α-flat (i.e. flat with respect to the α-connection) if and only if it is $-\alpha$-flat.

3.5. Important advances and key references

Overall, the study of differential geometry in statistics has led to important advances in a variety of fields, including the development of new geometries for statistics, higher order asymptotic theory, invariant asymptotic expansions and inference in nonlinear regression and curved exponential families.

Some further key references are Amari (1982a, b), Amari et al. (1987), Amari and Kumon (1983, 1988), Atkinson and Mitchell (1981), Barndorff-Nielsen (1983, 1986, 1987b), Barndorff-Nielsen and Cox (1989, 1994), Barndorff-Nielsen et al. (1986), Bates and Watts (1980, 1981), Burbea and Rao (1982a, b), Dawid (1975, 1977), Eguchi (1984, 1991), Kass (1984, 1989, 1990), Kass and Vos (1997), Murray and Rice (1993), Oller and Corcuera (1995), Oller and Cuadras (1985), Pistone and Sempi (1995), Rao (1961, 1962, 1963, 1987), Rao et al. (1982) and Vos (1989, 1991a, 1992).

3.6. A natural and fruitful marriage?

The overall goal of this activity could be said to be the establishing of a natural and fruitful marriage between differential geometry and statistical modelling which appropriately applies and extends the former so as to deepen our understanding and capabilities in the latter.

Some of the attractions are obvious. The two disciplines are compatible in a fundamental sense: in many situations, it is required that statistical inferences do not depend on the way that the statistical model has been parameterised, while one definition of geometry is the study of those things which are invariant under a change of coordinates. The geometric approach is well-suited for use in such inferential situations: coordinates are merely labels for points in the same way that parameters are merely labels for distributions. Again, many statistical procedures have very natural geometric interpretations. Three important examples are regression, dimension reduction of a statistic and minimisation of a statistical objective function under a smooth constraint. Further, with such procedures, the intuition which a picture gives can be an invaluable explanatory tool.

At the same time, there are points of tension in this developing relationship, as we review in Section 3.7. Preferred point geometry (Marriott, 1989; Critchley et al., 1992, 1993, 1994, 2000) is being developed as an attempt to ease these points of tension.

3.7. Yoke geometry

A variety of other approaches have also been developed since the mid-1980s. Prominent among these is that based on the concept of a yoke, first introduced in Barndorff-Nielsen (1987b). *Yoke geometry* shares several attractive features with *preferred point geometry*. In particular, the observed and expected yokes are natural statistically, while their associated geometries also extend statistical manifolds beyond third order.

3.8. Preferred point geometry

Our geometrical approach is distinguished by its emphasis on metric connections and by its preferred point nature. This latter is appropriate since, as we discuss in Section 5, much of statistics itself is preferred point in nature.

One profitable future research direction appears to be the development of points of contact between preferred point geometry and yoke geometry.

4. Points of tension

Not all geometrisations of statistics produce affirmative answers to the following key questions:

(T1) *Geometrically simple*? Notwithstanding their elegant formal properties, the mixing of metric and nonmetric connections can prove to be a large conceptual leap for statistical practitioners. For many people connections corresponding to minimum distance geodesics have an immediate, physically based, intuitive appeal (cf. Section 3.6) and so it can be helpful if the dual connections of Amari and others can be reformulated in such a framework.

(T2) *Statistically natural*? Barndorff-Nielsen et al. (1986) end their review paper on the rôle of differential geometry in statistical theory with the following remark:

"While the introduction of more specifically geometrical notions has considerable potential, it remains a challenging task to introduce such ideas in a way that is statistically wholly natural."

(T3) *Statistically interpretable*? The interesting formal duality structure of statistical manifolds and, in particular, of Amari's expected geometry is not well-understood statistically. By linearity, it will suffice to understand ± 1 duality.

(T4) *Are divergences and geodesic distances at least locally compatible*? The metric properties of Rao's geodesic distance contrast sharply with those of Kullback–Leibler divergence $d_{\mathrm{KL}}(\cdot\,,\,\cdot)$ defined by

$$d_{\mathrm{KL}}(\phi,\theta) := E_{\phi}\{\ln p(x,\phi) - \ln p(x,\theta)\}. \tag{1}$$

This measure, and many other proposed divergence or discrimination measures, appear quite different from the more geometric ideas of distance based on minimum path length. In particular, they do not obey either symmetry (M2) or the triangle inequality (M3). At the same time, they arise naturally in statistics. Can this apparent incompatibility be understood and resolved, at least locally, in a natural way? (After the necessary preliminaries, a formal definition of the local compatibility of a divergence and a preferred point metric is given in Section 7.2 below.)

Preferred point geometry seeks to provide affirmative answers to the above questions. The following section discusses the preferred point nature of much of statistics itself. Accordingly, adopting a corresponding preferred point *geometrisation* of statistics provides an affirmative answer to (T2). The fact that a *simple* (indeed, Riemannian-like)

preferred point geometry of statistics is sufficient for many purposes—in particular, for providing a metric-based understanding of how the nonmetric connections associated with statistical manifolds arise—means that (T1) is also answered in the affirmative. Overall, responding positively to (T1) and (T2) is embodied in the founding principles (P1) and (P2) of the preferred point approach to the geometrisation of statistics, announced at the end of Section 5. Finally, Section 6.1 tackles (T3), while (T4) is addressed next and, again, in Section 7.

