Statistical Science
1989, Vol. 4, No. 3, 188–234

The Geometry of Asymptotic Inference

Robert E. Kass

Abstract. Geometrical foundations of asymptotic inference are described in simple cases, without the machinery of differential geometry. A primary statistical goal is to provide a deeper understanding of the ideas of Fisher and Jeffreys. The role of differential geometry in generalizing results is indicated, further applications are mentioned, and geometrical methods in nonlinear regression are related to those developed for general parametric families.

Key words and phrases: Information, distance measures, Jeffreys' prior, Bayes factor, orthogonal parameters, curved exponential family, statistical curvature, approximate sufficiency, ancillary statistic, nonlinear regression.

1. INTRODUCTION

In statistical science, geometrical methods are ubiquitous, their analytical and aesthetic virtues taken for granted. Differential geometry, on the other hand, is largely unfamiliar to most statisticians and may seem rather technical. My purpose in this paper is to show how two simple and appealing ideas lead naturally to the introduction of differential geometrical structure in problems of parametric inference and, further, how the geometrical approach succeeds here, as elsewhere, adding clarification and insight as well as techniques that can produce new results. I will attempt this using very little differential geometry itself, presuming no previous knowledge on the part of the reader. As a pedagogical device, I will discuss extremely simple special cases, indicating only briefly the more complete treatment that may be compiled from the references.

The first of the two ideas is to base a local measure of distance between members of a family of distributions on what is most commonly called the Kullback-Leibler number or, equivalently, on Fisher information. This led Rao (1945) and Jeffreys (1946) to introduce a Riemannian metric defined by Fisher information—Rao in his paper on what is now called the Cramér-Rao lower bound, and Jeffreys in his paper on an invariant prior for estimation problems, which is sometimes known as Jeffreys' general rule. The second idea is to connect the special role of exponential families in statistical theory with their loglinear structure. This led Efron (1975) to quantify departures from exponentiality by defining the curvature of a statistical model. As pointed out by Dawid (1975) in his discussion of Efron's paper, the two ideas are related, but the appropriate foundation for Efron's measure of curvature actually involves non-Riemannian differential geometry.

Rao developed his work further in several papers, applying it in his studies of genetic diversity. (See Rao, 1987, for references.) Here, however, I will concentrate on Jeffreys' uses of the Fisher information metric, elaborating on the geometrical basis of his methods in greater detail than did Jeffreys himself. Meanwhile, the description I will give of the work initiated by Efron and developed further by others, primarily Amari (see Amari, 1987a; and Kass, 1987) will be oriented toward its most basic achievement: a thorough and concise geometrical interpretation of information loss, Fisher's fundamental quantification of departure from sufficiency, and information recovery, his justification for conditioning. Thus, in one sense this is a paper about Jeffreys and Fisher. On the other hand, it is also a highly selective review of the ideas and writing of many authors, some of whom I have borrowed heavily from in preparing my own presentation.

I am not intending to survey geometrical results in asymptotic inference, though I will cite some related references in two short bibliographical sections. Perhaps the most egregious deficiency is that I reduce to only a few words the line of research followed by Barndorff-Nielsen and his colleagues (see Barndorff-Nielsen, 1986b, 1987a). Nor will I have much to say about geometrical methods in nonlinear regression, though I will make some remarks about their relationship with the work that concerns asymptotic inference more broadly.

The justification I offer for a lengthy exposure to a comparatively narrow view of the field is twofold. First, many further geometrical investigations of

Robert E. Kass is Associate Professor of Statistics at Carnegie Mellon University. His mailing address is: Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

statistical inference originate from material reviewed here. Second, the aspects of Jeffreys' and Fisher's theories under discussion relate to issues that are, in my view, of ongoing vital importance. These are the role of reference priors in Bayesian inference and the use of conditioning in non-Bayesian inference. The paper is not directly about these issues, and it in no way resolves them. Still, the geometrical interpretations do lead to a deeper understanding of the concepts, and with this there is, as I will try to indicate, at least some sharpening of arguments through the elimination of irrelevant diversions. By providing elementary details in the discussion of a few topics, I hope to convey a sense of the statistical and mathematical concepts involved in the geometry of asymptotic inference.

1.1 Outline

Section 1.2 contains preliminary material, which should at least be skimmed since some basic notation and terminology is introduced. Section 2 concerns the Riemannian geometry based on Fisher information with Section 2.1 confined to the trinomial family and Section 2.2 covering generalizations. The generalizations are interwoven with descriptions and definitions of Riemannian geometrical objects. Technical details are omitted, and the intention is to provide just enough material for interested readers to get some feeling for the way the development proceeds. Subsequent sections do not depend on this one, so it may be skipped by those who are mainly interested in the statistical results. In Section 2.3 Jeffreys' uses of geometry are discussed.

Section 3 is devoted to the geometry of information loss and recovery. Sections 3.1 through 3.4 treat one-parameter families, with an emphasis on "statistical curvature." Generalizations are presented in Section 3.5, with results stated in terms of scalar curvatures; these scalar curvatures are compared with curvature measures in nonlinear regression in Section 3.5.6. The differential geometrical foundation is only briefly mentioned, in Section 3.6. I close, in Section 4, with a couple of brief comments on the role of geometry in asymptotic inference.

1.2 Preliminaries

1.2.1 Derivative Notation and Terminology

In addition to various common notations for derivatives, I will often use two conventions that may be unfamiliar to many readers. The first is to indicate partial derivatives by ∂_{ij} when the variables serving as arguments are identified by the context. The second is to use brackets rather than parentheses in identifying the point at which a derivative is evaluated. In addition, the point at which a derivative is evaluated will often be omitted. Thus, I might use any of the expressions in the equation

$$(D^2\psi)_{ij} = \partial_{ij}\psi = \partial_{ij}\psi[\eta] = \frac{\partial^2\psi}{\partial\eta_i\partial\eta_j}\bigg|_\eta,$$

and likewise any of those in the equation

$$D_\eta\psi = D\psi(\eta) = (\partial_i\psi[\eta]).$$

The rank of a differentiable transformation $\psi: R^m \to R^k$ is the rank of its derivative $D\psi$. To say that a transformation is of full rank is to say it is of full rank throughout its domain.

1.2.2 Parametric Families and Diffeomorphisms

Throughout the paper, parametric families of distributions $\mathscr{P} = \{P_\theta: \theta \in \Theta\}$ will be discussed. When Y_i is a random variable having distribution P_θ, its density will be denoted by $p(y_i \mid \theta)$. The likelihood function based on a sample $y = (y_1, \cdots, y_n)$ will be written as $L(\theta)$ or $L_y(\theta)$ (boldface for y will not be used), and the loglikelihood function will be written as $l(\theta)$ or $l_y(\theta)$ or, when an expectation is to be taken, as $l_Y(\theta)$. Regularity conditions will be mentioned only in a few places, but at the outset it is worth noting that for most purposes we will want to assume the identifiability condition, $P_\theta = P_{\theta'}$, if and only if $\theta = \theta'$ and the open parameter space condition, Θ, is an open subset of R^m (which would be better described as a Euclidean topology condition, see Section 2.3).

We may instead specify the family in terms of an alternative parameterization λ, writing $\mathscr{P} = \{P_\lambda: \lambda \in \Lambda\}$, as long as the new parameterization again satisfies the identifiability and open parameter space conditions. The first requires the mapping from Θ to Λ to be one-to-one, while the second is satisfied when the mapping is a homeomorphism (both it and its inverse are continuous). In addition, we will need to know that derivatives of functions of parameters are available regardless of the parameterization used. Thus, for example, when we wish to assume that the negative Hessian of the loglikelihood function at its maximum is positive definite, we will not want to stop and worry whether this assumption will hold if a new parameterization λ is substituted for an original parameterization θ. Since $D^2l(\hat{\lambda}) = (D_\lambda\theta)^T \cdot D^2l(\hat{\theta}) \cdot D_\lambda\theta$, the condition needed to insure this kind of invariance is that the one-to-one transformation from θ to λ and its inverse are smooth (infinitely differentiable) and of full rank (the Jacobians $D_\theta\lambda$ and $D_\lambda\theta$ are invertible). Such a transformation is called a *diffeomorphism.*

In asymptotics, many arguments use parameters that do not identify the full family $\mathscr{P}$, but only part of it, often in the neighborhood of some particular distribution P_{θ_0}. Such parameterizations are *local* as

opposed to *global*. When we consider at once all possible parameterizations (both local and global) that are related to each other as diffeomorphisms on their common domains, we are structuring $\mathscr{P}$ as a *smooth manifold*. A bit more will be said about this in Section 2.3, but an essential point is that if we wish results to be invariant with respect to differentiable transformations of parameters, we impose smooth manifold structure.

1.2.3 Subfamilies and Imbeddings

The restriction of the family $\mathscr{P}$ according to a null hypothesis H_0: $\theta \in \Theta_0$, where $\Theta_0 \subseteq \Theta$, may be written $\mathscr{P}_0 = \{P_\theta: \theta \in \Theta_0\}$. Here we will be interested in the situation in which Θ_0 is of dimension $k < m$, and part of the purpose of this subsection is to say precisely what this means. The simplest case occurs when the parameter space for $\mathscr{P}$ takes the form $\Psi \times B$ and the parameter spce for $\mathscr{P}_0$ is $\{(\psi, \beta) \in \Psi \times B: \psi = \psi_0\}$, i.e., we would write $H_0: \psi = \psi_0$ with $\beta \in B$ being a nuisance parameter. Note only are some arguments simplified when $\mathscr{P}$ is parameterized according to a product structure, but many basic results actually require the existence of such a structure, at least locally. A different way of stating this requirement is to assume Θ_0 is an *imbedded submanifold* of Θ.

Specifically, Θ_0 must be obtainable from an open subset B of R^k by a one-to-one mapping $\beta \to \theta(\beta)$ such that (1) the mapping is smooth and of full rank k and (2) 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 $\phi(\theta_0) \in B$.

Condition (1) is enough to give meaning to the statement that the dimension of Θ_0 is k; technically, it makes Θ_0 an immersed submanifold of Θ. Condition (2) is a continuity condition, which is needed to ensure, for example, consistency of maximum likelihood estimates. In the remainder of the subsection, I will elaborate somewhat on these remarks using two examples, which will recur in subsequent sections.

