Statistics and Differential Geometry

18-466 Mathematical Statistics

Jerome Le Ny

December 14, 2005

Abstract

It has been realised for several decades now, probably since Efron's paper introducing the concept of statistical curvature [Efr75], that most of the main concepts and methods of differential geometry are of substantial interest in connection with the theory of statistical inference. This report describes in simple cases the links existing between the two theories. It is based on an article introducing the topic, by R. Kass [Kas89]. The focus is on parametric statistical models.

Contents

Introduction

This term paper explores some connections between ideas in differential geometry and statistics. The basic idea is to view a parametric statistical model as a manifold, whose points are particular probability densities in the family under study. Then a particular parameterization of the family is a coordinate system on this manifold, and it turns out that the Fisher information matrix defines naturally a Riemannian metric on a statistical manifold. We can also study intrinsic properties of the manifold and their statistical interpretation, such as curvature, distance, etc. The most interesting applications of this differential geometrical view of statistical problems come only with the introduction of more involved techniques of tensor calculus, but in this report, we merely give a few illustrations of the geometry of some statistical procedures.

1 Elementary Differential Geometry

1.1 Manifolds

In this section we introduce (rather informally) a few concepts of differential geometry that will be useful in the following. Differential geometry studies the properties of differentiable manifolds. Intuitively, a manifold S is a Hausdorff topological space with *coordinate systems*, i.e., homeomorphisms from S or subsets of S onto open subsets of $\mathbb{R}^n$. For our purpose, we will consider the manifolds to be connected. n is called the dimension of S. In statistics, the elements or *points* of S will be probability distributions.

We call a coordinate system that has S as its domain a *global* coordinate system. In general a manifold may not have a global coordinate system (for example, the surface of a sphere cannot be homeomorphic to an open subset in $\mathbb{R}^n$), but only local coordinate systems on open subsets of S that cover S, which must be differentiably related on the intersection of their domains. However, since in statistics the focus is usually on local properties of manifolds, in the following we will consider only the case where there exists a global coordinate system, by considering only the neighborhood of interest.

Let S be a manifold and $\phi : S \to \mathbb{R}^n$ be a coordinate system for S. Then ϕ maps each point p of S to a vector $\phi(p) = [\xi^1(p), \ldots, \xi^n(p)]^T = [\xi^1, \ldots, \xi^n]^T$ of n real numbers called the coordinates of the point p. Each ξ^i may be viewed as a function $p \mapsto \xi^i(p)$ which maps p to its i^{th} coordinate. We call these n functions the coordinate functions, and we shall write the coordinate system ϕ as $\phi = [\xi^1, \ldots, \xi^n] = [\xi^i]$. Because we want to be able to change the coordinate systems and study properties that are invariant under coordinate transformations, we impose some restrictions on the allowed coordinate systems. Recall that a mapping f between open sets in a Euclidean space $\mathbb{R}^m$ is a C^∞ diffeomorphism if f is one-to-one, and f and f^{-1} are infinitely differentiable (i.e., C^∞).

Definition 1. *Let S be a set, and $\mathcal{A}$ a set of coordinate systems. We call S or $(S, \mathcal{A})$ an n-dimensional C^∞* differentiable manifold*, or more simply, a manifold, if:*

(i) each element ϕ of $\mathcal{A}$ is a homeomorphism from S to some open subset of $\mathbb{R}^n$.

(ii) For each pair ϕ, ψ in $\mathcal{A}$, the mapping $\psi \circ \phi^{-1}$ is a C^∞ diffeomorphism. $\psi \circ \phi^{-1}$ is called the coordinate transformation from ϕ to ψ.

Let f be a real-valued function on a C^∞ manifold S. The function f is called a C^∞ function on S if there exist a coordinate system $\phi = [\xi^i]$ such that $\bar{f} = f \circ \phi^{-1}$ is C^∞ on $\phi(S)$. This property does not depend on the choice of coordinate system. We define the partial derivatives of f to be $\frac{\partial f}{\partial \xi} := \frac{\partial \bar{f}}{\partial \xi} \circ \phi$ (and similarly for higher order derivatives) and denote by $\left(\frac{\partial f}{\partial \xi}\right)_p$ the value of this function at a point p. When the coordinate variables are clear (here ξ^i), we will also write $\partial_i := \frac{\partial}{\partial \xi^i}$. We will denote by $\mathcal{F}(S)$ or simply $\mathcal{F}$ the set of C^∞ functions on S. Then $\mathcal{F}$ is an algebra over $\mathbb{R}$ with the usual operations. Let S and Q be manifolds with coordinate systems $\phi : S \to \mathbb{R}^n$ and $\psi : Q \to \mathbb{R}^m$. A mapping $\lambda : S \to Q$ is said to be C^∞ or smooth if $\psi \circ \lambda \circ \phi^{-1}$ is C^∞.

Note that a coordinate transformation $\psi \circ \phi^{-1} : [\xi^1, \ldots, \xi^n] \mapsto [\rho_1, \ldots, \rho_n]$ defines new functions $\rho^i(\xi^1, \ldots, \xi^n)$ and $\xi^i(\rho^1, \ldots, \rho^n)$ which are C^∞. They satisfy

$$\sum_{j=1}^{n} \frac{\partial \xi^i}{\partial \rho^j} \frac{\partial \rho^j}{\partial \xi^k} = \sum_{j=1}^{n} \frac{\partial \rho^i}{\partial \xi^j} \frac{\partial \xi^j}{\partial \rho^k} = \delta^i_k,$$

where $\delta^i_k = 1$ if $k = i$ and 0 otherwise. In the following, we will adopt Einstein's summation convention, which means that we omit the summation sign $\sum$ corresponding to indices which are repeated in equations. Thus the previous relation could be rewritten $\frac{\partial \xi^i}{\partial \rho^j} \frac{\partial \rho^j}{\partial \xi^k} = \frac{\partial \rho^i}{\partial \xi^j} \frac{\partial \xi^j}{\partial \rho^k} = \delta^i_k$.

1.2 Tangent Vectors and Tangent Spaces

We start by defining curves and tangent vectors of curves on manifolds. Consider a one-to-one function $\gamma : I \to S$ from some interval $I \subset \mathbb{R}$ to S. By defining $\gamma^i(t) := \xi^i(\gamma(t))$, we may express the point $\gamma(t)$ using coordinates as $\bar{\gamma}(t) = [\gamma^1(t), \ldots, \gamma^n(t)]$. If $\bar{\gamma}(t)$ is C^∞ for $t \in I$, we call γ a C^∞ curve in S. Given a point $p \in S$, a curve γ such that $\gamma(a) = p$, and a function $f \in \mathcal{F}$ on S, we can consider the value of $f(\gamma(t))$ on the curve and define the derivative $\frac{d}{dt} f(\gamma(t))$ in the usual way. Using coordinates, we have $f(\gamma(t)) = \bar{f}(\bar{\gamma}(t)) = \bar{f}(\gamma^1(t), \ldots, \gamma^n(t))$, and we can rewrite the derivative using the chain rule as

$$\frac{d}{dt} f(\gamma(t)) = \left(\frac{\partial \bar{f}}{\partial \xi^i} \right)_{\bar{\gamma}(t)} \frac{d\gamma^i(t)}{dt} = \left(\frac{\partial f}{\partial \xi^i} \right)_{\gamma(t)} \frac{d\gamma^i(t)}{dt} , \tag{1}$$