5. The preferred point nature of much of statistics

From a preferred point perspective, it is sometimes natural statistically that the symmetry condition (M2) should fail and that the triangle inequality (M3) should be of less than central importance.

Consider first (M2). Asymmetry is natural statistically when we think of the 'preferred' or 'distinct' status given to some particular distribution as representing the true or hypothesised distribution. For example, the power of the Neyman–Pearson test of (any) fixed size between two simple hypotheses changes when the rôles of null and alternative are reversed. Otherwise said, how far apart H_a and H_b appear depends on which of them is regarded as specifying the true distribution.

Consider now the triangle inequality (M3). In many cases the "a" in $d(a,b)$ is a preferred point fixed by external considerations and so we are effectively only concerned with the function $d^a(\cdot)$ of a single argument defined by $d^a(b):=d(a,b)$. Leading instances of this arise as follows:

(PP1) We may take the preferred point $a \equiv \theta_0$ to represent the true or hypothesised distribution, and b a candidate distribution allowed by the model. For example, in a parametric likelihood context, we may assess the separation of $a \equiv \theta_0$ from a general parameter value $b \equiv \theta$ in terms of the single argument Kullback–Leibler divergence function $d^a(\cdot) \equiv d^{\theta_0}_{\mathrm{KL}}(\cdot)$ defined by

$$d^{\theta_0}_{\mathrm{KL}}(\theta):=d_{\mathrm{KL}}(\theta_0,\theta). \tag{2}$$

Or, in the same context and still taking $a \equiv \theta_0$, we may instead take b to be the maximum likelihood estimator $\hat{\theta}$ and work with the deviance function $d^a(\cdot) \equiv (\mathrm{X}^2)^{\theta_0}(\cdot)$ defined by

$$(\mathrm{X}^2)^{\theta_0}(\hat{\theta}):= -2\{\ln p(x,\theta_0) - \ln p(x,\hat{\theta})\} \tag{3}$$

whose observed values are the familiar asymptotic χ^2 test statistics.

(PP2) The preferred point a may represent the data and b any candidate from within a class specified by the model. For example, in linear least-squares regression, the separation (*squared* distance) between the observed vector of responses $a \equiv y$ and any point $b \equiv X\beta$ in the range space of the covariate matrix X is judged by the least squares function $d^a(\cdot) \equiv \mathrm{LS}^y(\cdot)$ defined by

$$\mathrm{LS}^y(X\beta):=||y - X\beta||^2_{\mathrm{E}} \tag{4}$$

where $||\cdot||_E$ denotes the usual Euclidean norm. Again, in parametric likelihood inference, we may take a to be the maximum likelihood estimate $\hat{\theta}$ and assess a general parameter value $b \equiv \theta$ in terms of the log-likelihood ratio function $d^a(\cdot) \equiv \mathrm{LLR}^{\hat{\theta}}(\cdot)$ defined by

$$\mathrm{LLR}^{\hat{\theta}}(\theta) := \{\ln p(x, \hat{\theta}) - \ln p(x, \theta)\}. \tag{5}$$

In all such cases, the triangle inequality (M3) is not of central importance, since it is not then directly relevant to consider all comparisons among general triples of points. Rather, in much of statistics, attention is naturally restricted to triples, written $(a; b_1, b_2)$, which contain the preferred point a and in which interest centres on comparisons, in terms of a suitable function $d^a(\cdot)$, between the preferred point a and each of b_1 and b_2. Direct comparisons between b_1 and b_2 are not of central interest here. Rather, comparing them indirectly via a—specifically, via the values of $d^a(b_1)$ and $d^a(b_2)$—will frequently be of interest.

The principal reason why the apparent incompatibility noted above arises is now clear. In the geometric tradition, all points in a manifold are treated equally. No point is singled out for special treatment, in which case we call the geometry *homogeneous*. Other geometrisations of statistics have followed this homogeneous approach. From some points of view, this is natural statistically. For example, all points θ in the parameter space Θ share the possibility of being the unknown true parameter θ_0 giving rise to the data. From other points of view, it is not. As we have seen, much of statistics has a preferred point nature, with the special point corresponding to the (hypothesised) true value or a (constrained) parameter estimate.

This diagnosis directly suggests a possible cure. In particular, a resolution of the apparent incompatibility noted in (T4). In such cases, why not define a geometry on the whole manifold which reflects the special status of the preferred point?

The definition and use of such a *preferred point geometry* are guided by twin principles:

(P1) Be as natural and simple (parsimonious) as possible from both the statistical and geometric perspectives.

(P2) Where appropriate (as indicated above), reflect in a natural way the special status of the preferred point.