When a null hypothetical subfamily is not given in product form, it is usually specified in one of two alternative forms. The first may be written $\Theta_0 = \{\theta \in \Theta: \theta = \theta(\beta), \beta \in B\}$ while the second may be written $\Theta_0 = \{\theta \in \Theta: \psi(\theta) = \psi_0\}$. Both are frequently used, the first being found most commonly in goodness-of-fit and curved exponential family applications. Each may be illustrated using subfamilies of the trinomial family $\mathscr{P} = \{P_{(\theta_1,\theta_2)}: 0 < \theta_1 < 1, 0 < \theta_2 < 1, 0 < 1 - \theta_1 - \theta_2 < 1\}$, where $P_{(\theta_1,\theta_2)}$ is the Trinomial $(n; \theta_1, \theta_2)$ distribution: if $y = (y_1, y_2, y_3)$ satisfies $y_1 + y_2 + y_3 = n$, we have $p(y \mid \theta_1, \theta_2) = (n!/y_1!y_2!y_3!)\theta_1^{y_1}\theta_2^{y_2}(1 - \theta_1 - \theta_2)^{y_3}$.

Example. Hardy–Weinberg model. The Binomial $(2; \beta)$ subfamily of the Trinomial $(n; \theta_1, \theta_2)$ defined by

$$\theta_1(\beta) = \beta^2 \quad \theta_2(\beta) = 2\beta(1 - \beta),$$

so that $\theta_3(\beta) \equiv 1 - \theta_1(\beta) - \theta_2(\beta) = (1 - \beta)^2$, is called the Hardy–Weinberg model. It furnishes an example of the first kind of specification of a subfamily, without product structure, mentioned above. The "null hypothesis" here is that the Hardy–Weinberg model holds, $\Theta_0 = \{\theta \in \Theta: \theta = \theta(\beta), \beta \in (0, 1)\}$.

Example. Symmetry model. The subfamily of the Trinomial $(n; \theta_1, \theta_2)$ defined by $\theta_1 = \theta_3$ with $\theta_3 \equiv 1 - \theta_1 - \theta_2$ will be called the symmetry model. Here we may consider the conditional probability that an observation falls in the first category, given that it is in either the first or the third, $\psi(\theta) = \theta_1/(1 - \theta_2)$. The null hypothetical submodel is then specified by $\Theta_0 = \{\theta \in \Theta: \psi(\theta) = \tfrac{1}{2}\}$.

In each of these examples, there are many apparent ways of defining the additional parameter, ψ in the first case, β in the second, so that the Trinomial space will be filled out. For instance, in the Hardy–Weinberg model we could define $\psi(\theta) = \theta_2/(2\theta_1^{1/2}(1 - \theta_1^{1/2}))$, and in the symmetry model we could define $\beta(\theta) = \theta_2$. In each example, we could then begin an analysis with the (ψ, β) parameterization and a null hypothetical specification $H_0: \psi = \psi_0$ ($\psi_0 = 1$ corresponding to the Hardy–Weinberg model).

Asymptotic theory of subfamilies often relies on the existence of a *local* reparameterization, in product form, in some neighborhood of each point θ_0 in the null hypothetical space Θ_0. Writing $\mathscr{P}^U = \{P_\theta \in \mathscr{P}: \theta \in U\}$, the condition may be stated formally as follows:

> For each $\theta_0 \in \Theta_0$ there exists an open neighborhood $U \subseteq \Theta$ and a parameterization (ψ, β) of $\mathscr{P}^U$, meaning $\mathscr{P}^U = \{P_{(\psi,\beta)}: (\psi, \beta) \in \Psi \times B\}$, such that $\theta_0 \in U$ and $\mathscr{P}_0^U = \{P_{(\psi,\beta)}: (\psi, \beta) \in \Psi \times B, \psi = \psi_0\}$.

Using the inverse function theorem, this condition may be shown to be equivalent to the imbedded submanifold assumption. Thus the imbedded submanifold assumption amounts to an assertion that, for local (asymptotic) results, "without loss of generality" the null hypothesis may be written $H_0: \psi = \psi_0$ for some (ψ, β).

1.2.4 Arclength and Curvature of Curves

The simplest definition of the curvature of a curve is in terms of the arclength parameterization. If $c: I \to R^k$ is a twice differentiable curve with nonzero derivative (where I is an interval), it may be parameterized by arclength s, for which

$$\| D_s c \| = 1 \tag{1.1}$$

and then its *curvature* at s^* is $\kappa = \kappa_{s^*} = \| D_s^2 c[s^*] \|$. One way to motivate this definition of curvature is to

assume the curve is twice continuously differentiable and then show that the circle that best "fits" the curve at $c(s^*)$ (the "osculating circle") has radius κ^{-1}; since the inverse of the radius is an obvious measure of the curvature of a circle, the inverse of the "instantaneous radius" of the curve, i.e., the inverse of the radius of the best-fitting circle, is an intuitive general measure.

The arclength parameterization is useful for some purposes, but not others. An alternative expression, in terms of an arbitrary parameter, is obtained by decomposing the second derivative vector $D_t^2 c$ into a component $(D_t^2 c)_T$ tangent to the curve, i.e., in the direction of $D_t c$, and the residual normal component $(D_t^2 c)_N = D_t^2 c - (D_t^2 c)_T$. We have the curvature formula

$$\| D_s^2 c \| = \| (D_t^2 c)_N \| \cdot \| D_t c \|^{-2} \tag{1.2}$$

which will be used in Section 3.

1.2.5 Exponential Families

In this section, I set up notation while reviewing a few relevant properties of exponential families. Detailed treatments may be found in Barndorff-Nielsen (1978) and Brown (1986). Let $\mathscr{E}$ be a *canonical* or *natural exponential family* of order k, having densities of the form

$$p(y \mid \eta) = \exp[y^T \eta - \psi(\eta)] h(y)$$

with respect to a dominating measure ν. The elements of $\mathscr{E}$ will be denoted by Q_η. The natural parameter space will be denoted by N, i.e.,

$$N = \left\{ \eta \in R^k : \int \exp[y^T \eta] h(y)\, d\nu(y) < \infty \right\}.$$

If for each η in N there exists Q_η in $\mathscr{E}$, then $\mathscr{E}$ is said to be full; if, in addition, N is open as a subset of R^k, then $\mathscr{E}$ is said to be *regular*.

In a regular exponential family, N is a convex subset of R^k, ψ is a strictly convex analytic function on N, the moments of Y of all orders exist and the mean and variance are given by

$$E_\eta(Y) = D\psi[\eta],$$

$$V_\eta(Y) = D^2\psi[\eta]$$

with $D^2\psi[\eta]$ being positive definite. Let $\mu = \mu(\eta) = E_\eta(Y)$, so that μ may be considered a mapping of N into R^k. In a regular exponential family, the mapping $\mu\colon N \to R^k$ is one-to-one and smooth (infinitely differentiable). Since $D\mu = D^2\psi$ is positive definite, it then follows by the Inverse Function Theorem that the inverse mapping is also smooth, and $\mu(\cdot)$ is a diffeomorphism of N onto its image. The image space $\mu(N)$, which is the mean-value parameter space, will be denoted by M. Exponential families discussed throughout the paper will be assumed to be regular.

The basic result needed for asymptotics is the following. Suppose $Y_1, \cdots, Y_n$ are iid observations from an element Q_η of $\mathscr{E}$. With probability one, for sufficiently large n there exists a unique MLE $\hat{\eta}$, which may be found as the unique root of the likelihood equations; in addition, $\hat{\eta}$ is strongly consistent and

$$[ni(\eta)]^{1/2}(\hat{\eta} - \eta) \xrightarrow{\mathscr{L}} N_k(0, I_k)$$

where $i(\eta)^{1/2}$ is the positive definite square-root of the information matrix $i(\eta) = D^2\psi[\eta]$.

2. INFORMATION-METRIC RIEMANNIAN GEOMETRY

How far apart are two members of a parametric family of probability distributions? From the point of view of asymptotics, it would make sense to measure distance using asymptotic standard deviation units of the best estimators (or posterior distributions) of the parameter. In the one-dimensional case, the asymptotic standard deviation is $[ni(\theta)]^{-1/2}$ so that something like $[i(\theta)]^{1/2}$ times the magnitude of the deviation of the two parameter values would seem appropriate. I say "something like" because it is not clear what value of θ should be used in $[i(\theta)(\theta_1 - \theta_2)^2]^{1/2}$. One could fix an arbitrary θ_0 and measure all distances relative to $i(\theta_0)$, but the resulting numbers would depend on both the parameterization and the value θ_0 that were chosen. An alternative is to integrate the infinitesimal version of this quantity, i.e., to use

$$d(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} i(\theta)^{1/2}\, d\theta \tag{2.1}$$

as a distance measure. This is the *information distance*.

Jeffreys observed that the Kullback–Leibler number generates the information distance via the *information metric* (in the terminology used here), and this led to his general rule for determining priors, which is to take the prior density to be proportional to the square root of the determinant of the information matrix. Jeffreys also revealed his geometrical viewpoint in his treatment of odds factors for hypothesis tests based on "orthogonal" parameters, an orthogonal parameterization being one in which the information matrix is diagonal. I will return to Jeffreys' methods after first discussing the geometry of the information metric. The main example I will treat will be the multinomial, because no abstract geometrical notions are needed in that case.

2.1 Spherical Multinomial Geometry

In this section, I illustrate many aspects of information-metric geometry with a spherical representation of the trinomial family.

2.1.1 Spherical Representation

Let $\mathscr{E}$ be the trinomial family with $n = 1$ and let $\mathscr{E}_0$ be a one-dimensional imbedded subfamily (see Section 1.2.3). Since $\mathscr{E}$ is an exponential family, it may be represented in terms of the mean value and natural parameter spaces, but a common alternative is to use the simplex $\{(p_1, p_2, p_3): p_1 + p_2 + p_3 = 1, p_i > 0; i = 1, 2, 3\}$ in R^3. Instead of the simplex, consider the positive-orthant portion of the sphere of radius 2, defined by

$$z_i = 2\sqrt{p_i}, \quad i = 1, 2, 3, \tag{2.2}$$

so that the simplex relation becomes $z_1^2 + z_2^2 + z_3^2 = 4$ with $z_i > 0$, $i = 1, 2, 3$. The subfamily $\mathscr{E}_0$ may then be represented as a curve

$$\mathbf{z}(\beta) = (z_1(\beta), z_2(\beta), z_3(\beta)), \quad \beta \in B,$$

on the sphere. The reason for doing this comes from the following calculation.