We call this the directional derivative of f along the curve γ. Then we define the tangent vector of γ at p to be the operator: $\mathcal{F} \to \mathbb{R}$ which maps $f \in \mathcal{F}$ to $\frac{d}{dt} f(\gamma(t))|_{t=a}$, and define $\left(\frac{d\gamma}{dt} \right)_p = \dot{\gamma}(a)$ to be this operator. We can rewrite equation (1) as:

$$\left(\frac{d\gamma}{dt} \right)_p = \dot{\gamma}(a) = \dot{\gamma}^i(a) \left(\frac{\partial}{\partial \xi^i} \right)_p , \tag{2}$$

where $\left(\frac{\partial}{\partial \xi^i} \right)_p$ is the operator which maps $f \mapsto \left(\frac{\partial f}{\partial \xi^i} \right)_p$, and $\dot{\gamma}^i(a) = \frac{d\gamma^i}{dt}(a)$. The operator $\left(\frac{\partial}{\partial \xi^i} \right)_p$ is also the tangent vector at the point p of the i^{th} coordinate curve, obtained by fixing the values of all ξ^j for $j \neq i$ and varying only the value of ξ^i.

Now consider all curves that pass through the point p. We denote the set of all tangent vectors correponding to these curves by T_p or $T_p(S)$. We see from equation (2) that

$$T_p(S) = \left\{ c^i \left(\frac{\partial}{\partial \xi^i} \right)_p \;\middle|\; [c^1, \ldots, c^n] \in \mathbb{R}^n \right\}. \tag{3}$$

So $T_p(S)$ is a linear space of dimension $n = \dim S$. It is called the tangent space of S at p and its elements are called the tangent vectors of S at p. In addition, we call $\left(\frac{\partial}{\partial \xi^i} \right)_p$ the natural basis of $T_p(S)$ with respect to the coordinate system $[\xi^i]$.

1.3 Vector Fields

Let $X : p \mapsto X_p$ be a mapping which maps each point p in the manifold S to a tangent vector $X_p \in T_p(S)$. We call such a mapping a *vector field*. Given a coordinate system $[\xi^i]$ and the corresponding natural basis, for each point p there exist n real numbers $[X^1_p, \ldots, X^n_p]$ which uniquely determine

$$X_p = X^i_p (\partial_i)_p. \tag{4}$$

Hence we may define the functions $X^i : p \mapsto X^i_p$ on S. We call these n functions $\{X^1, \ldots, X^n\}$ the components of X with respect to $[\xi^i]$. If the components of a vector field are C^∞ with respect to some coordinate system, then they are C^∞ with respect to any other, and the vector field is called a C^∞ vector field. We shall denote these vector fields by $\mathcal{T}(S)$. Note that $\partial_i \in \mathcal{T}$ for $i = 1, \ldots, n$.

1.4 Submanifolds

Definition 2. *Let M and S be manifolds, where M is a subset of S. Let $[\xi^1, \ldots, \xi^n]$ and $[u^1, \ldots, u^m]$ be coordinate systems for S and M, respectively, where $n = \dim S$ and $m = \dim M$. We call M a submanifold of S if the following conditions hold:*

(i) The restriction $\xi^i_{|M}$ of each ξ to M is a C^∞ function on M.

(ii) Let $B^i_a := \left(\frac{\partial \xi^i_{|M}}{\partial u^a}\right)_p$ and $B_a := [B^1_a, \ldots, B^n_a] \in \mathbb{R}^n$. Then for each point p in M, $\{B_1, \ldots, B_m\}$ are linearly independent (hence $m \leq n$).

(iii) For any open subset W of M, there exists U, an open subset of S, such that $W = M \cap U$.

The conditions are independent of the choice of coordinate systems. A connected open subset of S is a manifold and also a submanifold of S. We can construct an example of a submanifold of dimension $m < n$ in the following way. Let $[\xi^i]$ be a coordinate system of S and $\{c^{m+1}, \ldots, c^n\}$ be $n - m$ real numbers. Define

$$M := \left\{p \in S \mid \xi^i(p) = c^i, m+1 \leq i \leq n\right\}. \tag{5}$$

$(M, [\xi^i_{|M}])$ is the required submanifold, assuming it is non empty. Conversely, every m-dimensional submanifold of S can be locally constructed this way.

Let M be a submanifold of the manifold S. For a point $p \in M$ we may view $T_p(M)$ as a linear subspace of $T_p(S)$. If $[\xi^i]$ and $[u^a]$ are coordinate systems for S and M, we have the equality of the differential operators: for all $f \in \mathcal{F}$, $\left(\frac{\partial f}{\partial u^a}\right)_p = \left(\frac{\partial \xi^i}{\partial u^a}\right)_p \left(\frac{\partial f}{\partial \xi^i}\right)_p$.

1.5 Riemannian metrics

Let S be a manifold of dimension n. For each point p in S, let us assume that an inner product $\langle \cdot, \cdot \rangle_p$ has been defined on the tangent space $T_p(S)$. The mapping $g : p \to \langle \cdot, \cdot \rangle_p$, associating points to corresponding inner products (which are positive definite bilinear forms) is a *Riemannian metric* on S. Given a Riemannian metric g on S, we call S or (S, g) a Riemannian manifold.

Let $[\xi^i]$ be a coordinate system for S and let $\partial_i := \frac{\partial}{\partial \xi^i}$. At each point p of S, since $(\partial_i)_p$ is a basis $T_p(M)$, the components of a Riemannian metric with respect to $[\xi^i]$ are given by a symmetric positive definite matrix $G(p) = \{g_{ij}(p)\}_{i,j=1}^n$ with $g_{ij}(p) = \langle (\partial_i)_p, (\partial_j)_p \rangle$. We require that the functions: $p \mapsto g_{ij}(p)$ be in $\mathcal{F}(S)$ (or at least can be differentiated as needed). Also, if M is a submanifold of S, $g(p)$ naturally defines an inner product on $T_p(M)$ which is a subspace of $T_p(S)$ and so we obtain a Riemannian metric on M.