6. Preferred point geometry and statistical manifolds

6.1. Preferred point extensions of statistical manifolds and interpretation of duality in Amari's expected geometry

Following Marriott (1989), Critchley et al. (1993) introduced a preferred point geometry (M, ρ^ϕ). This is a Riemannian-like structure but one in which the metric ρ^ϕ depends smoothly on the preferred point $\phi \in M$. This dependence is natural statistically

in the many cases reviewed in the previous section where we are principally interested in geodesic distances away from the preferred point ϕ.

An essential feature of preferred point geometry is that the distance between two points will typically depend upon which of them is taken as preferred. This happens because, in general, both the geodesic path joining a and b, and its length, will be different in the $\phi = a$ and $\phi = b$ geometries. Thus, the lack of symmetry (M2) and of a triangle inequality (M3) found in core statistical divergence functions, but not in standard geometry, are mirrored naturally in preferred point geometry. This collapses the tension (T4).

Any preferred point geometry has a homogeneous geometry associated with it, obtained by restricting attention to the diagonal where $\phi = \theta$. A symmetry condition characterises when this homogeneous geometry subsumes a statistical manifold. At the same time, it provides a natural higher order extension of such structures. Details are given in Critchley et al. (1993). There are corresponding links with strings (Barndorff-Nielsen and Blaesild, 1987a, b, 1988).

In an expected geometry in which ϕ denotes the true parameter, three statistically natural choices of the preferred point metric ρ^ϕ—to be denoted by g^ϕ, h^ϕ and k^ϕ, respectively—are as follows:

(PPM 1) $g^\phi(\theta) := \mathrm{cov}_\phi(s(x,\theta))$, the ϕ-covariance matrix of the score vector

$$s(x,\theta) := [\partial \ln p(x,\zeta)/\partial\zeta]|_{\zeta=\theta}.$$

(PPM 2) The ϕ-expectation of minus the hessian of the log-likelihood at θ does not transform appropriately under reparameterisation $\theta \to \theta^*$ and so cannot be used as a preferred point metric. However, using the standard differential geometrical way to fix this, the g^ϕ-covariant version of this ϕ-expected negative hessian—denoted $h^\phi(\theta)$—defines our second preferred point metric.

(PPM 3) The usual derivation of the asymptotic distribution of the maximum likelihood estimator as $\sqrt{n}(\hat\theta - \phi) \overset{a}{\sim} \mathrm{N}(0, I(\phi)^{-1})$ is based on applying the central limit theorem to the score vector $s(x, \cdot)$ at the true value ϕ which occurs on the right-hand side of the asymptotic linear relation

$$\sqrt{n} I(\phi)(\hat\theta - \phi) \approx (1/\sqrt{n}) s(x,\phi). \tag{6}$$

Alternatively, we may consider the asymptotic distribution of $(\hat\theta - \theta)$ for *any* value of θ using the score vector $s(x, \cdot)$ at θ. Expanding the score vector in a covariant Taylor expansion about $\hat\theta$ leads to the improved approximation to the asymptotic distribution of $\hat\theta$ *locally to* θ,

$$\sqrt{n}(\hat\theta - \theta) \overset{a}{\sim} \mathrm{N}(\sqrt{n} I(\theta)^{-1} \mu^\phi(\theta), k^\phi(\theta)^{-1}), \tag{7}$$

where

$$\mu^\phi(\theta) := E_\phi(s(x,\theta)) \tag{8}$$

and

$$k^\phi(\theta) := h^\phi(\theta) g^\phi(\theta)^{-1} h^\phi(\theta). \tag{9}$$

We interpret (9) as saying that g^{ϕ} and k^{ϕ} are dual with respect to h^{ϕ}. That is, (9) expresses a certain natural general duality between the score vector and the maximum likelihood estimator in terms of their preferred point metrics $g^{\phi}(\theta)$ and $k^{\phi}(\theta)$ defined as above. In this duality, the hessian of the log-likelihood plays a central (pivotal) rôle via $h^{\phi}(\theta)$. In particular, rearranging (9), we have $g^{\phi}(\theta)=h^{\phi}(\theta)k^{\phi}(\theta)^{-1}h^{\phi}(\theta)$. See Critchley et al. (1993) for details, and for duality theorems for arbitrary preferred point geometries.

When $\phi=\theta$, the matrices associated with each of g^{ϕ}, h^{ϕ} and k^{ϕ} reduce to the Fisher information matrix $I(\theta)$. Their metric connections reduce there, respectively, to Amari's $+1$, 0 and -1 connections. In full exponential families, the g^{ϕ} and k^{ϕ} metric connections agree *everywhere* with Amari's $+1$ and -1 connections respectively. In this way, the ± 1 duality of Amari's geometry, previously rather ineluctable statistically (T3), can now be interpreted as reflecting the above duality between the score vector and the maximum likelihood estimator. Natural extensions to $\pm\alpha$ duality have been studied by Zhu and Wei (1997a, b).

6.2. Full exponential family examples

Consider a full exponential family whose density with respect to some carrier measure can be written

$$p(x,\theta)=\exp(x^{\mathrm{T}}\theta-\psi(\theta)),$$

where θ is the canonical (or natural) parameter. Recall that the expectation parameter $\eta(\theta):=E_{\theta}(x)$ and the Fisher information matrix are given, respectively, by $\eta(\theta)=\psi'(\theta)$ and $I(\theta)=\psi''(\theta)$.