The tangent vector to the curve $\mathbf{z}$ is $D_\beta \mathbf{z} = (D_\beta z_1, D_\beta z_2, D_\beta z_3)$. Its squared length is

$$\begin{aligned} \langle D_\beta \mathbf{z}, D_\beta \mathbf{z} \rangle &= \sum_{i=1}^{3} (D_\beta z_i)^2 \\ &= \sum_{i=1}^{3} (D_\beta (2p_i^{1/2}))^2 \\ &= \sum_{i=1}^{3} p_i(\beta)^{-1} (D_\beta p_i)^2 \qquad (2.3) \\ &= \sum_{i=1}^{3} p_i(\beta) (D_\beta \log p_i)^2 \\ &= i(\beta). \end{aligned}$$

Thus, the length of the tangent vector to the curve $\mathbf{z}$ is

$$\| D_\beta \mathbf{z} \| = i(\beta)^{1/2}. \tag{2.4}$$

Combining (2.1) and (2.4), the information distance between two elements Q and Q^* of $\mathscr{E}_0$ is the length of the curve $\mathbf{z}$ between $\mathbf{z}(\beta)$ and $\mathbf{z}(\beta^*)$,

$$d(Q, Q^*) = \int_\beta^{\beta^*} \| D_\beta \mathbf{z} \| \cdot d\beta. \tag{2.5}$$

Note that the length of the curve $\mathbf{z}$ between $\mathbf{z}(\beta)$ and $\mathbf{z}(\beta^*)$ does not depend on the parameterization of the curve, and thus the information distance also does not depend on the parameterization of the model in (2.1). In addition, it is immediate from (1.1) and (2.4) that arclength is a variance-stabilizing parameterization, i.e., a parameterization in which $i(\beta)$ is constant.

Expression (2.4) may be extended to an interpretation of Fisher information for the full trinomial family. Suppose θ is a parameterization for $\mathscr{E}$ (i.e., the mapping from the (p_1, p_2) parameter space to Θ is a diffeomorphism, see Section 1.2.2), so that Θ is in R^2 and $\mathbf{z}$ becomes a function of $\theta = (\theta_1, \theta_2)$. Then, working backwards in (2.3), and considering p_i as a function of θ,

$$\begin{aligned} i(\theta)_{jk} &= \sum_{i=1}^{3} p_i(\theta) (\partial_j \log p_i)(\partial_k \log p_i) \\ &= \sum_{i=1}^{3} (\partial_j z_i)(\partial_k z_i) \qquad (2.6) \\ &= \langle \partial_j \mathbf{z}, \partial_k \mathbf{z} \rangle \end{aligned}$$

where $\partial_j \mathbf{z} = \partial_{\theta_j} \mathbf{z}$, etc., and j and k are either 1 or 2. That is, the (j, k)-component of Fisher information is the inner product of the jth and kth coordinate tangent vectors on the surface of the sphere defined by (2.2).

To define the information distance between two multinomial distributions Q and Q^*, let q and q^* be the points on the sphere that correspond to Q and Q^* according to (2.2) and consider all possible curves connecting q and q^*. Each curve represents a one-parameter subfamily $\mathscr{E}_0$ within which the information distance between Q and Q^* may be defined, using (2.5). Let $d_c(Q, Q^*)$ denote the information distance between Q and Q^* as members of the one-parameter family represented by the curve c. Then the information distance between Q and Q^* as members of the full trinomial family $\mathscr{E}$ is

$$d(Q, Q^*) = \min d_c(Q, Q^*) \tag{2.7}$$

where the minimum is taken over all curves c connecting the points q and q^*. The curve that achieves this minimum is an arc of the great circle through q and q^*, and is called a *geodesic*.

Since the information distance between Q and Q^* in $\mathscr{E}$ having values of (p_1, p_2, p_3) and (p_1^*, p_2^*, p_3^*), respectively, is the length of the great circle connecting the corresponding $\mathbf{z}$ and $\mathbf{z}^*$ vectors defined by (2.2), it is equal to the angle between $\mathbf{z}$ and $\mathbf{z}^*$ multiplied by 2 (the radius of the sphere and, therefore, also the circle). The dot product of the unit vectors $\mathbf{z}/2$ and $\mathbf{z}^*/2$ is the cosine of the desired angle and, therefore, the information distance is

$$d(Q, Q^*) = 2 \cdot \arccos \sum_{i=1}^{3} (p_i p_i^*)^{1/2}. \tag{2.8}$$

2.1.2 Information Distance, Hellinger Distance and Kullback–Leibler Number

Information distance is a simple transformation of Hellinger distance, here denoted by d_H:

$$d_H(Q, Q^*) = \left(\sum_{i=1}^{3} (p_i^{1/2} - p_i^{*1/2})^2 \right)^{1/2}$$
$$= (\tfrac{1}{2}) \| \mathbf{z} - \mathbf{z}^* \|$$
$$= 2 \cdot \sin(d(Q, Q^*)/4).$$

It may be noted, in addition, that

$$d_H(Q, Q^*) = (\tfrac{1}{2}) d(Q, Q^*) + O(d(Q, Q^*)^3) \tag{2.9}$$

as $Q^* \to Q$, so that for small distances the two distance functions are essentially the same (other than the factor ½).

There is also a close relationship between information distance and the Kullback–Leibler number, or information divergence, defined by

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

As $Q^* \to Q$ there is,

$$-\log(p_i^*/p_i) = -\log(1 + (p_i^* - p_i)/p_i)$$
$$= -(p_i^*)/p_i - (\tfrac{1}{2})(p_i^* - p_i)^2/p_i^2$$
$$+ O((p_i^* - p_i)^3)$$

so that

$$K(Q, Q^*)$$
$$= (\tfrac{1}{2}) \sum_{i=1}^{3} (p_i^* - p_i)^2/p_i + O\left(\sum_{i=1}^{3} |p_i^* - p_i|^3 \right). \tag{2.10}$$

Now as Q^* approaches Q, it does so along some path. Parameterize this path by arclength (of the corresponding path of $\mathbf{z}^*$ as it approaches $\mathbf{z}$ on the sphere). As one consequence, from (1.1) and (2.3),

$$\sum p_i(s)(D_s \log p_i)^2 = 1 \tag{2.11}$$

where s is arclength. As another, $s^* - s = d(Q, Q^*)$ for $\mathbf{z} = \mathbf{z}(s)$ and $\mathbf{z}^* = \mathbf{z}(s^*)$. Thus,

$$(p_i(s^*) - p_i(s))^2$$
$$= (D_s p_i(s^* - s))^2 + O(|s^* - s|^3) \tag{2.12}$$

and combining (2.10)–(2.12),

$$K(Q, Q^*) = (\tfrac{1}{2}) d(Q, Q^*)^2 + O(d(Q, Q^*)^3). \tag{2.13}$$

Equation (2.13) makes precise the idea that the Kullback–Leibler number behaves locally like the square of a distance function (which was the starting point for Jeffreys). Notice, too, that (2.10) provides the well-known relationship between the Kullback–Leibler number and the chi-squared discrepancy measure. From (2.13) we also obtain approximate equality of information distance and chi-squared discrepancy. These identities will lead to a geometrical derivation of familiar asymptotic chi-squared distributional results in Section 2.1.5.

2.1.3 Jeffreys' Prior

A different use for the spherical representation is in picturing Jeffreys' prior $\pi(\theta) \propto \det(i(\theta))^{1/2}$. To calculate this determinant for the multinomial, let us introduce spherical polar coordinates (θ, ϕ) defined by

$$\theta = \arccos(\delta^{1/2}),$$
$$\phi = \arcsin(\gamma^{1/2}),$$

where

$$\delta = p_3,$$
$$\gamma = p_2/(p_1 + p_2).$$

See Figure 1. The trinomial likelihood becomes

$$L(\delta, \gamma) = \gamma^{y_2}(1 - \gamma)^{y_1} \delta^{1 - y_1 - y_2}(1 - \delta)^{y_1 + y_2}$$

and the information matrix determinant is

$$\det(i(\delta, \gamma)) = |\delta\gamma(1 - \gamma)|^{-1}.$$

Changing to (θ, ϕ) parameters then yields

$$\det(i(\theta, \phi))^{1/2} d\theta d\phi$$
$$= \det(i(\delta, \gamma))^{1/2} d\delta d\gamma$$
$$= \frac{4 \cos(\theta)\sin(\theta)\cos(\phi)\sin(\phi) d\theta d\phi}{\cos(\theta)\sin(\phi)\cos(\phi)} \tag{2.14}$$
$$= 4 \sin(\theta) d\theta d\phi,$$

which is the element of surface area on the sphere of radius 2. That is, (2.14) shows that Jeffreys' prior is uniform on this sphere.

2.1.4 Orthogonal Parameters

Let (θ_0, ϕ_0) be the spherical polar coordinates of the point q_0 on the sphere (2.2), as shown in Figure 1. The vectors tangent to the coordinate curves $\phi = \phi_0$ and $\theta = \theta_0$ are denoted by $\partial/\partial\theta = \partial/\partial\theta\,|_{(\theta_0, \phi_0)}$ and $\partial/\partial\phi = \partial/\partial\phi\,|_{(\theta_0, \phi_0)}$, respectively, and are also shown in Figure 1. The coordinates θ and ϕ are orthogonal in the sense that $\partial/\partial\theta$ and $\partial/\partial\phi$ are perpendicular. Since δ and γ are obtained by separate transformation of θ and ϕ, they too are orthogonal. By (2.6), a parameter pair is orthogonal whenever the information matrix in terms of that pair is diagonal.