Let $X, X' \in T_p$ be tangent vectors at p, $X = X^i(\partial_i)_p$, $X' = X'^i(\partial_i)_p$. Then we have

$$\langle X, X' \rangle_p = g_{ij}(p) X^i X'^j,$$

and the length $\|X\|$ of the tangent vector X is given by $\|X\| = \sqrt{\langle X, X \rangle}$. We can easily obtain the relationships between the components g_{ij} and the components $\tilde{g}_{kl}$ with respect to another coordinate system $[\rho^k]$, as

$$\tilde{g}_{kl} = g_{ij} \left(\frac{\partial \xi^i}{\partial \rho^k}\right) \left(\frac{\partial \xi^j}{\partial \rho^l}\right) \quad \text{and} \quad g_{ij} = \tilde{g}_{kl} \left(\frac{\partial \rho^k}{\partial \xi^i}\right) \left(\frac{\partial \rho^l}{\partial \xi^j}\right). \tag{6}$$

Now let $\gamma : (a, b) \to S$ be a curve in S. We define its length $\|\gamma\|$ to be

$$\|\gamma\| := \int_a^b \left\| \frac{d\gamma}{dt} \right\| dt = \int_a^b \sqrt{g_{ij}\dot{\gamma}^i\dot{\gamma}^j} dt. \tag{7}$$

Using (6), we can show that this length does not depend on the choice of coordinate system. Then if t_0 is an arbitrary point in (a, b), we can define the arc length s on the curve by

$$s(t) = \int_{t_0}^t \|\dot{\gamma}(u)\| du.$$

If the tangent vector of γ is nonzero throughout the domain of γ, then s is differentiable with nonzero derivative. By the inverse function theorem, there exist an inverse transformation $c : s((a, b)) \to (a, b)$, and s defines a parameterization of $\gamma \circ c$, such that $\gamma \circ c$ and γ have the same image. In practice, we may speak of the arc length parameterization of γ when we really mean the parameterization of the curve $\gamma \circ c$. With this convention, we have $\dot{\gamma}(t) = \dot{\gamma}(s)\ s'(t)$, and since $s'(t) = \|\dot{\gamma}(t)\| \neq 0$ we see that

$$\|\dot{\gamma}(s)\| = 1. \tag{8}$$

2 Information-Metric Riemannian Geometry

2.1 Manifolds of Densities

Differential geometry emphasizes intrisic properties of manifolds, that is, properties that do not depend on extrinsic coordinate expressions. In statistics, we can structure a parametric family of probability densities as a smooth manifold, and consider at once all possible parameterizations of the parameter space.

Let $\{P_\theta, \theta \in \Theta\}$ be a family of laws on the sample space $(\mathbf{X}, \mathcal{B})$, dominated by a σ-finite measure ν on $(\mathbf{X}, \mathcal{B})$, with $p_\theta(x) = p(\theta, x) := \frac{dP_\theta}{d\nu}(x)$. We assume that Θ is an open subset of $\mathbb{R}^n$ and that the mapping $\theta \to p_\theta$ is injective and C^∞ and that when necessary we can take derivatives inside the integral sign (see section 2.3). Then we call $S = \{p_\theta \mid \theta = [\theta^1, \ldots, \theta^n] \in \Theta\}$ a statistical model, a parametric model, or simply a model on $\mathbf{X}$. So S is parameterized using n real-valued variables $[\theta^1, \ldots, \theta^n]$, and we may want to consider other parameterizations. We will also consider only families such that $p_\theta(x) > 0$, for all $\theta \in \Theta$ and $x \in \mathbf{X}$, and therefore S is a subset of

$$\mathcal{P}(\mathbf{X}) := \left\{ p : \mathbf{X} \to \mathbb{R} \mid\ p(x) > 0\ \forall x \in \mathbf{X}, \int p(x) d\nu(x) = 1 \right\}.$$

Given a statistical model $S = \{p_\theta \mid \theta \in \Theta\}$, the mapping $\phi : S \to \mathbb{R}^n$ defined by $\phi(p_\theta) = \theta$ allows us to consider $\phi = [\theta^i]$ as a coordinate system for S. There are some topological issues involved to satisfy definition 1 (we need in particular a topology on the densities, such as the weak topology), but here we will just assume that S can indeed be considered as a manifold, that we may call a statistical manifold. In this case, a parameterization of S is in fact also a coordinate system of S. This allows us to consider points in S without reference to a particular parameterization. Then we can define the likelihood function based on x as

$$L_x : S \to \mathbb{R}$$
$$L_x(p) = p(x)$$

and the log-likelihood becomes $l_x(p) = \log(L_x(p))$. A maximum likelihood point, if one exists, is a point $\hat{p}$ for which $L_x(\hat{p}) = \max_{p \in S}\ L_x(p)$.

2.2 Curved Exponential Families (CEF)

Exponential families and curved exponential families (CEF), i.e., subfamilies of exponential families, constitute important background for the ideas of information geometry. In particular, there is a geometric characterization of exponential families and an associated computable criterion, which allows us to decide exactly when a family is exponential (see [MR93], chapter 1; we will not pursue this discussion here).

Let $S = \{p_\theta \mid \theta \in \Theta\}$ be a statistical model corresponding to an exponential family in a minimal representation, where Θ is the natural parameter space. We sometimes use the term *full* exponential family to emphasize the fact that the entire parameter space Θ is considered. Moreover, if Θ is an open subset of $\mathbb{R}^n$, as it will be assumed here, the exponential family is called *regular* using the terminology in [Kas89], following [BN78]. We write the densities $p_\theta(x) = e^{\theta \cdot T(x) - j(\theta)}$ with respect to a σ-finite dominating measure μ, with $K(\theta) = \{\int e^{\theta \cdot T(x)} d\mu(x)\}$ and $j(\theta) = \log K(\theta)$. Recall from Theorems 2.5.5, 2.5.7 and Corollary 2.5.8 in the course notes that Θ is a convex set, $j(\theta)$ is strictly convex on Θ, K and j have derivatives of all orders that may be computed by differentiating under the integral sign, the moments of T of all orders exist and the mean vector and covariance matrix of j are given by (see also equation (14)):

$$E_\theta T = \nabla j(\theta) \tag{9}$$

$$V_\theta T = \nabla^2 j(\theta), \tag{10}$$

where $\nabla^2 j(\theta)$ is the Hessian matrix of j at θ. From the convexity of j we deduce immediately that the log-likelihood function $l_x(\theta) := \log p(\theta, x)$ is concave.

For convenience in section 3.2, we rename the random variables $Y = T(X)$ and let the corresponding new sample space be $\mathbf{Y}$. We let $\mu(\theta) = E_\theta Y$, and call the image space of this mapping the *mean-value parameter space*, denoted M. For an exponential family, a nice fact is that the mean value can be used as a parameterization. So we may also take μ to stand for the mean-value, when used as parameter for the family.

Theorem 3. *For an exponential family, the mapping $\mu : \theta \to \mu(\theta) = E_\theta Y$ is a C^∞ diffeomorphism.*

Proof. First, it follows from the lemma below that μ is one-to-one.

Lemma 4. *For all $\theta, \theta^* \in \Theta$,*

$$(\theta - \theta^*)^T \{\mu(\theta) - \mu(\theta^*)\} \geq 0$$

and equality holds if and only if $\theta = \theta^*$.