Observe also that, in this full exponential family case, the ϕ-mean score and the expectation parameterisations are affinely related by $\mu^{\phi}(\theta)=\eta(\phi)-\eta(\theta)$.

We also have

(a) In θ-coordinates, $g^{\phi}(\theta)=I(\phi)$, a constant independent of θ.

(b) In η-coordinates, $k^{\phi}(\eta)=I^{-1}(\phi)$, a constant independent of η.

(c) In θ-coordinates, $h^{\phi}(\theta)=I(\theta)$, which varies with θ.

In particular, whatever the preferred point ϕ, the canonical parameterisation is g^{ϕ}-affine while the expectation parameterisation is k^{ϕ}-affine.

Properties (a) and (b) respectively, subsume the celebrated results that:

(a) the full exponential family is $+1$-flat and the canonical θ-coordinates are $+1$-affine.

(b) the full exponential family is -1-flat and the expectation η-coordinates are -1-affine.

6.3. Equivalence of expected preferred point geometries

Critchley et al. (1994) considered the statistically natural preferred point geometry (M,μ^{ϕ},g^{ϕ}) defined by $\mu^{\phi}(\theta)$ and $g^{\phi}(\theta)$, the ϕ-mean and the ϕ-covariance, respectively, of the score vector $s(x,\theta)$. We observe here that this geometry is equivalent to

(contains the same information as) three other statistically natural expected preferred point geometries. Insightful in itself, this result will also be useful later (Section 7.3).

In a mild abuse of notation we write, for example, (M, g^ϕ, h^ϕ) for the pair of Riemannian preferred point geometries $((M, g^\phi), (M, h^\phi))$. Denoting equivalences between geometries by '$\leftrightarrow$', we have:

$$(M, \mu^\phi, g^\phi) \leftrightarrow (M, g^\phi, h^\phi) \leftrightarrow (M, g^\phi, h^\phi, k^\phi) \leftrightarrow (M, h^\phi, k^\phi).$$

The proof is straightforward. For the first equivalence we need to show that, given g^ϕ, $\mu^\phi \leftrightarrow h^\phi$. But $\rightarrow$ follows by g^ϕ-covariant differentiation, while $\leftarrow$ follows by integration and the boundary condition $\mu^\phi(\phi) \equiv 0$. The second and third equivalences are immediate from (9).

7. Divergence functions

7.1. Definition

Following Critchley et al. (1994), we define a divergence function $d(\cdot\,,\,\cdot)$ to be a smooth function on pairs of points in M which satisfies:

(D1) $d(\phi, \theta) \geqslant 0$ with equality if and only if $\phi = \theta$;

(D2) $\partial_i d(\phi, \theta)|_{\phi=\theta} = \partial'_i d(\phi, \theta)|_{\phi=\theta} = 0$ where $\partial_i = \partial/\partial\phi_i$ and $\partial'_i = \partial/\partial\theta_i$;

(D3) $\partial'_i \partial'_j d(\phi, \theta)|_{\phi=\theta} = (I(\phi))_{ij}$.

The divergences of Chentsov (1972) and Amari (1990) are special cases in which the evaluation at $\phi = \theta$ in (D3) is dropped, it being assumed that a parameterisation with this stronger property exists.

These conditions imply that, for θ in a neighbourhood of any given point ϕ, a divergence function behaves quadratically with hessian the Fisher information at ϕ. That is, locally to $\phi = \theta$,

$$d(\phi, \theta) \approx \tfrac{1}{2}(\theta - \phi)^{\mathrm{T}} I(\phi)(\theta - \phi).$$

This definition is close to that of a normed yoke (Barndorff-Nielsen, 1989) where condition (D3) is relaxed to nonsingularity of the hessian.

We observe that, apart from smoothness, condition (D1)—which is exactly the non-negativity metric axiom (M1)—is the only essential condition. Condition (D2) can be achieved with any function satisfying (D1) by squaring it if necessary. Condition (D3) can be achieved by rescaling, provided the hessian is nonsingular. In this sense, divergences can be thought of as regular extensions of metrics in which the symmetry (M2) and triangle inequality (M3) axioms are dropped. The local quadratic nature of divergences means that they are analogues—not of distances—but of (half) *squared* distances. Thus, we would not *expect* the triangle inequality (M3) to hold for them. Rather, under analogues of orthogonality, it is natural to look for Pythagorean relationships between divergences (see Section 7.4).

Well-known examples of divergence functions include the Kullback–Leibler divergence defined above, the Hellinger 'distance' and Renyi α-information (see Amari, 1990, p. 88).

7.2. The local differential geometry of the Kullback–Leibler divergence

Amari's celebrated projection theorem (Amari, 1990, p. 90) states that the point $\tilde{\theta}$ in a submanifold $\tilde{M}$ of a full exponential family M which minimises the Kullback–Leibler divergence from a given point $\theta \in M$ is joined to θ by a -1-geodesic which cuts $\tilde{M}$ orthogonally in Rao's Fisher metric at $\tilde{\theta}$.