2.1.5 Imbedded Subfamilies and Asymptotically Chi-squared Statistics

Suppose $\mathscr{E}_0$ is a one-parameter imbedded subfamily of the trinomial family $\mathscr{E}$ (as in the examples of Section 1.2.3). To say that $\mathscr{E}_0$ is an imbedded

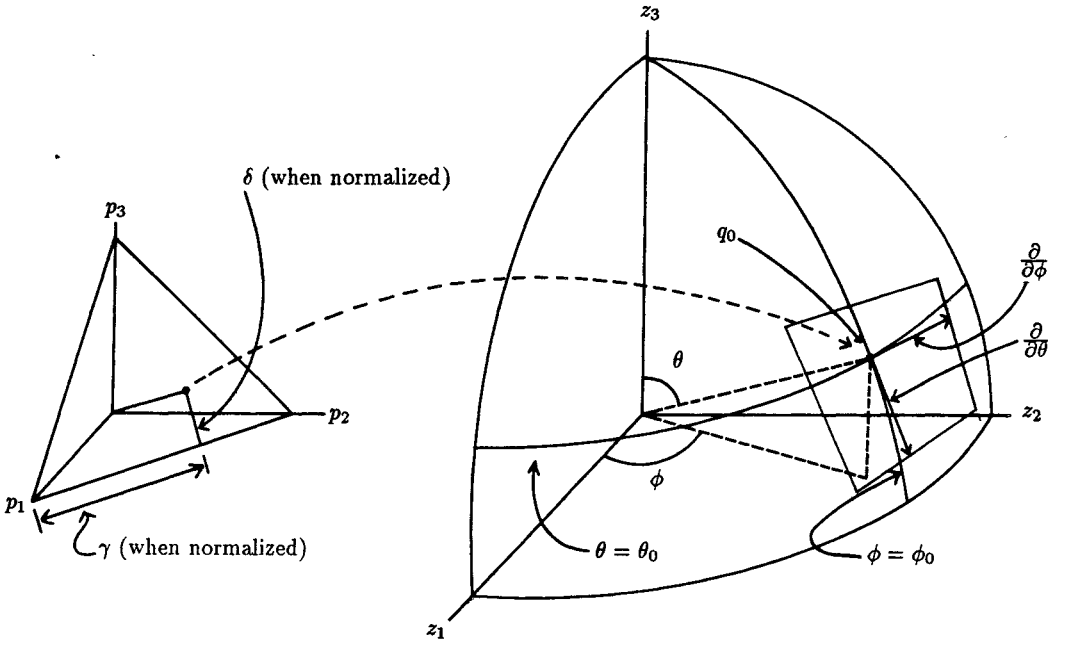

FIG. 1. *The transformation from* (p_1, p_2, p_3)*-space to* (z_1, z_2, z_3)*-space, with polar coordinates* $\theta = \arccos(p_3^{1/2})$, $\phi = \arcsin((p_2/(p_2 + p_2))^{1/2})$. *The vectors* $\partial/\partial\theta$ *and* $\partial/\partial\phi$ *are tangent to the coordinate curves defined by the polar coordinates* (θ, ϕ).

subfamily of a multinomial family is just a simple way of saying that it satisfies Birch's conditions (Birch, 1964; Bishop, Fienberg and Holland, 1975, page 510). In this section, I give a geometrical interpretation of distribution theory based on the imbedded subfamily assumption.

Let $\mathscr{M}$ denote the positive-orthant portion of the sphere (2.2), let $\mathscr{M}_0$ denote the subspace of $\mathscr{M}$ corresponding to $\mathscr{Q}_0$, and let $\hat{q}_n$ be the point in $\mathscr{M}$ corresponding to the MLE $\hat{Q} = \hat{Q}_n$ in $\mathscr{Q}$, based on a sample of size n. If we define the Kullback–Leibler minimum

$$K(\hat{Q}, \mathscr{Q}_0) = \min_{Q \in \mathscr{Q}_0} K(\hat{Q}, Q)$$

(assuming the minimum exists), the likelihood ratio test statistic for testing $\mathscr{Q}_0$ against $\mathscr{Q}$ is $2n \cdot K(\hat{Q}, \mathscr{Q}_0)$. If we likewise define

$$d(\hat{Q}, \mathscr{Q}_0) = \min_{Q \in \mathscr{Q}_0} d(\hat{Q}, Q)$$

where $d(Q, Q^*)$ is the information distance of (2.7) and (2.8), then it follows from (2.13) that

$$2n \cdot K(\hat{Q}, \mathscr{Q}_0) = n \cdot d(\hat{Q}, \mathscr{Q}_0)^2 + o_p(1). \tag{2.15}$$

Equation (2.15) provides a geometrical understanding of the asymptotic chi-squared distribution of the likelihood ratio test (and, using (2.10), Pearson's chi-squared statistic, as well). In the theory of linear models, the squared length of the residual from an orthogonal projection of a spherical Normal variable onto a linear subspace has a chi-squared distribution. Here, letting q_0 in $\mathscr{M}$ be the point corresponding to the true distribution Q_0 in $\mathscr{Q}_0$, we may use the tangent plane V at q_0 to linearly approximate the sphere $\mathscr{M}$, and a line V_0 in that plane to approximate $\mathscr{M}_0$ (see Figure 2). When $\hat{q}_n$ is mapped to the tangent plane (by a mapping described below) its image Y_n is asymptotically spherically Normal, and

$$n \cdot \| (I - P_{V_0}) Y_n \|^2 \xrightarrow{\mathscr{L}} \chi_1^2 \tag{2.16}$$

where P_{V_0} is the orthogonal projection onto V_0 and I is the identity. But

$$d(\hat{Q}, \mathscr{Q}_0)^2 = \| (I - P_{V_0}) Y_n \|^2 + o_p\left(\frac{1}{n}\right) \tag{2.17}$$

which, together with (2.16), implies that $n \cdot d(\hat{Q}, \mathscr{Q}_0)^2$

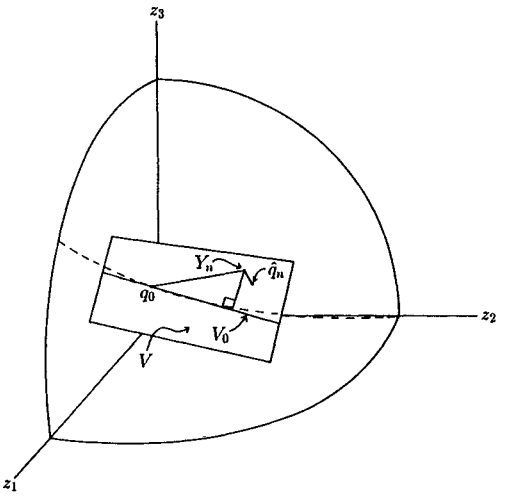

FIG. 2. *The subspace* V_0 *of the tangent plane* V *at* q_0. Y_n *is the image of* $\hat{q}_n$ *under the mapping* F *defined by* (2.18).

is asymptotically chi-squared and, therefore, by (2.15), so is $2n \cdot K(\hat{Q}, \mathscr{E}_0)$.

Some details. I now fill in a few of the details leading to (2.16) and (2.17). For simplicity, assume $\mathscr{E}_0$ is the symmetry model. Each point q on $\mathscr{M}$ may be connected to q_0 by an arc of a great circle, which has a unit tangent vector $\hat{v}$ at q_0 in the direction of q. Thus, q may be identified with a vector $v = d(q_0, q) \cdot \hat{v}$ in the tangent space V to $\mathscr{M}$ at q_0. This identification defines a mapping $F: \mathscr{M} \to V$ according to

$$F(q) = d(q_0, q) \cdot \hat{v}. \tag{2.18}$$

Now let $\{v_1, v_2\}$ be an orthonormal basis for V. The pair (θ_1, θ_2) defined by

$$F(q) = \theta_1 v_1 + \theta_2 v_2 \tag{2.19}$$

is a parameterization of $\mathscr{E}$ and the coordinate tangent vectors are v_1 and v_2. Since v_1 and v_2 are orthonormal, it follows from (2.6) that the parameterization defined by (2.19) has the identity as its information matrix, and, therefore, the limiting Normal distribution of its MLE, based on repeated sampling, is spherical.

Now, when $\mathscr{E}_0$ is the symmetry model, the image V_0 of $\mathscr{M}_0$ under F becomes a linear subspace of V. Letting $Y_n = \hat{\theta}_n$ be the MLE for the parameterization of (2.19), (2.16) follows immediately. Once it is shown that

$$\| (I - P_{V_0})v \|^2 = d(q, \mathscr{E}_0)^2 + o(d(q_0, q)^2) \tag{2.20}$$

as $d(q_0, q) \to 0$ where $v = d(q_0, q) \cdot \hat{v}$, (2.17) will follow. Equation (2.20) may be derived using

$$\lim \frac{\| F(q^*) - F(q) \|}{d(q^*, q)} = 1$$

as $(q^*, q) \to (q_0, q_0)$. This is a very general result, but here it follows from (2.8) (e.g., using (2.9)).

Although the symmetry model is rather special, the argument will go through with minor modifications for any imbedded subfamily (Kass, 1980). Of course, in the general setting in which $\mathscr{E}_0$ is an imbedded subfamily of a multinomial family $\mathscr{E}$, the degrees of freedom are $\dim(\mathscr{E}) - \dim(\mathscr{E}_0)$.

The condition that $\mathscr{E}_0$ be an imbedded subfamily of $\mathscr{E}$ is important. In the definition given in Section 1.2.3, it was noted that without the continuity condition (2) $\mathscr{E}_0$ becomes an immersed but not imbedded subfamily. In this case, the MLE may be inconsistent and the various asymptotically equivalent goodness-of-fit statistics no longer have limiting chi-squared distributions for all elements of the null hypothetical family $\mathscr{E}_0$. It is easy to construct families that exhibit MLE inconsistency using the asymptotic spherical Normality of $Y_n = \hat{\theta}_n$ for the parameterization (2.19); we may take any smooth rank-one curve in R^2 that doubles back on itself at the origin as in Figure 3, and then consider (θ_1, θ_2) to be the coordinates (possibly after shrinking the curve toward the origin so that it will fit within $\Theta \subseteq R^2$) and use F^{-1} (according to (2.18), with q_0 being arbitrary) to define $\mathscr{E}_0$. Figure 3 then becomes a picture of the tangent space at q_0, with $\hat{\theta}_n$ being asymptotically spherically normal. For any $\hat{\theta}_n$ within the dotted lines, $\hat{\beta}_n$ will lie on the "tail" portion of $\mathscr{E}_0$, and not near β_0. This occurs with positive probability for all n, so that $\hat{\beta}_n$ will not converge in probability to β_0. (The topological nature of consistency of the MLE is quite general, e.g., Bahadur, 1971, Section 9.) Similarly, the limiting distribution of $n \cdot d(\hat{Q}, \mathscr{E}_0)^2$ will be based on the squared length of a projection onto a space consisting of the union of a ray through the origin and a line through the origin, rather than the line alone.