Proof of the lemma. Because Θ is open, we can extend define $f(\alpha) = j(\alpha\theta + (1-\alpha)\theta^*) = j(\theta^* + \alpha(\theta - \theta^*))$ for α in $[-\epsilon_1, 1 + \epsilon 2]$, for some $\epsilon_1, \epsilon_2 > 0$. We have

$$\begin{aligned} f'(\alpha) =& \nabla j(\theta^* + \alpha(\theta - \theta^*))^T (\theta - \theta^*) \\ f''(\alpha) =& (\theta - \theta^*)^T \nabla^2 j(\theta^* + \alpha(\theta - \theta^*))(\theta - \theta^*). \end{aligned}$$

In corollary 2.5.8 in the lecture notes, it is shown that $\nabla^2 j(\theta)$ is positive definite. Therefore for $\theta \neq \theta^*$, we get $f''(\alpha) > 0$ and so $f'(\alpha)$ is increasing on $(0, 1)$. In particular, we have $f'(0) < f'(1)$, which is the inequality claimed, using (9). □

So from the lemma, $\mu(\theta) = \mu(\theta^*)$ implies $\theta = \theta^*$, i.e., the mapping is one-to-one. Smoothness was already mentioned as part of Theorem 2.5.7 in the course notes. Since $D\mu(\theta) = \nabla^2 j(\theta)$ is positive definite, it follows from the inverse function theorem that the inverse mapping is also smooth.

□

By a *curved exponential family*, we mean a set of probability densities which forms a submanifold within a full exponential family. [Kas89] requires an additional topological condition to avoid inconsistencies of MLE's, which will appear in definition 5. We can think of a subfamily as the subset of distributions of the full exponential family for which the parameter θ is restricted to a subspace Θ_0 of Θ. A natural way of generating subfamilies, to which we will limit our discussion, is when Θ_0 is obtainable from an open subset B of $\mathbb{R}^k$ by a one-to-one mapping $\beta \to \theta(\beta)$, which must satisfy certain regularity conditions. We will restrict our attention to curves (i.e., one-dimensional subfamilies) within the full parameter space.

Definition 5. *[One-Parameter Curved Exponential Family] A subfamily of a full exponential family is a one-parameter curved exponential family if B is an open interval in $\mathbb{R}$ and*

(i) the mapping $\beta \to \theta(\beta)$ is one-to-one, C^∞, and $\partial_\beta \theta(\beta)$ is nowhere equal to the zero vector;

(ii) writing $\phi : \Theta_0 \to B$ for the inverse mapping, if a sequence $\{\theta_n \in \Theta_0\}$ converges to a point $\theta_0 \in \Theta_0$, then the corresponding sequence $\{\phi(\theta_n) \in B\}$ must converge to $\phi(\theta_0) \in B$.

The condition that the gradient does not vanish is present to ensure the consistency of the likelihood equations under the two different parameters. Under the conditions in the definition, $\beta \to \theta(\beta)$ is said to be an imbedding and Θ_0 is imbedded in Θ.

2.3 The Fisher metric as a Riemannian Metric

Let $S = \{p_\theta \mid \theta \in \Theta\}$ be a statistical model, and denote $l(\theta, x) := \log p(\theta, x)$. Moreover, we assume that Θ is an open set in $\mathbb{R}^m$, denote $\theta = (\theta_1, \ldots, \theta_m)$, and recall from the lecture notes that the Fisher information matrix is defined as:

$$I(\theta)_{ij} := E_\theta \left[\partial_i l(\theta, x) \, \partial_j l(\theta, x) \right], \tag{11}$$

if the partial derivatives exist and have finite variance (by definition, $\partial_i := \frac{\partial}{\partial \theta_i}$). Here is an alternate expression for the matrix $I(\theta)$, assuming we can take derivatives under the integral sign and that the needed derivatives exist. We have then

$$E_\theta[\partial_i l] = \int \partial_i p(\theta, x) d\nu(x) = \partial_i \int p(\theta, x) d\nu(x) = 0.$$

Applying ∂_j we obtain

$$\begin{aligned} 0 = &\int \partial_j [\partial_i l(\theta, x) p(\theta, x)] d\nu(x) \\ = &E_\theta[\partial_j \partial_i l(\theta, x)] + E_\theta[\partial_i l(\theta, x) \partial_j l(\theta, x)] \end{aligned}$$

and so

$$I(\theta)_{ij} = -E_\theta[\partial_i \partial_j l(\theta, x)]. \tag{12}$$

Another useful representation is

$$I(\theta)_{ij} = 4 \int \partial_i \sqrt{p(\theta, x)} \, \partial_j \sqrt{p(\theta, x)} d\nu(x). \tag{13}$$

Finally in the case of an exponential family, we have an explicit form of the log-likelihood function in (11); we get $I(\theta) = E_\theta[(T - \nabla j(\theta))(T - \nabla j(\theta))^T]$ and since $E_\theta T = \nabla j(\theta)$ we have

$$I(\theta) = V_\theta T = \nabla^2 j(\theta). \tag{14}$$

Now the matrix $I(\theta)$ is symmetric, and in general it is positive semi-definite, since for an m-dimensional vector $c^T = [c^1, \ldots, c^m]^T$ we have

$$c^T I(\theta) c = \int \left\{ \sum_{i=1}^m c^i \partial_i l(\theta, x) \right\}^2 p(\theta, x) d\nu(x) \geq 0.$$

We assume further that $I(\theta)$ is positive definite, for all θ in Θ. This happens for example in the case of an exponential family, as was proved in Corollary 2.5.8 in the lecture notes. Then $I(\theta)$ may be used to define a Riemannian metric $\langle \cdot, \cdot \rangle$ on S, which is known as the *Fisher metric* or *information metric*. Like any Riemannian metric, the information metric is invariant over the choice of coordinate system (i.e., parameterization). Using (4), a coordinate free expression may be written

$$\begin{aligned} &\langle ., . \rangle : \mathcal{T}(S) \times \mathcal{T}(S) \to \mathbb{R} \\ &\langle X_p, Y_p \rangle_p = E_p \left[X_p(l_x) Y_p(l_x) \right]. \end{aligned}$$

Indeed, if $X_p = X_p^i (\partial_i)_p$ and $Y_p = Y_p^j (\partial_j)_p$, we obtain

$$\langle X_p, Y_p \rangle_p = \sum_{i,j} X_p^i Y_p^j E_p[(\partial_i(l_x))_p (\partial_j(l_x))_p] = \sum_{i,j} X_p^i Y_p^j \; I(p)_{ij}.$$

Note that if x becomes a vector of N i.i.d. observations, we have a corresponding statistical model $S_N = \{p_\theta^N \mid \theta \in \Theta\}$. We know that the information matrix is simply multiplied by N (so distances are multiplied by $\sqrt{N}$) and so it is not necessary to distinguish between the geometries of S_N and S, which are simply related by a scale factor of N. Also, an important property of the information metric, and a consequence of its definition as a Riemannian metric, is that it defines a metric on any submanifold of S. This will be implicitly used in the next section.

2.4 Example: Geometry of the Trinomial Family

We can describe some of the concepts of information-metric geometry using the case of the trinomial family, which turns out to be a sphere in that geometry.