It is important to note, however, that there is no concept of geodesic distance involved here, since the connection concerned is nonmetric. Thus, this projection theorem does not establish a relationship between the divergence function and a squared geodesic distance. To get this, we use preferred point geometry.

Critchley et al. (1994) use preferred point geometry ideas to investigate the local differential geometry of divergence functions, focusing especially on the Kullback–Leibler divergence.

A first result is that, given any divergence function, there exists a preferred point metric locally compatible with it. That is, for all points θ in a neighbourhood of the preferred point ϕ, half the squared preferred point metric's geodesic distance from ϕ to θ equals $d(\phi, \theta)$.

However, such a locally compatible metric is far from unique. We concentrate now on local compatibility of the Kullback–Leibler divergence and the statistically natural preferred point metric g^{ϕ} defined above. As we observe here (in Section 7.3), it turns out that h^{ϕ} also plays an important rôle.

Now, $d_{\mathrm{KL}}(\cdot\,,\,\cdot)$ is in fact well-defined on infinite dimensional spaces of densities of the form $N := \{p(x)\}$, the set of all mutually absolutely continuous regular densities on the sample space Ω_x with respect to some fixed measure P. Although we do not develop its implications here, we note in passing that the preferred point can be in N rather than M. This important fact is of interest, for example, in studying mis-specified models.

This simple observation enables us to draw an important basic distinction:

Kullback–Leibler divergences measure separations of points in N;
(preferred point) geodesic distances measure separations of points in M.

The Kullback–Leibler divergence $d_{\mathrm{KL}}(\phi, \theta)$ is purely a function of the distributions labelled by ϕ and θ. It is independent of the particular manifold M considered. In contrast, any geodesic distance between ϕ and θ depends not only on the distributions they label but also on the particular, finite dimensional manifold M in which they are considered to lie.

This dependence on the manifold M means that the preferred point metric g^{ϕ} will not, in general, be locally compatible with d_{KL}. There are two geometries on M which

do not in general agree: its *intrinsic* geometry (M, g^{ϕ}) and what, reflecting $M \subset N$, we might call its *embedding* geometry (M, d_{KL}), (cf. Sections 3.2 and 3.3 above, respectively). Some extra condition will therefore be needed for local compatibility. Intuitively, if M itself is intrinsically "flat" (that is, if M is g^{ϕ}-flat), it will be enough if M then also "sits flat" inside N in some sense. The idea of *total flatness* cashes this intuition.

Again, re-considering Amari's projection theorem in the light of this flatness intuition, we can now interpret it as stating that the concepts of minimising d_{KL} in N and g^{ϕ}-geodesic projection in M coincide because of a particular flatness (zero curvature) property of the full exponential family.

Defining the preferred point geometry $(M, \mu^{\phi}, g^{\phi})$ to be totally flat if there is a single coordinate system θ with the property that, for every preferred point ϕ, simultaneously both $g^{\phi}(\theta)$ is constant as θ varies and $\mu^{\phi}(\theta)$ is a linear function of $(\theta - \phi)$, Critchley et al. (1994) prove the following result.

Let (M, g^{ϕ}) be g^{ϕ}-flat and let θ-coordinates be g^{ϕ}-affine. Then the following three statements are equivalent:

(i) Locally to ϕ, the manifold M is totally flat.

(ii) Locally to ϕ, the Kullback–Leibler divergence $d_{\mathrm{KL}}(\phi, \theta)$ equals half the squared g^{ϕ}-geodesic distance from ϕ to θ.

(iii) Locally to ϕ, the Kullback–Leibler divergence is an exact quadratic function of the θ-coordinates given by $d_{\mathrm{KL}}(\phi, \theta) = \frac{1}{2}(\theta - \phi)^{\mathrm{T}} I(\phi)(\theta - \phi)$.

An important corollary of this result is that, whenever a manifold is *not* totally flat, minimising Kullback–Leibler divergence will *not* in general be equivalent to minimising the g^{ϕ}-geodesic distance. The choice between these measures will then matter. Clearly, the former enjoys certain robustness properties, being independent of the parametric manifold considered, while the latter might be expected to be more efficient when the data generation process does lie in or close to the chosen parametric manifold. *Caveat emptor*! See also the discussion on influence analysis in Section 8.

It is clear that total flatness is a strong condition. It is of interest to enquire which full exponential families are totally flat. It follows from the above that, if M is a full exponential family induced by some fixed measure P, then the following three statements are equivalent:

(i) M is totally flat.

(ii) The covariance of the canonical statistic does not depend upon the canonical parameter.

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

In particular, taking P to be Lebesgue measure, the family of p-variate normal distributions with (any) constant covariance matrix is totally flat.

7.3. The rôle of h^{ϕ} in total flatness

Here, we consider instead the preferred point geometry (M, g^{ϕ}, h^{ϕ}) and say that it is totally flat if there exists a single coordinate system with the property of being,

for every preferred point ϕ, simultaneously affine for both g^ϕ and h^ϕ. We call such a coordinate system *co-affine*.