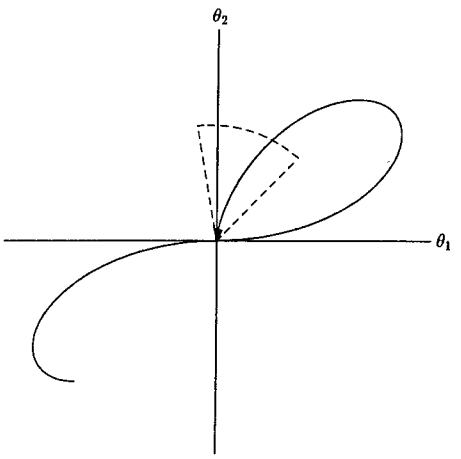

FIG. 3. *A smooth curve that doubles back on itself at the origin:* $(\theta_1(t), \theta_2(t)) = (sin(t) \cdot (cos(t) + 1), sign(t) \cdot sin^2(t))$ *for* t *in* $(-\pi/2, \pi)$, *with the arrow indicating that the curve approaches the origin as* t *approaches* π.

2.1.6 Inference Regions Based on Information Distance

The information distance can also be used to define an inference region as the set of all points in $\mathscr{E}$ within a distance δ of the maximum likelihood point $\hat{Q}$. By inverting the test of H_0: $Q = Q_0$ based on the criterion $n \cdot d(\hat{Q}, Q_0)^2$ and using the asymptotic chi-squared distribution of $n \cdot d(\hat{Q}, Q_0)^2$, on two degrees of freedom (which follows from the argument in Section 2.1.5), we find the asymptotic coverage probability of such a region to be $1 - \alpha$ when δ is chosen to satisfy $P\{\chi_2^2 > \delta^2\} = \alpha$. It is also an approximate highest posterior density region on the sphere (2.2), and its boundary becomes an approximate likelihood contour when the likelihood is defined on that sphere. Equations defining this region in terms of p_1, p_2, and p_3 are easily derived.

2.2 Parametric Families as Riemannian Manifolds

Most of the geometrical relationships in Section 2.2 hold for arbitrary regular parametric families. In this

section, I will briefly indicate the generalizations by describing some of the basic elements of Riemannian geometry, while showing how families of densities are given the structure of Riemannian manifolds. Those seeking to learn the mathematics may consult Spivak (1979), volumes I and II. Another recommended exposition is Stoker (1969), which is written in the classical, i.e., nonabstract, spirit and devotes most of its attention to the differential geometry of surfaces in three-dimensional Euclidean space. Boothby (1975) gives a very straightforward treatment of the relevant material, while Kobayashi and Nomizu (1969) go somewhat deeper (as do the subsequent volumes by Spivak).

For emphasis, I first summarize the most important aspects of the description I am about to give. A smooth manifold may be intuited as an abstract surface on which functions are defined, with calculations made in terms of some coordinate system on the manifold. It is important, however, in geometry and also in geometrical developments in statistics, that the choice of coordinates plays no essential role in the theory. In statistics, parametric families of densities become manifolds and parameterizations become coordinate systems. A Riemannian metric is a collection of inner products, one defined at each point of a manifold on the tangent space to the manifold at that point. The information metric is the collection of inner products defined by the Fisher information matrix, which may vary continuously from point to point. A Riemannian manifold is a manifold on which a Riemannian metric is defined.

The term "surface" usually refers to a smooth m-dimensional subspace of R^n, described by n-m equations in the coordinate variables $x_1, x_2, \cdots, x_n$. We may imagine the surface sitting within the surrounding or "ambient" space R^n, just as a two-dimensional surface sits within R^3. The space R^n and the coordinates $x_1, x_2, \cdots, x_n$ are not part of the surface itself, but are external to it, and properties that are characterized in terms of the ambient space and its coordinates are called extrinsic. Modern differential geometry, since Gauss and Riemann, has emphasized instead the intrinsic properties of surfaces, which do not depend on extrinsic coordinate expressions. In studying surfaces as spaces on which functions are defined and analyzed directly, without reference to the ambient space, a first step is to introduce coordinate systems of the proper dimension. For instance, the two-dimensional unit sphere S^2 centered at the origin of R^3 has a positive portion S^2_+ that lies in the interior of the first octant; this surface may be identified using any of the coordinate pairs (x, y), (x, z), (y, z), (θ, ϕ) (spherical), or (r, ϕ) (cylindrical) (Figure 4).

A coordinate mapping is a homeomorphism (a continuous, one-to-one mapping with a continuous inverse) of an open subset of the surface onto an open subset of R^m, for some m. The mapping (θ, ϕ) maps all of S^2_+ onto $(0, \pi/2) \times (0, \pi/2)$. The term "coordinate system" refers to the pair comprised of a coordinate mapping and its domain. When the domain is the entire surface, the coordinate system is called global. The sphere S^2 is compact and the homeomorphic image of a compact set is compact. Therefore, there does not exist a global coordinate system on S^2. The coordinate mapping (θ, ϕ) is global on S^2_+ but must be defined locally on S^2. Note that the mapping (x, y, z) which gives S^2_+ its familiar algebraic characterization, $x^2 + y^2 + z^2 = 1$, is not a coordinate mapping because its image is not an open subset of R^3. The surface S^2_+ is intrinsically two-dimensional; this is expressed by the openness in R^2 of the image spaces of the various coordinate mappings, (x, y), (θ, ϕ), etc. For a property to be intrinsic it must be expressible indifferently with respect to the choice among the infinitely many possible coordinate systems. This consideration leads to the construction of manifolds and their substitution for surfaces as the main object of study.

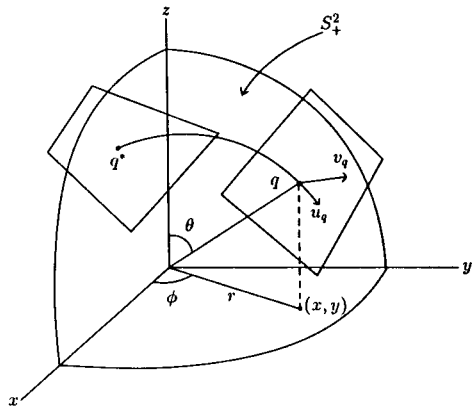

FIG. 4. *The first-octant portion of the unit sphere, and its tangent spaces at points q and q*. Labels identifying alternative coordinate systems (x, y), (θ, ϕ), and (r, ϕ) are shown; u_q and v_q are tangent vectors at q, with u_q being tangent to the great circle connecting q* and q.*

The basic requirement of a topological manifold is that each of its points must have a neighborhood that is homeomorphic to R^m, for a fixed m which becomes the dimension of the manifold. In order to exclude pathologies, topological manifolds are also required to be second-countable Hausdorff spaces. Differentiation is carried out in terms of coordinates, and coordinate systems could be introduced on topological manifolds, but without additional structure the definition of differentiability could depend on the coordinate system used to define it. To secure irrelevance of the choice of coordinate systems, ϕ and ψ are said to be compatible if both $\phi \circ \psi^{-1}$ and $\psi \circ \phi^{-1}$ are smooth (infinitely differentiable) on their respective domains, which are open subsets of R^m (or are empty if the domains of ϕ and ψ do not overlap). That is, ϕ and ψ are compatible if the transformation from ϕ-coordinates

to ψ-coordinates is a diffeomorphism (see also Section 1.2.2). A collection of compatible coordinate systems that cover a topological manifold is called a smooth structure (or an atlas) on the manifold. When all possible compatible coordinate systems are included, the collection is called a maximal smooth structure (a maximal atlas). Finally, a topological manifold together with a maximal smooth structure is a smooth manifold. (Terminology varies; for instance, smooth manifolds are not always required to be second-countable).

Since asymptotic statistical theory is predominantly local (e.g., concerning results that hold in a neighborhood of a true value of a parameter), the distinction between local and global parameterizations is rarely of interest. On the other hand, it is very important to be able to change parameters without affecting regularity conditions that involve differentiability. Thus, rather than making inferences on a particular parameter space, one might consider at once all possible compatible local parameterizations. *This is precisely what is done when a family of densities is structured as a smooth manifold.* (In fact, in one construction of a manifold, each point in the abstract space is the equivalence class of coordinate system image values—in statistics, each point would be the equivalence class of corresponding parameter values for all possible compatible local parameterizations.)

The manifold of densities. As already done implicitly in Section 1.2.2, let us assume that the family of distributions $\mathscr{P}$ may be considered, instead, a family of densities. Let us also suppose that the family of densities has the structure of an m-dimensional smooth manifold, and relabel it $\mathscr{E}$. In most cases, the manifold structure of $\mathscr{E}$ would be defined in terms of a global parameterization, which would induce a "usual" Euclidean topology. (Note, however, that location families on spheres do not have global parameterizations.) Beginning with any familiar family $\mathscr{P}$, it would be possible to introduce some strange topology for $\mathscr{E}$, but when common parameterizations become coordinate systems for $\mathscr{E}$, the topology of $\mathscr{E}$ becomes the weak topology.

Letting q denote a generic element of $\mathscr{E}$, the likelihood function may be defined on $\mathscr{E}$ according to

$$L_y\colon \mathscr{E} \to R,$$

$$L_y(q) = q(y).$$

Similarly, the loglikelihood function may be defined directly on $\mathscr{E}$, with no reference to a particular parameterization. A maximum likelihood estimate, if one exists, may also be considered a point $\hat{q}$ in $\mathscr{E}$ defined by

$$L_y(\hat{q}) = \max_{q \in \mathscr{E}} L_y(q).$$

The purpose of mentioning these definitions is to emphasize the possibility of introducing geometrical structure in a "coordinate-free" or parameterization-invariant manner.