Let $\mathcal{Q}$ be the trinomial family with $n = 1$ trial. Let $p^i > 0,\ i \in \{1, 2, 3\}$, be the probabilities of each outcome, $p^1 + p^2 + p^3 = 1$: the multinomial distribution with $n = 1$ has distribution $P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = (p^1)^{x_1} (p^2)^{x_2} (p^3)^{x_3}$, for $x_i \in \{0, 1\}$, $i \in \{1, 2, 3\}$ and $x_1 + x_2 + x_3 = 1$. It is a two-dimensional exponential family since we can rewrite

$$(p^1)^{x_1} (p^2)^{x_2} (p^3)^{x_3} = \exp[x_1 \log(p^2/p^1) + x_2 \log(p^3/p^1) + \log\, p^1].$$

Instead of the natural parameter space, we can consider the parameterization

$$z^i = 2\sqrt{p^i}$$

such that to each triple (p^1, p^2, p^3) in the simplex $\{(p^1, p^2, p^3) : p^1 + p^2 + p^3 = 1\}$ corresponds a point on the positive orthant portion of the sphere of radius 2. A one-dimensional imbedded subfamily $\mathcal{Q}_0$ (see definition 5) may be represented as a curve c having components $\mathbf{z}(\beta) = (z^1(\beta), z^2(\beta), z^3(\beta)),\ \beta \in B$, on the sphere. The tangent vector to the curve c is $\partial_\beta z = (\partial_\beta z^1, \partial_\beta z^2, \partial_\beta z^3)$, and its squared length is $\langle \partial_\beta z, \partial_\beta z \rangle$. Now we have $\sqrt{p(\beta, x)} = (z^1/2)^{x_1} (z^2/2)^{x_2} (z^3/2)^{x_3}$, so we can compute the Fisher information using (13)

$$\begin{aligned} I(\beta) =& 4 \sum_{x_1 + x_2 + x_3 = 1} \left(\partial_\beta \sqrt{p(\beta, x)} \right)^2 = (\partial_\beta z^1)^2 + (\partial_\beta z^2)^2 + (\partial_\beta z^3)^2 \\ =& \langle \partial_\beta z, \partial_\beta z \rangle_2, \end{aligned} \tag{15}$$

where $\langle \cdot, \cdot \rangle_2$ is here the standard Euclidean inner product. Thus the Euclidean length of the tangent vector to the curve c is $\|\partial_\beta z\|_2 = I(\beta)^{1/2}$. Consequently, from equation (7) the associated Euclidean length of the curve c between $z(\beta)$ and $z(\beta^*)$ corresponding to two elements Q and Q^* of $\mathcal{Q}$ is

$$d_c(Q, Q^*) = \int_\beta^{\beta^*} \|\partial_\beta z\|_2 \; d\beta = \int_\beta^{\beta^*} I(\beta)^{1/2} d\beta.$$

In general for a one-parameter subfamily corresponding to a curve c, we call the distance $d_c(\beta_1, \beta_2) = \int_{\beta_1}^{\beta_2} I(\beta)^{1/2} d\beta$ between two distributions associated to β_1 and β_2 the *information distance*. As mentioned earlier, this distance does not depend on the choice of parameterization of the curve.

Now for the full trinomial family, we can again consider a coordinate system $\theta = [\theta_1, \theta_2]$ on the 2-dimensional manifold $\mathcal{Q}$ and compute the information matrix the same way:

$$\begin{aligned} I(\theta)_{ij} =& 4 \sum_{x_1+x_2+x_3=1} \left(\partial_i \sqrt{f(\theta, x)}\right) \left(\partial_j \sqrt{f(\theta, x)}\right) = \sum_{k=1}^{3} \left(\partial_i z^k\right) \left(\partial_j z^k\right) \\ =& \langle \partial_i z, \partial_j z \rangle_2, \end{aligned}$$

for $i, j \in \{1, 2\}$. That is, the (i, j)-component of the Fisher information matrix is the inner product of the i^{th} and j^{th} coordinate tangent vectors on the surface of the sphere.

To define the information distance between two multinomial distributions Q and Q^*, we consider all possible curves on the sphere connecting the two corresponding points. Each curve represents a one-parameter subfamily for which the information distance between Q and Q^* can be defined as before. Then the information distance between Q and Q^* as members of the full trinomial family is defined as the minimum of the distances taken over all curves connecting the two points. The curve that achieves this minimum is called a *geodesic*. In this case it is an arc of the great circle through the points z and z^* on the sphere. From this, we deduce that the information distance between Q and Q^* is the angle between z and z^* (which is $\langle z/2, z^*/2 \rangle_2$) multiplied by 2, the radius of the sphere. Therefore

$$d(Q, Q^*) = 2 \arccos \sum_{i=1}^{3} (p_i p_i^*)^{1/2}.$$

2.4.1 Information Distance, Hellinger Distance and Kullback-Leibler Information

In this section, we use the multinomial geometry to show the link between the Hellinger distance, the Kullback-Leibler information and the information distance defined in the previous section.

If P and Q are probability measures with densities p and q with respect to a σ-finite measure μ, the Hellinger distance between the two is $h(P, Q)$ defined by

$$h^2(P, Q) = \int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mu = 2 - 2 \int \sqrt{pq} d\mu,$$

and the Kullback-Leibler information is

$$K(P, Q) = \int log\left(\frac{p}{q}\right) p \, d\mu.$$

For the case of two distributions Q and Q^* in the trinomial family, we get

$$h(Q, Q^*) = \left(\sum_{i=1}^{3} ({p^i}^{1/2} - {p^{i*}}^{1/2})^2\right)^{1/2} = \frac{1}{2} \|z - z^*\|_2 = 2 \sin(d(Q, Q^*)/4),$$

where the last equality comes from the fact already mentioned that the angle between z and z^* is $d(Q, Q^*)/2$. Hence as $d(Q, Q^*) \to 0$ we have $h(Q, Q^*) \sim \frac{1}{2} d(Q, Q^*)$, and so the two distances behave essentially identically.

For the Kullback-Leibler information, we have

$$K(Q, Q^*) = \sum_{i=1}^{3} p^i \log(p^i / p^{i*}).$$

Now as $p \to p^*$, we have by Taylor expansion

$$\log(p^{i^*}) - \log(p^i) = -\frac{1}{p^i}(p^{i*} - p^i) + \frac{1}{2{p^i}^2}(p^{i*} - p^i)^2 + O((p^{i*} - p^i)^3)$$

so that

$$K(Q, Q^*) = \frac{1}{2}\sum_{i=1}^{3} \frac{(p^{i*} - p^i)^2}{p^i} + O(\|p^* - p\|^3).$$

Let the point Q^* approach Q along a curve c (assumed smooth), that we can parametrize by the arc length s. Then

$$\begin{aligned}(p^i(s^*) - p^i(s))^2 =& (\partial_s p^i(s))^2(s^* - s)^2 + O(|s^* - s|^3)\\ =& (\partial_s p^i(s))^2 d_c(Q, Q^*)^2 + O(|s^* - s|^3),\end{aligned}$$

the second equality being obtained by definition of arc length. Now note that

$$\begin{aligned}\sum_{i=1}^{3} \frac{1}{p^i(s)}(\partial_s p^i(s))^2 =& \sum_{i=1}^{3} (\partial_s \log p^i(s))^2 p^i(s)\\ =& \sum_{x_1+x_2+x_3=1} (\partial_s \log p(s,x))^2\, p(s,x)\\ =& I(s)\\ =& \|\partial_s z(s)\|^2 \qquad \text{by (15)}\\ =& 1 \qquad \text{by (8)}.\end{aligned}$$

Therefore we obtain

$$K(Q, Q^*) = \frac{1}{2} d_c(Q, Q^*)^2 + O(d_c(Q, Q^*)^3).$$

In particular if c is the geodesic curve between Q and Q^*, we have $K(Q, Q^*) \sim \frac{1}{2} d(Q, Q^*)^2$ and we see that the Kullback-Leibler information behaves locally like the square of a distance function (however, $K(Q, Q^*) \neq K(Q^*, Q)$ in general).