The two preferred point geometries (M, μ^ϕ, g^ϕ) and (M, g^ϕ, h^ϕ) are equivalent (Section 6.3). Again, these two definitions of total flatness are also equivalent. The former implies the latter by covariant g^ϕ-differentiation. The reverse implication follows by integration and the boundary condition (D2).

The latter definition has the advantage that, in view of the duality relation (9), it follows straightforwardly that total flatness is logically equivalent to the existence of a coordinate system that is simultaneously affine for all three preferred point metrics g^ϕ, h^ϕ and k^ϕ (or, again, just for h^ϕ and k^ϕ). Thus, in co-affine coordinates θ in a totally flat manifold, all three preferred point geometries (M, ρ^ϕ) reduce to the Euclidean geometry determined by the Fisher information matrix evaluated at the preferred point ϕ. In particular, their geodesics are line segments (convex combinations) in θ-coordinates and the g^ϕ, h^ϕ and k^ϕ squared geodesics from ϕ to θ are *all* simply $(\theta-\phi)^{\mathrm{T}} I(\phi)(\theta-\phi)$. This observation gives, at once, obvious alternative forms of the above result characterising total flatness.

Again, defining total flatness in terms of h^ϕ enables us to quantify the extent to which minimisation of d_{KL} and of g^ϕ differ. Let M be g^ϕ-flat and let θ-coordinates be g^ϕ-affine. Then, differences between $d_{\mathrm{KL}}(\phi, \theta)$ and $\frac{1}{2}(\theta-\phi)^{\mathrm{T}} I(\phi)(\theta-\phi)$ reflect departures from constancy in θ-coordinates of the metric h^ϕ locally to ϕ. There is a natural geometric way to quantify such departures (namely, the corresponding Christoffel symbols for the metric h^ϕ: see Amari, (1990, p. 41–42)). These measures of *total curvature* (departure from total flatness) provide the information sought.

7.4. Preferred point Pythagoras theorem

As Amari et al. (1990) affirm, the projection theorem (see above) and the generalised Pythagorean theorem for divergences (Amari, 1990, p. 86) are the highlights of the theory of dually flat manifolds, such as Amari's expected α-geometries.

A preferred point Pythagoras theorem established in Critchley et al. (1994) provides a strengthening of this latter result.

8. Preferred point geometry of influence analysis

Following Critchley (1998), we briefly indicate how a general preferred point geometry of influence analysis in statistics might be developed. We hope to report more fully on this work in progress in the near future. For related work, see Vos (1991b, 1994) and Kass and Vos (1997).

Statistical science often proceeds by adopting a working formulation of a problem. We may stylise this as follows. Having defined a question of interest, the scientist/statistician team decide on an appropriate statistical model for the context involving one or

more unknown parameters θ and on an associated inference method, collect an optimal feasible set of relevant data, and then use their working problem formulation:

$$\mathrm{PF}^{\mathrm{o}} = (Q,\ \text{data},\ \text{model},\ \text{inference method})$$

to provide an answer A^{o} to the question of interest Q.

Now, it is natural to think of such a formulation as a *preferred point* formulation: *preferred* because it represents the team's current best shot at a 'good' (parsimonious yet realistic, etc.) description of the problem and *point* because it will only ever be one of many possible (neighbouring) problem formulations. Perturbations of problem formulation are always pertinent. Accordingly, sensitivity analyses are sensible.

In this broad conception, the rôle of influence analysis is to explore interesting alternative problem formulations PF and their effect—if any—on the answer provided to the question of interest. With ω^{o} denoting the null case of a perturbation parameter vector $\omega \in \Omega$, a change $\omega^{\mathrm{o}} \to \omega$ brings a change $\mathrm{PF}^{\mathrm{o}} \to \mathrm{PF}$ in problem formulation. Its effect $A^{\mathrm{o}} \to A$ on the answer provided to the question of interest is monitored by tracking the induced change $\tau(\omega^{\mathrm{o}}) \to \tau(\omega)$ in a suitable target function $\tau(\cdot)$. In particular contexts this may, for example, be Cook's likelihood displacement function

$$\mathrm{LD}(\omega) = 2\{l(\hat{\theta}; \omega^{\mathrm{o}}) - l(\hat{\theta}(\omega); \omega^{\mathrm{o}})\} \tag{10}$$

reflecting the change $\hat{\theta} \to \hat{\theta}(\omega)$ where $\hat{\theta}(\omega)$ maximises the perturbed log-likelihood $l(\cdot\,; \omega)$ and $\hat{\theta} := \hat{\theta}(\omega^{\mathrm{o}})$, the Kullback–Leibler divergence between posterior distributions under ω^{o} and ω, or (the expected utility of) the optimal decision procedure under ω^{o} and ω. Overall, we wish to compare the size of the perturbation $\omega^{\mathrm{o}} \to \omega$ to the size of the change $\tau(\omega^{\mathrm{o}}) \to \tau(\omega)$ it causes. Invariance of such influence analyses to reparameterisations $\omega \to \omega^{*}$ of the perturbation is highly desirable, since ω is merely a label for a problem formulation. Unfortunately, as Loynes (1986) pointed out, Cook's (1986) analysis does not have this property.