Subfamilies as submanifolds. In Section 1.2.3, the imbedded submanifold assumption on null hypothetical subfamilies was discussed. There it was described in terms of a parameter space Θ, but we may avoid working with a particular parameter space simply by requiring the subfamily to form an imbedded submanifold $\mathscr{E}_0$ of $\mathscr{E}$. Although I have not specifically said how imbedded submanifolds are defined in general, the local product structure mentioned at the end of Section 1.2.3 continues to be a characterization in abstract settings.

The next construct is tangent vector. Tangent vectors may be described in several ways. The tangent vectors at a given point on a smooth manifold form the tangent space to the manifold at that point. In the case of a two-dimensional surface, such as S_+^2, imbedded in R^3, it is easy to "see" the tangent spaces (two are sketched in Figure 4): they are planes—importantly, two-dimensional vector spaces—tangent to the surface at some point. It is irrelevant, and even meaningless in the abstract constructions in which R^3 plays no role, that points in R^3 away from the surface may lie on more than one tangent plane. The primary concept is abstract but simple: at each point there is defined a vector space of the same dimension as the manifold. On R^m, each tangent space is canonically identified with R^m itself: the tangent space at $q \in R^m$ may be pictured as the space of vectors whose tails are anchored at q; we are able to characterize vectors in R^m in terms of their direction and length, ignoring their location, only because R^m is flat.

Intuitively, tangent vectors are tangent to curves on the manifold. This intuition is fully justified when the manifold is imbedded in R^n, for in that case the curves on the manifold are also curves in R^n to which the tangent vectors are tangent. One intrinsic formulation defines tangent vectors at a point q as equivalence classes of curves through q, the curves being equivalent if they have the same directional derivatives in their coordinate expressions. The simplest definition of tangent vectors characterizes them according to the formal properties of directional derivative operators. Thus, a tangent vector at q is a linear operator v_q on the space of real functions that are smooth at q such that v_q satisfies the Leibniz rule: if f and g are smooth at q, then $v_q(f \cdot g) = f \cdot v_q(g) + g \cdot v_q(f)$.

Tangent vectors are often written in the notation used above, e.g., v_q is a tangent vector at q, but then coordinate representations are also needed. A coordinate mapping, x, defined on an open set $U \subseteq \mathscr{E}$ with $x\colon U \to R^m$, may be written in component form: $x = (x^1, x^2, \cdots x^m)$. (Superscripts are standard; subscripts are reserved for other related objects.) Then, the

tangent space at a point $q \in U$ is spanned by the vectors that are tangent to the coordinate curves, the ith coordinate curve satisfying x^j = constant for all $j \neq i$. The ith x-coordinate basis tangent vector at the point q is written $\partial/\partial x^i|_q$. (See Figure 2.) The dependence on the point at which the tangent vector lies is often eliminated from the notation, leaving $\partial/\partial x^i$. This tangent vector is the directional derivative operator that takes the directional derivative along the ith coordinate curve; thus the partial derivative notation. When the coordinate system is specified by the context, the notation $\partial_i = \partial/\partial x^i$ is often used.

Vector fields on smooth manifolds are analogous to those on R^m; they are collections of tangent vectors v_q, one at each point q. A vector field X, made up of tangent vectors X_q, is smooth if the mapping $q \to X_q(f)$ is smooth whenever f is a smooth real function. In terms of a coordinate basis $\{\partial_i\}$ the vector field X has components a_i, i.e., $X_q = \Sigma\, a_i(q) \cdot \partial/\partial x^i|_q$, and for smooth vector fields these components are smooth functions. The space of smooth vector fields on $\mathscr{M}$ is denoted by $\mathscr{X}(\mathscr{M})$.

Now, if an inner product is defined on each tangent space, there being no necessary relationship among the inner products (although some smoothness is usually assumed), then the collection of these inner products is a Riemannian metric; it is denoted here by $\langle , \rangle = \{\langle , \rangle_q \mid q \in \mathscr{M}\}$, where $\langle , \rangle_q$ is an inner product on the tangent space at q. Together with $\langle , \rangle$, $\mathscr{M}$ then becomes a Riemannian manifold. The usual notation for the coordinate expression of a Riemannian metric is $g_{ij} = \langle \partial_i, \partial_j \rangle$.

On R^m, the usual Euclidean metric assigns to each tangent space the Euclidean inner product. On surfaces in R^m, there is a usual metric "inherited" from the Euclidean metric on the ambient R^m. For instance, the usual metric on S^2_+ is inherited from the Euclidean metric on R^3. In Figure 4, the vectors u_q and v_q in the tangent plane at q may be considered vectors in R^3 (that is, in the tangent space to R^3 at q, but, as noted earlier, the various relocations of R^m are not usually distinguished). As such they have a usual inner product, and their inner product as elements of the tangent space may be defined to be equal to this extrinsic, usual inner product.

More generally, but omitting details, smoothness and rank of mappings from one manifold to another are easily defined using arbitrary coordinate systems. From this, if $\mathscr{M}_0$ is an imbedded submanifold of $\mathscr{M}$, and $\mathscr{M}$ has a metric $\langle , \rangle$, then the inherited metric on $\mathscr{M}_0$ may be defined. In addition, the Riemannian metric structure-preserving mappings may be defined; these are called isometries, and when there is an isometry between two manifolds they are called isometric. The intrinsic geometries of isometric manifolds are identical. In terms of coordinate components, a diffeomorphism ϕ from one Riemannian manifold $\mathscr{M}^{(1)}$ to another $\mathscr{M}^{(2)}$ is an isometry when there exist coordinate systems such that $g^{(2)}_{ij}(\phi(q)) = g^{(1)}_{ij}(q)$, i.e., the matrices representing the metrics have the same form. These remarks will be used in deriving the spherical multinomial geometry from a Euclidean Poisson geometry later in this subsection; there, the matrices representing metrics will be Fisher information matrices.

The information metric. The *information metric* is easily defined in terms of any parameterization θ according to

$$\langle \partial_i, \partial_j \rangle = i(\theta)_{ij}$$

where $i(\theta)_{ij}$ is the (i, j)-component of the Fisher information matrix $i(\theta)_{ij} = E_\theta(\partial_i l_Y \partial_j l_Y)$. This serves as a definition as long as $E_\theta(\partial_i l_Y \partial_j l_Y)$ is finite for all θ and the matrix $i(\theta)$ is positive definite and continuous in θ. A coordinate-free definition may be written

$$\langle , \rangle : \mathscr{X}(\mathscr{M}) \times \mathscr{X}(\mathscr{M}) \to R,$$

$$\langle X_q, Z_q \rangle_q = E_q(X_q(l_Y) Z_q(l_Y))$$

under the analogous assumptions stated in terms of general tangent vector fields X and Z, that is, positive definiteness of the bilinear mappings $\langle , \rangle_q$ and continuity of the mapping $q \to \langle X_q, Z_q \rangle_q$. More than continuity is required for many purposes, geometrical as well as statistical. Thus, it is convenient to assume that the information metric is smooth. Standard conditions for asymptotic theory of maximum likelihood, Cramer's conditions or some variant (Lehmann, 1983) also may be easily formulated directly in terms of the manifold of densities, in a coordinate-free manner.

An important property of the information metric is its consistency under inheritance on a submanifold. That is, if $\mathscr{M}_0$ is an imbedded submanifold of a manifold of densities $\mathscr{M}$ on which the information metric is defined, then the information metric may also be defined on $\mathscr{M}_0$ and it is equal to the metric inherited on $\mathscr{M}_0$ from the information metric on $\mathscr{M}$. An illustrative application of this property is given next.

Multinomial geometry, via Poisson geometry. In Section 2.2, the sphere (2.2) was shown to represent the Fisher information metric geometry of the trinomial family according to (2.3) and (2.6). Here an alternative derivation of the spherical representation is given. Consider the manifold $\mathscr{M}$ of the joint densities of $m + 1$ independent Poisson random variables Y_i with Poisson parameters $\lambda_i > 0$, $i = 1, \cdots, m + 1$. If interest centers on the relative magnitudes of the λ_i,

$$q_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_{m+1}},$$

then we may reparameterize using $(q_1, \cdots, q_m, \beta)$, where $\beta = \sum_{i=1}^{m+1} \lambda_i$, and, in the terminology of Cox

and Hinkley (1974, page 35), the sufficient MLE $(\hat{q}_1, \cdots, \hat{q}_m, \hat{\beta})$ may be decomposed into the conditionally sufficient component $(\hat{q}_1, \cdots, \hat{q}_m)$ and the ancillary component $\hat{\beta} = \sum_{i=1}^{m+1} \hat{\lambda}_i$. Inference about $(q_1, \cdots, q_m)$ may proceed by conditioning on $n \equiv \sum_{i=1}^{m+1} y_i = \sum_{i=1}^{m+1} \hat{\lambda}_i$ to yield a joint multinomial distribution. Writing

$$\mathscr{S}^m \equiv \left\{ (q_1, q_2, \cdots, q_m) \;\middle|\; \sum_{i=1}^{m} q_i < 1; q_i > 0, i = 1, \cdots, m \right\}$$

and $q_{m+1} = 1 - \sum_{i=1}^{m} q_i$, let $\mathscr{E}_m(n)$ denote the manifold of Multinomial$(n; q_1, \cdots, q_m)$ densities, with $(q_1, q_2, \cdots, q_m) \in \mathscr{S}^m$. Returning now to the full $(m + 1)$-dimensional Poisson family, let $\mathscr{E}_0$ be the subspace defined by

$$\mathscr{E}_0 \equiv \left\{ \prod_{i=1}^{m+1} e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} \;\middle|\; \sum_{i=1}^{m+1} \lambda_i = n \right\}.$$

Using $(q_1, q_2, \cdots, q_m)$-coordinates $\mathscr{E}_0$ may be written

$$\mathscr{E}_0 = \left\{ \left(e^{-n} \prod_{i=1}^{m+1} \frac{n^{y_i}}{y_i!} \right) \cdot q_1^{y_1} q_2^{y_2} \cdots q_m^{y_m} q_{m+1}^{y_{m+1}} \mid (q_1, \cdots, q_m) \in \mathscr{S}^m \right\}.$$

Although the sample spaces of $\mathscr{E}_0$ and $\mathscr{E}_m(n)$ are different, the information matrices in (q_1, q_m)-coordinates are identical, so the two spaces, with the information metrics, are isometric. That is, the information-metric geometries of $\mathscr{E}_0$ and $\mathscr{E}_m(n)$ are the same. Meanwhile, from the inheritance property of the information metric, the information geometry of $\mathscr{E}_0$ is the geometry it inherits from the information geometry of $\mathscr{E}$. The latter geometry is Euclidean: the coordinates defined by the Poisson "variance-stabilizing" transformation

$$\xi_i \equiv \sqrt{4\lambda_i}, \quad i = 1, \cdots, m + 1,$$

are Euclidean, since the information matrix becomes the identity. But in these coordinates $\mathscr{E}_0$ is the set $\xi_1^2 + \xi_2^2 + \cdots + \xi_{m+1}^2 = 4n$, $\xi_i > 0$, $i = 1, \cdots, m + 1$. Therefore, $\mathscr{E}_0$ is isometric to the positive-orthant portion of the m-dimensional sphere of radius $R = \sqrt{4n}$. Thus, the information metric geometry of $\mathscr{E}_m(n)$ is that of the positive-orthant portion of the m-dimensional sphere of radius $R = \sqrt{4n}$ (which is described by (2.2) where $n = 1$ and $m = 2$).