2.4.2 Jeffreys' Prior

Within the field of Bayesian statistics, it is natural to ask how to select a prior on the parameter space. Let $S = \{p_\theta | \theta = [\theta^1, \ldots, \theta^n] \in \Theta\}$ be a statistical model, and suppose that the volume $V := \int \sqrt{\det(I(\theta))} d\theta$ with respect to the Fisher metric is finite (the integral is n-fold). Then $\pi(\theta) = \frac{1}{V}\sqrt{\det(I(\theta))}$ defines a probability distribution on Θ, called Jeffreys' prior. Since it is invariant over the choice of the coordinate system $[\theta^i]$, we may consider it as a probability distribution on the model S. We can calculate it for the trinomial family, introducing spherical polar coordinates, and find that it is uniform on the sphere.

3 Geometry of Information Loss and Recovery

3.1 Fisher's Measure of Information

Here we review some classical properties of the Fisher information matrix related to information loss. Let $S = \{p_\theta \mid \theta \in \Theta\}$ be a statistical model on a sample space $(\mathbf{X}, \mathcal{B})$, and let T and A be statistics, i.e., measurable functions from $\mathbf{X}$ to another measurable space $(\mathbf{Y}, \mathcal{C})$. A statistic T induces a new model $S_T := \{q_\theta \mid \theta \in \Theta\}$, where $q_\theta(y)$ is the density of the statistic T obtained from p_θ. Similarly to (11), we can define the Fisher information matrix $I_T(\theta)$ of the induced model. In the following theorem, for two symmetric matrices M and N, by $M \preceq N$ we mean $N - M$ is positive semi-definite. We have:

Theorem 6. *The Fisher information matrices $I(\theta)$, $I_T(\theta)$ and $I_A(\theta)$ of the original and induced models satisfy:*

(i) $0 \preceq I_T(\theta) \preceq I(\theta)$.

(ii) $I_T(\theta) = I(\theta)$ *if and only if* T *is sufficient.*

(iii) $I_A(\theta) = 0$ *if and only if the distribution of* A *does not depend on* θ*, i.e.,* A *is* ancillary.

(iv) The information loss $\Delta I(\theta) = I(\theta) - I_T(\theta)$ *caused by summarizing the data* x *into* $y = T(x)$ *is given by*

$$\Delta I(\theta)_{ij} = E_\theta \left[Cov_\theta[\partial_i l(\theta, X), \partial_j l(\theta, X) \mid T(X)] \right]$$

(v) $I_{T,A}(\theta) = I_T(\theta) + I_A(\theta)$ *if* T *and* A *are independent. In general* $I_{T,A}(\theta) = I_{T|A}(\theta) + I_A(\theta)$.

Proof. We will start by proving (iv). Let P_θ and Q_θ be probability distributions with densities p_θ and q_θ with respect to a σ-finite measure ν, as defined at the beginning of this section ($Q_\theta = P_\theta \circ T^{-1}$). We assume as before that derivation with respect to θ^i and integration can be interchanged. First for all $C \in \mathcal{C}$, we have

$$\begin{aligned}\int_{T^{-1}(C)} \partial_i \log(q_\theta \circ T(x))\, dP_\theta(x) = \int_C \partial_i \log(q_\theta(y))\, dQ_\theta(y) = \partial_i \int_C q_\theta(y) d\nu(y) \\ = \partial_i \int_C dQ_\theta(y) = \partial_i \int_{T^{-1}(C)} dP_\theta(x) \\ = \int_{T^{-1}(C)} \partial_i \log(p_\theta(x)) dP_\theta(x),\end{aligned}$$

so we have

$$\partial_i \log q_\theta(T(X)) = E_\theta[\partial_i l(\theta, X) | T(X)]. \tag{16}$$

Now if we define $r_\theta(x)$ by

$$p_\theta(x) = r_\theta(x) q_\theta(T(x)), \tag{17}$$

we have

$$\partial_i l(\theta, x) = \partial_i \log q_\theta(T(x)) + \partial_i \log r_\theta(x), \tag{18}$$

and so by the previous equality we get $E_\theta[\partial_i \log r_\theta(X) | T(X)] = 0$. Thus it follows from [Dud03], Theorem 10.2.9. that $\log r_\theta(X)$ is orthogonal to any function $f(T(X))$ with respect to the inner product $\langle\langle \Phi, \Psi \rangle\rangle_\theta = E_\theta[\Phi(X)\Psi(X)]$. In particular we have

$$E_\theta[\partial_i \log r_\theta(X)\ \partial_j \log q_\theta(T(X))] = 0, \quad \forall i, j. \tag{19}$$

Now by definition

$$\begin{aligned}&\mathrm{Cov}_\theta[\partial_i l(\theta, X), \partial_j l(\theta, X) \mid T(X)] \\ &= E_\theta \left[\{\partial_i l(\theta, X) - E_\theta[\partial_i l(\theta, X) | T(X)]\} \{\partial_j l(\theta, X) - E_\theta[\partial_j l(\theta, X) | T(X)]\} | T(X) \right] \\ &= E_\theta \left[\partial_i \log r_\theta(X)\ \partial_j \log r_\theta(X) | T(X) \right], \quad \text{by (18),(16).}\end{aligned}$$

Therefore

$$\begin{aligned}E_\theta \left[\mathrm{Cov}_\theta[\partial_i l(\theta, X), \partial_j l(\theta, X) \mid T(X)] \right] &= E_\theta \left[\partial_i \log r_\theta(X)\ \partial_j \log r_\theta(X) \right] \\ &= I(\theta)_{ij} - I_T(\theta)_{ij}, \quad \text{by (18),(19).}\end{aligned}$$

This finishes the proof of (iv). Then (i) becomes obvious. The condition $\Delta I(\theta) = 0$ in (ii) is that $\partial_i \log r_\theta(x) = 0$ for all θ, i, x, and this is equivalent to T being sufficient by the factorization theorem (see (17)). (iii) is clear from the definition of $I_A(\theta)$. (v) follows from the fact that if $p_\theta(x) = p_{\theta,1}(x_1) p_{\theta,2}(x_2)$, then the cross products in $I_{12}(\theta)$ vanish under our assumptions:

$$\int \int \partial_i \log p_{\theta,1}(x_1) \partial_j \log p_{\theta,2}(x_2)\ p_{\theta,1}(x_1) p_{\theta,2}(x_2) d\nu(x) = \partial_i \partial_j \int \int p_\theta(x) d\nu(x) = 0.$$

□

Information loss was defined in the theorem above. Now suppose (T, A) is sufficient and A is ancillary, then $I_{T|A}(\theta) = I(\theta)$. Thus, the information lost by an insufficient estimator T can be recovered by conditioning on an appropriate statistic A.