In this rather general set up, it is not immediately obvious what it means to go 'straight' from one problem formulation to another, nor how large such a perturbation is. A geometrically natural way to answer these questions invariantly is to put an appropriate metric tensor ρ^{o} on perturbation space Ω and to use the geodesics of the Riemannian preferred point geometry $(\Omega, \rho^{\mathrm{o}})$ that it induces. The preferred point nature of the metric ρ^{o} reflects the preferred nature of the working problem formulation PF^{o}. Preliminary work along these lines has been encouraging.

References

Amari, S., 1982a. Differential geometry of curved exponential families—curvatures and information loss. Ann. Statist. 10, 357–387.

Amari, S., 1982b. Geometrical theory of asymptotic ancillarity and conditional inference. Biometrika 69, 1–17.

Amari, S., 1990. Differential-Geometric Methods in Statistics, 1st Edition Lecture Notes in Statistics, Vol. 28, Springer, Berlin, 1985.

Amari, S., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L., Rao, C.R., 1987. Differential Geometry in Statistical Inference. Institute of Mathematical Statistics, Hayward, California.

Amari, S., Kumon, M., 1983. Differential geometry of Edgeworth expansions in curved exponential families. Ann. Inst. Statist. Math. 35A, 1–24.
Amari, S., Kumon, M., 1988. Estimation in the presence of infinitely many nuisance parameters—geometry of estimating functions. Ann. Statist. 16, 1044–1068.
Amari, S., Kurata, K., Nagaoka, H., 1990. Differential geometry of Boltzmann machines, Technical Report METR 90-19, Dept. Mathematical Engineering, University of Tokyo.
Atkinson, C., Mitchell, A.F.S., 1981. Rao's distance measure. Sankhya A 43, 345–365.
Barndorff-Nielsen, O.E., 1983. On a formula for the distribution of the maximum likelihood estimator. Biometrika 70, 343–365.
Barndorff-Nielsen, O.E., 1986. Likelihood and observed geometries. Ann. Statist. 14, 856–873.
Barndorff-Nielsen, O.E., 1987a. Differential and integral geometry in statistical inference. In: Amari, S., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L., Rao, C.R. (Eds.), Differential Geometry in Statistical Inference. Institute of Mathematical Statistics, Hayward, California, 95–161.
Barndorff-Nielsen, O.E., 1987b. Differential geometry and statistics: some mathematical aspects. Indian J. Math. 29, 335–350.
Barndorff-Nielsen, O.E., 1988. Parametric Statistical Models and Likelihood. Lecture Notes in Statistics 50. Springer, Berlin.
Barndorff-Nielsen, O.E., Blaesild, P., 1987a. Strings: mathematical theory and statistical aspects. Proc. Roy. Soc. London, A 411, 155–176.
Barndorff-Nielsen, O.E., Blaesild, P., 1987b. Derivative strings: contravariant aspects. Proc. Roy. Soc. London, A 411, 421–444.
Barndorff-Nielsen, O.E., Blaesild, P., 1988. Coordinate-free definition of structurally symmetric derivative strings. Adv. Appl. Math. 9, 1–6.
Barndorff-Nielsen, O.E., Cox, D.R., 1989. Asymptotic Techniques for Use in Statistics. Chapman and Hall, London.
Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Chapman and Hall, London.
Barndorff-Nielsen, O.E., Cox, D.R., Reid, N., 1986. The rôle of differential geometry in statistical theory. Internat. Statist. Rev. 54, 83–96.
Bates, D.M., Watts, D.G., 1980. Relative curvature measures of nonlinearity. J. Roy. Statist. Soc. B 42, 1–25.
Bates, D.M., Watts, D.G., 1981. Parameter transformations for improved approximate confidence intervals in nonlinear least squares. Ann. Statist. 9, 1152–1167.
Bhattacharrya, A., 1943. On discrimination and divergence. Proceedings of the 29th Indian Science Congress Vol. 13, Part III.
Bhattacharrya, A., 1946. On a measure of divergence between two multinomial populations. Sankhya 7, 401–406.
Burbea, J., Rao, C.R., 1982a. Entropy differential metrics, distance and divergence measures in probability spaces: a unified approach. J. Multivariate Anal. 12, 575–596.
Burbea, J., Rao, C.R., 1982b. On the convexity of some divergence measures based on entropy functions. IEEE Trans. Inform. Theory 28, 489–495.
Chentsov, N.N., 1972. Statistical Decision Rules and Optimal Inference, Nuaka, Moscow, translated from Russian into English (1982), AMS, Rhode Island.
Cook, R.D., 1986. Assessment of local influence (with Discussion) J. Roy. Statist. Soc. B 48, 133–169.
Critchley, F., 1998. Discussion of some algebra and geometry for hierarchical models, applied to diagnostics by J. S. Hodges. J. Roy. Statist. Soc. B 60, 528–529.
Critchley, F., Marriott, P.K., Salmon, M.H., 1992. Distances in statistics. Proceedings of the 36th Meeting Italian Statistical Society, CISU, Rome, pp. 36–60.
Critchley, F., Marriott, P.K., Salmon, M.H., 1993. Preferred point geometry and statistical manifolds. Ann. Statist. 21, 1197–1224.
Critchley, F., Marriott, P.K., Salmon, M.H., 1994. Preferred point geometry and the local differential geometry of the Kullback–Leibler divergence. Ann. Statist. 22, 1587–1602.
Critchley, F., Marriott, P.K., Salmon, M.H., 2000. An elementary treatment of Amari's expected geometry. In: Marriott, P.K., Salmon, M.H. (Eds.), Applications of Differential Geometry to Econometrics, Cambridge University Press, Cambridge, 294–315.
Dawid, A.P., 1975. Discussion of Efron's paper. Ann. Statist. 3, 1231–1234.
Dawid, A.P., 1977. Further comments on a paper by Bradley Efron. Ann. Statist. 5, 1249.