There is a connection between the uses of "metric" in "Riemannian metric" and "topological metric space." At each point of a parameterized curve on a manifold there is a tangent vector, "tangent to the curve," in the tangent space. A Riemannian metric gives this tangent vector a length and when length is integrated along the curve the result is called the arclength of the curve. Given any two points on a connected manifold, the arclength of the shortest curve connecting them defines a distance between them. With this distance, the manifold becomes a metric space and, it turns out, the metric space topology coincides with its original topology. Meanwhile, the shortest curve connecting two points, when parameterized by arclength, is called a geodesic.

The information distance. The *information distance* is the distance function determined by the information metric. An explicit expression for it was given in the multinomial case in Section 2.1.1. The squared information distance $d(\hat{q}, \mathscr{E}_0)^2$ of an MLE $\hat{q}$ in a manifold of densities $\mathscr{E}$ away from a null hypothetical imbedded subfamily $\mathscr{E}_0$ becomes a general goodness-of-fit statistic; its asymptotic chi-square distribution may be derived essentially according to the argument of Section 2.1.5, with the mapping F of (2.18) becoming, more generally, the inverse of what is called the "exponential mapping." The terminology makes sense when the manifold is a matrix group such as the orthogonal group: the exponential map on the tangent space of "infinitesimal elements of the group," e.g., on the space of infinitesimal orthogonal matrices, the anti-symmetric matrices, may be written in the familiar matrix Taylor series form. The set of densities in $\mathscr{E}$ within a distance $d = \delta$ of the maximum likelihood point $\hat{q}$ is, in general, an inference region having asymptotic coverage probability $1 - \alpha$ when δ is chosen to satisfy $P\{\chi_m^2 > \delta^2\} = \alpha$. It also has approximate posterior probability $1 - \alpha$.

Corresponding to each Riemannian metric on a smooth manifold there is a uniquely defined "infinitesimal" element of volume on each tangent space. Taken together, these define what is called the natural volume element (with respect to the given metric) on the manifold. In the case of the usual Euclidean metric on R^m, we usually write the natural volume element in terms of rectangular coordinates: $dV = dx_1 dx_2 \cdots dx_m$. When $m = 2$, in polar coordinates we write: $dV = rdrd\theta$. On S_+^2 with the usual inherited metric, in spherical coordinates we write: $dV = \sin(\theta) d\theta d\phi$. As is true in general for surfaces in R^m having the inherited metric, the natural volume element on S_+^2 corresponds to the familiar element of surface area of advanced calculus. The general form of the natural volume element is

$$dV = \det(G(\mathbf{z}))^{1/2} \cdot dz_1 dz_2 \cdots dz_m$$

where $G(\mathbf{z})$ is the $\mathbf{z}$-coordinate matrix representation of the given metric. The natural volume element generates a measure on the manifold, which may be considered "uniform" with respect to the metric (see Section 2.3.1 for discussion of the implications of this).

Jeffreys' rule. The measure determined by the information metric is, of course, that determined by Jeffreys' general rule.

Orthogonal parameters. Whenever the (i, j)-component of the information matrix $i(\theta)$ is zero, ∂_i and ∂_j are orthogonal, and θ_i and θ_j are called orthogonal parameters. Usuallly this term refers to global orthogonality, i.e., $i(\theta)_{ij} = 0$ for all θ, though in some cases it can be useful to consider parameters that are orthogonal only on restricted regions.

It is interesting that the information metric geometry of the most basic family of distributions, the multinomial family, is the simplest of all non-Euclidean geometries. It may also be shown that the information metric geometries of location-scale families constitute the next simple class of non-Euclidean geometries, the hyperbolic spaces, which have constant negative curvature. According to Gauss' "Theorema Egregium" and its generalization, this curvature (Gaussian curvature and, more generally, Riemannian scalar curvature) is intrinsic: it is determined solely by the metric. As a consequence, a space of non-zero curvature is not isometric to Euclidean space, and so cannot have its metric be represented by a constant matrix in any coordinate system. Thus, for a family of densities that has non-zero curvature with respect to the information metric, there does not exist a parameterization in which the information matrix is constant. That is, for such a family there does not exist a "covariance stabilizing transformation." (As noted by Reeds, 1975, this remark solves a problem posed by Holland, 1973.)

2.3 Elaboration of Jeffreys' Methods

2.3.1 Reference Priors: Uniformity, Symmetry and Jeffreys' Rules

As reviewed in Section 2.2 (and Section 2.1.3), the prior of Jeffreys' general rule $\pi(\theta) \propto \det(i(\theta))^{1/2}$ is generated by the natural volume element of the information metric. I now briefly consider this scheme as a general method of getting priors, comparing measures determined by metrics of Lebesgue measure on R^m.

The motivation here is that the uniform measure on some parameter space (i.e., Lebesgue measure) seems, at first glance, to be a good choice for a reference prior (that is, for a prior chosen according to a formal rule, without detailed consideration of the context). It is objectionable, however, because a prior that is uniform on one parameter space is not uniform on others. Jeffreys' rule, on the other hand, is invariant while also being uniform in a meaningful sense, e.g., on the sphere (2.2). I suggest that this kind of uniformity is what makes Lebesgue measure seem appealing, and the fact that measures determined by metrics may be considered uniform provides heuristic motivation for using them as reference priors.

It might be argued that the appeal of Lebesgue measure comes from its translation invariance, which is to say from symmetry. Within Riemannian geometry, however, uniformity is more general than invariance under transformations that respect symmetry. Uniformity and translation invariance coincide on R^m, but do not necessarily coincide in general: there is an intuition behind the meaning of a uniform distribution on a surface that exhibits no symmetry. One would interpret, it seems to me, a uniform application of paint on the surface of an irregularly shaped object, as one in which equal amounts of paint were applied to regions of equal surface area. Uniform distributions on surfaces are special cases of measures that apply equal mass to sets of equal volume, that is, of measures that are determined by Riemannian metrics.

Furthermore, when symmetry is present, the metric may be required to respect it. If, for instance, the space in question has differentiable group structure, i.e., is a Lie group, then the symmetry-preserving metric determines (left) invariant measures, that is, Haar measures. This means that familiar uses of symmetry in determining reference priors (see Villegas, 1971, 1977) may be incorporated into the geometrical rule of defining the prior in terms of the volume element. Thus, the concepts of uniformity and symmetry may be distinguished within Riemannian geometry; uniformity is an attribute of all measures determined by metrics, while invariance under transformation by elements of a group is an attribute of measures determined by metrics that respect symmetry.

When the manifold of densities is a Lie group, the information metric is, in fact, left invariant, so that Jeffreys' general rule determines left Haar measure. In treating problems having location parameters, Jeffreys modified this rule (1946, pages 458–459; 1961, pages 182–183): if μ is a location parameter and α is some other (possibly multidimensional) parameter, Jeffreys' modified rule becomes $\pi(\mu, \alpha) = |(g_{ik})|^{1/2}$ where (g_{ik}) is the information matrix based on α alone. (The modified rule also applies when the location parameter is multidimensional.) In the special case of the Normal(μ, σ^2) family, this entails replacing $\pi(\mu, \sigma) = \det(i(\mu, \sigma))^{1/2} = \sigma^{-2}$ with $\pi(\mu, \sigma) = i(\sigma)^{-1/2} = \sigma^{-1}$. His argument, as I understand it, is simply that location parameters should be considered separately. In terms of the manifold of densities $\mathscr{E}$, there is a factorization $\mathscr{E} = \mathscr{E}_1 \times \mathscr{E}_2$ where $\mathscr{E}_1$ is identified with the location group G (using a smooth isomorphism so that $\mathscr{E}_1$, is "essentially identical" to G). Then the appropriate metric on the first factor, $\mathscr{E}_1$, is the one that preserves the symmetry, i.e., the invariant metric. If the metric on the second factor $\mathscr{E}_2$ is the information metric, then the volume element on $\mathscr{E} = \mathscr{E}_1 \times \mathscr{E}_2$ is a product volume element and the measure determined by the product metric is a product measure: the measure on the first factor being a Haar

measure, and the measure on the second factor being that generated by information metric. Thus, by taking the volume element determined by the product metric we arrive at Jeffreys' modified rule. Note that μ and α are a priori independent as a consequence of the product form of the metric. The point is that geometry accommodates the analysis and, although Jeffreys did introduce a new element, symmetry, into his argument, there is a similarity of method in the general rule and its modification in that both determine the prior from a metric.

The purpose of these remarks is to indicate that the method of determining reference priors from metrics is intuitive and can generate prior densities other than the square-root of the information determinant. The only suitable metrics I know of, however, are those just discussed, producing Jeffreys' rule and its modification. It would be interesting to find others. Whether a reference prior determined by a metric is appealing depends on whether the metric is appealing. The information metric is the seemingly natural choice, from the point of view of asymptotics. In practice, of course, one must always consider the possible distortion in representing knowledge by a reference prior, and the extent to which this distortion affects inferences. (See Kass, Tierney and Kadane (1989) for discussion of asymptotic methods for assessing the sensitivity of inferences to the choice of prior, and Kass (1988) for some additional discussion of invariance, Jeffreys' rules, and data-translated likelihood.)