Let us consider for simplicity the case where $\Theta \subset \mathbb{R}$. The information lost by an estimator T can be quantified in terms of $I_T(\theta)/nI_1(\theta)$ or $nI_1(\theta) - I_T(\theta)$, where $I_1(\theta)$ is the information per observation in a sample of size n. It is easier to look only at the limiting values (as $n \to \infty$), which is 1 for the ratio form for efficient estimators in an exponential family (see next section). Then we can distinguish among the different estimators using $nI_1(\theta) - I_T(\theta)$ (which will typically tend to infinity for inefficient estimators). What Fisher called information loss is $\lim_{n\to\infty}(nI_1(\theta) - I_T(\theta))$, and he claimed that among all efficient estimators, the MLE minimizes the information loss. There are geometric interpretations of this claim, as illustrated in the next section.

3.2 Geometrical Interpretation of Estimation in CEF

3.2.1 Auxiliary Space Associated with an Estimator

Recall the definition of a one-parameter curved exponential family with parameter $\beta \in B$ (definition 5); it corresponds to a submanifold S_0 in the manifold S of the full exponential family. We have defined the mean-value parameter space M for the full exponential family, and the submanifold S_0 can be specified by a subset $M_0 \subset M$. Recall also the notation introduced before theorem 3 for exponential families. Usually an estimator is a mapping from the sample space $\mathbf{Y}$ into the the parameter space B, but we will restrict the class of possible estimators T that we consider. Suppose we have a sample of i.i.d. observations $Y_1, \ldots, Y_n$, then their mean $\bar{Y}_n$ determines a point $\hat{\mu}$ in S whose $\mu-$coordinates are $\bar{Y}_n$. Since $\bar{Y}_n$ is a sufficient statistic for S, hence also for S_0, we will only consider estimators which are functions of $\hat{\mu}$. Now really these estimators are mappings from S to B, such that $T(\hat{\mu}) = \hat{\beta}$. However, since we will not deal with advanced concepts of differential geometry, here we can identify S with the space M, and consider the estimators to be maps from M to B, as is done in [Kas89]. This definition includes many important estimators, and simplifies the estimation process: for any given estimator, we do not need to consider a sequence of mappings, but rather need only examine a single mapping, and then the geometrical interpretation becomes revealing.

Let V be an open neighborhood in M such that $V \cap M_0 \neq \emptyset$. An estimator $T : V \to B$ will be called regular if T is smooth with nonzero derivative and for all β such that $\mu(\beta) \in V$, we have $T(\mu(\beta)) = \beta$. Let $\mu_0 \in M_0$, W be a neighborhood of μ_0 in M, C an open subset of B, and $f : C \times W \to \mathbb{R}$ a smooth function with $f(\beta, \cdot) : W \to \mathbb{R}$ having everywhere a nonzero gradient, such that $f(\beta, \mu(\beta)) = 0$ for all $\beta \in C$. Then by the implicit function theorem there exists an open neighborhood V of μ_0 in W on which a regular estimator T is uniquely defined by the equation $f(T(y), y) = 0$. For example, about each point μ_0 in a CEF there is a neighborhood on which the MLE is regular, the function f being given by the likelihood equation

$$f(\beta, y) = (y - \mu(\beta))^T \partial_\beta \theta(\beta),$$

corresponding to the minimization of $\theta(\beta)^T y - j(\theta(\beta))$, with $\mu(\beta) = \nabla j(\theta(\beta))$.

Now consider the replacement of $\bar{Y}_n$ by a sufficient statistic (T, A). From the relation $I_{T,A}(\theta) = I_T(\theta) + I_{A|T}(\theta)$, we see that the information not contained in T must be contained in the conditional distribution of A given T. The geometrical interpretation of this is based on the following proposition.

Proposition 7. *If T is a regular estimator, then for each $\beta_0 \in B$, there exists a neighborhood U of $\mu(\beta_0)$ in M and a coordinate system (T_U, A) of U onto an open subset of $\mathbb{R} \times \mathbb{R}^{k-1}$ such that T_U is the restriction of T to U and $U \cap M_0 = \{\mu \in U : A(\mu) = 0\}$.*

Note that it does not seem that A is ancillary in general. This proposition corresponds to the local construction of a submanifold mentioned in (5). We will not elaborate on this point which consists in the construction of imbedded submanifolds in differential geometry. What is

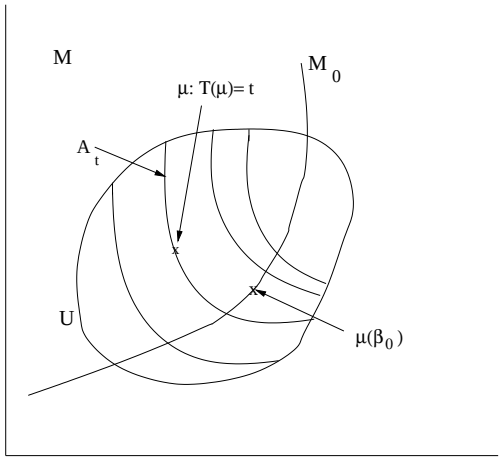

Figure 1: Local decomposition of M near $\mu(\beta_0)$.

important to us is that locally, the estimator T determines a decomposition of the manifold of the full exponential family into subspaces $A_t = \{\mu \in U : T(\mu) = t\}$ (see figure 1). A_t is called the auxiliary subspace associated with the estimator T at t. Within the neighborhood U of $\mu(\beta_0)$, the local coordinate system maps points μ to $(T(\mu), A(\mu))$, with $A(\mu) = 0$ for points on the curve M_0. If the true parameter of the CEF is β_0, then with probability one $\bar{Y}_n$ will fall in U for n sufficiently large, as follows from the law of large numbers. Then $(T(\bar{Y}_n), A(\bar{Y}_n))$ becomes sufficient and the amount of information lost by T is the amount contained in the distribution of $A(\bar{Y}_n)$ given T.

3.2.2 Efficiency

First we consider the relationship of the Fisher information matrices for exponential families in terms of parameterization by the natural parameter θ and by μ, and show that they are inverses of each other. This is a straightforward application of the change of variable formula (6). Let $l_y(\mu) = \log(p(y, \theta(\mu)))$ be the log likelihood on M.