Efron, B., 1975. Defining the curvature of a statistical problem (with Discussion). Ann. Statist. 3, 1189–1217.
Efron, B., 1978. The geometry of exponential families. Ann. Statist. 6, 362–376.
Eguchi, S., 1983. Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Statist. 11, 793–803.
Eguchi, S., 1984. A characterisation of second order efficiency in a curved exponential family. Ann. Inst. Statist. Math. 36A, 199–206.
Eguchi, S., 1991. A geometric look at nuisance parameter effect of local powers in testing hypotheses. Ann. Inst. Statist. Math. 43A, 245–260.
Jeffreys, H., 1948. Theory of Probability, 2nd Edition. Clarendon Press, Oxford.
Kass, R.E., 1984. Canonical parameterizations and zero parameter effects curvature. J. Roy. Statist. Soc. B 46, 86–92.
Kass, R.E., 1989. The geometry of asymptotic inference (with Discussion). Statist. Sci. 4, 188–234.
Kass, R.E., 1990. Data-translated likelihoods and Jeffrey's rules. Biometrika 77, 107–114.
Kass, R.E., Vos, P.W., 1997. Geometrical Foundations of Asymptotic Inference. Wiley, New York.
Kullback, S.L., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22, 79–86.
Lauritzen, S.L., 1987. Statistical manifolds. In: Amari, S., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L., Rao, C.R. (Eds.), Differential Geometry in Statistical Inference. Institute of Mathematical Statistics, Hayward, California, pp. 163–216.
Loynes, R.L., 1986. Discussion of Assessment of local influence by R. D. Cook. J. Roy. Statist. Soc. B 48, 156–157.
Marriott, P.K., 1989. Applications of Differential Geometry to Statistics, Ph.D. Thesis, University of Warwick.
Murray, M.K., Rice, J.W., 1993. Differential Geometry and Statistics. Chapman and Hall, London.
Oller, J.M., Corcuera, J.M., 1995. Intrinsic analysis of statistical estimation. Ann. Statist. 23, 1562–1581.
Oller, J.M., Cuadras, C.M., 1985. Rao's distance for negative multinomial distributions. Sankhya, A 47, 75–83.
Pistone, G., Sempi, C., 1995. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Statist. 23, 1543–1561.
Rao, C.R., 1945. Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. 37, 81–89.
Rao, C.R., 1961. Asymptotic information and limiting information. Proceedings of the Fourth Berkeley Symposium of Mathematics and Statistical Probability vol. 1, pp. 531–545.
Rao, C.R., 1962. Efficient estimates and optimum inference procedures in large samples (with Discussion). J. Roy. Statist. Soc. B 24, 46–72.
Rao, C.R., 1963. Criteria of estimation in large samples. Sankhya, A 25, 189–206.
Rao, C.R., 1987. Differential metrics in probability spaces. In: Amari, S., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L., Rao, C.R. (Eds.), Differential Geometry in Statistical Inference. Institute of Mathematical Statistics, Hayward, California, pp. 217–240.
Rao, C.R., Sinha, B.K., Subramanyam, K., 1982. Third order efficiency of the maximum likelihood estimator in the multinomial distribution. Statist. Decisions 1, 1–16.
Vos, P.W., 1989. Fundamental equations for statistical submanifolds with applications to the Bartlett correction. Ann. Inst. Statist. Math. 41, 429–450.
Vos, P.W., 1991a. Geometry of f-divergence. Ann. Inst. Statist. Math. 43, 515–537.
Vos, P.W., 1991b. A geometrical approach to identifying influential cases. Ann. Statist. 19, 1570–1581.
Vos, P.W., 1992. Minimum f-divergence estimators and quasi-likelihood functions. Ann. Inst. Statist. Math. 44, 261–279.
Vos, P.W., 1994. Likelihood-based measures of influence for generalized linear models. Comm. Statist. A 23, 3477–3490.
Zhu, H.T., Wei, B.C., 1997a. Some notes on preferred point alpha-geometry and alpha-divergence function. Statist. Probab. Lett. 33, 427–437.
Zhu, H.T., Wei, B.C., 1997b. Preferred point alpha-manifold and Amari's alpha-connections. Statist. Probab. Lett. 36, 219–229.