Proposition 8. *In an exponential family, the Fisher information matrices $I(\theta)$ and $I(\mu)$ satisfy*

$$\begin{aligned} I(\theta) =& D\mu(\theta) = \nabla^2 j(\theta), \\ I(\mu) =& D\theta(\mu) = I(\theta(\mu))^{-1}. \end{aligned}$$

Proof. The first relations are immediate from (9) and (14). The chain rule (6) can be written

$$I(\mu) = D\theta(\mu) I(\theta(\mu)) D\theta(\mu)^T. \tag{20}$$

A consequence of Theorem 3 and the inverse function theorem is

$$D\theta(\mu) = D\mu(\theta)^{-1}. \tag{21}$$

Now we can combine (20), (21) and the first relations to finish the proof. □

As a consequence of the Delta method, and of the fact that $I(\theta) = \nabla^2 j(\theta) = V_\theta(n\bar{Y}_n)$ for an exponential family, we have the following theorem:

Theorem 9. *For a regular estimator T that is consistent for a parameter β of a CEF,*

$$avar_\beta(T)^{-1/2}\sqrt{n}(T(\bar{Y}_n) - \beta) \to N(0,1) \quad \textit{in distribution}$$

where $avar_\beta(T) = [\nabla_\mu T(\mu(\beta))]^T I(\mu(\beta))^{-1} [\nabla_\mu T(\mu(\beta))]$ is called the asymptotic variance of T.

We define the *efficiency* of a regular estimator T to be the ratio $I(\beta)^{-1}/\text{avar}_\beta(T)$. We can interpret $\text{avar}_\beta(T)$ geometrically and see that its minimum value is $I(\beta)^{-1}$. Suppose T is defined by an estimating equation $f(T(\bar{y}), \bar{y}) = 0$, with $f : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$. If we denote the partial derivatives of f with respect to its two arguments by $D_1 f$ and $D_2 f$ ($D_1 f$ is a scalar), by differentiating $f(T(\mu), \mu) = 0$ with respect to μ we get

$$0 = D_1 f(\nabla_\mu T(\mu)) + D_2 f,$$

and so $\nabla_\mu T(\mu) = -(D_1 f)^{-1}(D_2 f)$. Then we can rewrite

$$\text{avar}_\beta(T) = (D_2 f)^T I(\mu(\beta))^{-1}(D_2 f)(D_1 f)^{-2}. \tag{22}$$

By definition the auxiliary space at β is given by $A_\beta = \{\mu : f(\beta, \mu) = 0\}$. Now $D_2 f$ is the gradient of $f(\beta, \cdot)$ and so it is normal to A_β with respect to the Euclidean inner product. Denoting as before $\langle \cdot, \cdot \rangle_{\mu(\beta)}$ the inner product associated to the Fisher information matrix at $\mu(\beta)$, we have then for any vector v tangent to A_β:

$$\begin{aligned} 0 =& v^T(D_2 f(\beta, \mu(\beta))) = v^T I(\mu(\beta))[I(\mu(\beta))^{-1} D_2 f(\beta, \mu(\beta))] \\ 0 =& \langle v, n_\beta \rangle_{\mu(\beta)}, \end{aligned}$$

where $n_\beta = I(\mu(\beta))^{-1} D_2 f(\beta, \mu(\beta))$, which is thus seen to be normal to A_β at $\mu(\beta)$ with respect to the $\langle \cdot, \cdot \rangle_{\mu(\beta)}$ inner product. This will be useful in the proof of the following theorem.

Theorem 10. *The efficiency of a regular estimator T based on i.i.d. observations from a CEF is given by*

$$\frac{I(\beta)^{-1}}{avar_\beta(T)} = \sin^2(\phi),$$

where ϕ is the angle between M_0 and A_β (i.e., between $\partial_\beta \mu(\beta)$ and A_β), with respect to $\langle \cdot, \cdot \rangle_{\mu(\beta)}$.

Proof. We begin with

$$\langle n_\beta, n_\beta \rangle_{\mu(\beta)} = (D_2 f)^T I(\mu(\beta))^{-1}(D_2 f).$$

Next we rewrite $D_1 f$ by differentiating $f(\beta, \mu(\beta)) = 0$ with respect to β:

$$0 = D_1 f + (D_2 f)^T (\partial_\beta \mu(\beta))$$

and so

$$-D_1 f = \langle n_\beta, \partial_\beta \mu(\beta) \rangle_{\mu(\beta)}.$$

We use these expressions in (22) to obtain

$$\text{avar}_\beta(T) = \langle n_\beta, n_\beta \rangle_{\mu(\beta)} \left(\langle n_\beta, \partial_\beta \mu(\beta) \rangle_{\mu(\beta)} \right)^{-2}.$$

Now we also have $I(\beta) = \langle \partial_\beta \mu(\beta), \partial_\beta \mu(\beta) \rangle_{\mu(\beta)}$, as in (20), and so we get

$$\frac{I(\beta)^{-1}}{\text{avar}_\beta(T)} = \frac{(\langle n_\beta, \partial_\beta \mu(\beta) \rangle_{\mu(\beta)})^2}{\langle n_\beta, n_\beta \rangle_{\mu(\beta)} \langle \partial_\beta \mu(\beta), \partial_\beta \mu(\beta) \rangle_{\mu(\beta)}}.$$

The right-hand side is the squared cosine of the angle between n_β and $\partial_\beta \mu(\beta)$ with respect to $\langle , \rangle_{\mu(\beta)}$, which is $(\pi/2) - \phi$. This proves the theorem using $\cos(\pi/2 - \phi) = \sin(\phi)$. □

Finally in the case of the MLE, from the likelihood equations,

$$0 = f(\beta, \bar{Y}_n) = (\bar{Y}_n - \mu(\beta))^T (\partial_\beta \theta)$$

we get using proposition 8

$$D_2 f = \partial_\beta \theta = D\theta(\mu)\, \partial_\beta \mu = I(\mu(\theta))\, \partial_\beta \mu,$$

and thus $n_\beta = \partial_\beta \mu$. Thus the angle between M_0 and A_θ is $\pi/2$, which maximizes $\sin^2 \phi$. So from Theorem 10, we have a geometric interpretation that the MLE minimizes the asymptotic variance, i.e., is efficient: in this case the auxiliary space associated with the MLE and the CEF are orthogonal. Efficiency of maximum likelihood estimators is proved more generally in section 3.9 of the course notes.

Conclusion

We have given examples of geometrical arguments that can provide insight into some statistical procedures. A drawback is that in general, this requires strong regularity conditions, but the motivation is that geometrical understanding can then be used as a basis for other proofs. Moreover, it appears that more advanced concepts of differential geometry than those presented here, in particular involving covariant differentiation and some affine connections introduced by Amari actually provide results that would be difficult to obtain otherwise (see [AN00] for an introduction).

References

[AN00] S-I. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[BN78] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, New York, 1978.

[Dud03] R.M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2 edition, 2003.

[Efr75] B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, (3):1189–1217, 1975.

[Kas89] R.E. Kass. The geometry of asymptotic inference. *Statistical Science*, 4(3):188–234, 1989. With discussion.

[MR93] M.K. Murray and J.W. Rice. *Differential Geometry and Statistics*. Chapman and Hall, 1993.