2.3.2 Odds Factors and Orthogonal Parameters

Jeffreys introduced the term "orthogonal parameters" to refer to a parameterization in which the information matrix $i(\theta)$ is diagonal. The geometrical interpretation was given in Section 2.1.4. Jeffreys' chief application was in his theory of hypothesis tests (which he called "significance tests") based on what is now often known as odds factors. The "odds factor" or "Bayes factor" is the ratio of the posterior odds of a null hypothesis H_0 (versus an alternative H_A) to the prior odds of H_0: schematically, we have

$$\frac{P\{H_0 \mid \text{data}\}}{P\{H_A \mid \text{data}\}} = \frac{P\{\text{data} \mid H_0\}}{P\{\text{data} \mid H_A\}} \cdot \frac{P\{H_0\}}{P\{H_A\}}$$

where $P\{H_0\} = 1 - P\{H_A\}$ is the prior probability of H_0, and the odds factor is

$$K = P\{\text{data} \mid H_0\}/P\{\text{data} \mid H_A\}.$$

Suppose we have a parameterization (β, ψ) and wish to test H_0: $\psi = \psi_0$ vs. H_A: $\psi \neq \psi_0$, using a prior $\pi(\beta)$ under H_0 and $\pi(\beta)\pi(\psi \mid \beta)$ under H_A. The odds factor becomes

$$K = \frac{\int p(y \mid \beta, \psi_0)\pi(\beta)\, d\beta}{\iint p(y \mid \beta, \psi)\pi(\beta)\pi(\psi \mid \beta)\, d\beta\, d\psi}.$$

For the case in which the parameter spaces B and Ψ are one-dimensional, Jeffreys provided an asymptotic approximation to K, noted an important simplification that occurs when β and ψ are orthogonal, and recommended a particular Cauchy reference prior.

Jeffreys' results may be derived, sharpened, and generalized using Laplace's method. (For general discussion and references on related uses of Laplace's method in statistics see Reid (1988) and Kass, Tierney and Kadane (1988).) Writing $p(y \mid \beta, \psi) = \exp[l(\beta, \psi)]$ and $l^*(\beta) = l(\beta, \psi_0)$, and letting $(\hat{\beta}, \hat{\psi})$ and β^* be the maxima of l and l^*,

$$K = \frac{\det(\Sigma(\psi_0))^{1/2}\exp[l(\beta^*, \psi_0)]\pi(\beta^*)}{(2\pi)^{1/2}\det(\Sigma)^{1/2}\exp[l(\hat{\beta}, \hat{\psi})]\pi(\hat{\beta})\pi(\hat{\psi} \mid \hat{\beta})} \cdot \{1 + O(n^{-1})\} \tag{2.21}$$

where

$$\Sigma(\psi_0) = (-D^2 l^*[\beta^*])^{-1} \text{ and } \Sigma = (-D^2 l[(\hat{\beta}, \hat{\psi})])^{-1}.$$

Now suppose that β and ψ are one-dimensional orthogonal parameters, and assume that the "true" value of ψ, is either ψ_0 or a neighboring alternative satisfying $\psi - \psi_0 = O(n^{-1/2})$, so that $\psi_0 - \hat{\psi} = O(n^{-1/2})$. By expanding $\partial_\beta l(\beta^*, \psi_0)$ about $(\hat{\beta}, \hat{\psi})$,

$$\beta^* - \hat{\beta} = O(n^{-1}) \tag{2.22}$$

and using this in (2.21) yields

$$\hat{K} = (2\pi)^{-1/2}\sigma \det(\Sigma)^{-1/2} \cdot \exp[l(\beta^*, \psi_0) - l(\hat{\beta}, \hat{\psi})] \cdot \pi(\hat{\psi} \mid \hat{\beta})^{-1} \tag{2.23}$$

with

$$K = \hat{K}\{1 + O(n^{-1})\}$$

where $\sigma = \Sigma(\psi_0)^{1/2}$.

When β and ψ are independent under the prior, expression (2.23) is a substantial simplification of (2.21): the prior will then enter the approximation to the odds factor only through $\pi(\hat{\psi}) = \pi(\hat{\psi} \mid \hat{\beta})$. Thus, to order $O(n^{-1})$, when the parameters are orthogonal and a priori independent, only the parameter quantifying departure from the null hypothesis need be contemplated. In fact, the parameters need not be globally orthogonal, but merely orthogonal at (β, ψ_0) for all β. (Such parameters may be called "null-orthogonal.") This is enough to ensure (2.22).

There remains the problem of determining the prior. An appealing possibility is to put a prior on the distance from the null hypothetical submodel to a particular alternative (β, ψ). In multidimensional problems, one could use a spherically symmetric distribution on the slices of a tube around the model, with the spheres being defined by the information metric. Interestingly, in the two-parameter case he treated, this is essentially what Jeffreys proposed. Jeffreys did not use (2.23), but instead made a further

approximation, replacing σ and Σ with their expected information counterparts, and $l(\theta^*, \psi_0) - l(\hat{\theta}, \hat{\psi})$ with its corresponding quadratic approximation (expected information again replacing observed). This produced an expression accurate only to order $O(n^{-1/2})$. To this order, however, it may be shown that the prior Jeffreys used (1961, page 275) was

$$\pi(\delta \mid \beta) = \frac{1}{\pi} \cdot \frac{1}{1+\delta^2}$$

where $\delta = \text{sign}(\psi) \cdot d((\psi_0, \beta), (\psi, \beta))$ is the signed information distance away from the null hypothetical submodel.

Except in special cases, this information distance is somewhat difficult to compute. A more practical procedure might be to use an elliptically contoured distribution on ψ in (2.23), with the ellipses defined by Fisher information $i(\beta, \psi)_{\psi\psi}$. Doing so, we would again achieve the interpretation that a prior was being put, approximately, on the distance from the model (with an accuracy of order $O(n^{-1/2})$). I intend to provide further references and details of these matters elsewhere. Cox and Reid (1987) have also used orthogonal parameters to define a conditional profile likelihood. See their paper for additional references, as well.

2.4 Bibliographical Notes

With the exception of Section 2.3.2, the material in this section is based on Kass (1980) and an unpublished 1981 technical report, which had the same title as the present paper. Most of the results were also obtained by other authors. As noted in the introduction, the other investigations of Fisher information Riemannian geometry were motivated by the paper by Rao (1945; see also Bhattacharyya 1943, 1946). The spherical Multinomial and hyperbolic Normal geometries were discovered or re-discovered by various authors (including Atkinson and Mitchell (1981), while Amari (1985) references an unpublished report in 1959), and several papers have described the information metric geometry for other families (Yoshizawa, 1971; Atkinson and Mitchell, 1981; Mitchell and Krzanowski, 1985; Oller and Cuadras, 1985; Skovgaard, 1984). A version of the relationship between information distance and Pearson's chi-squared statistic was given by Bhattacharyya (1946), and a geometrical derivation of the limiting chi-squared distribution of certain goodness-of-fit statistics appeared in Dudley (1979). The use of alternatives to the Kullback–Leibler number in generating alternative metrics was mentioned in Good (1969), and geometries of other entropy-like measures were explored in some detail by Burbea and Rao (1982a, b; see also Burbea, 1986). See Rao (1987) for additional references and discussion of applications to the study of genetic diversity. Stein (1965) also introduced the information metric in his study of admissibility of Bayes estimates based on improper priors, but I do not know of any related subsequent use of it. Finally, I note that the multinomial geometry of Section 2.2 may be considered a simplified version of infinite-dimensional geometry (see Dawid, 1977), which is used in nonparametric statistics (see, e.g., Beran, (1977) and Pfanzagl, (1982, especially Chapter 6) for comparison).

3. GEOMETRY OF INFORMATION LOSS AND RECOVERY

This section reviews the interpretation of Fisher's theory of estimation within a restricted context, that of iid observations from one-parameter subfamilies of exponential families, which Efron (1975) called *curved exponential families*. These families are of interest partly because their regularity ensures validity of many formal manipulations, and also because in this setting geometrical understanding of the estimation process becomes more readily apparent: an estimate may be considered a kind of projection of the exponential family observation onto the subfamily, in a manner roughly analogous to that of a projection of a data point onto a regression subspace in the theory of linear models. The benefit of such analyses is not only that results established apply to interesting examples of curved exponential families but, more importantly, the families might be considered archetypes for parametric inference.

One way that curved exponential families could serve as archetypes is by including an important special class, the subfamilies of multinomial distributions. Multinomials themselves could be considered archetypical in the sense that they can represent any continuous distribution through discretization, that is, by dividing the sample space into disjoint regions and taking the multinomial probabilities to be the probabilities assigned to the regions by the continuous distribution. This was routinely used by Fisher (and by Jeffreys) without comment. That is, he often performed calculations for subfamilies of multinomials and then applied them to continuous distributions.

In addition to the multinomial representation of general distributions, it has been shown by Amari (1987a), following a suggestion of Efron (1975), that there is a specific sense in which curved exponential families are local approximations to general families and that the geometrical analysis may be carried out using the local approximation method. Furthermore, one would hope that the insights gained from analysis of this special class might be of wider use, as well.

Curved exponential families will be used here to interpret Fisher's concepts of efficiency, information loss, and information recovery, and the claims he made about them.

3.1 Fisher's Principles

Fisher articulated his theory primarily in three papers published in 1922, 1925, and 1934, and in his book *Statistical Methods and Scientific Inference* (1956). Fisher's starting point is concisely presented in his statement, "Briefly, and in its most concrete form, the object of statistical methods is the reduction of data" (1922, page 300). Although this phrase may sound unobjectionable, it puts him on a path that will diverge from that taken by decision theory; decision theory considers crucial the purpose to which the data or their reduction are to be put. In particular, Fisher's concept of estimation is different from "point estimation" in that it involves data summary rather than optimal decision-making. (See Efron (1982) for some discussion of this.) As a matter of emphasis, the amount of "information" that a statistic can "summarize" is primary, and sufficiency, efficiency and ancillarity should be recognized as relating to information summary, as well as being understood from the more technical definitions.

3.1.1 Fisher's Measure of Information

Suppose θ is one-dimensional, and let $i^Y(\theta) = E_\theta((D_\theta l_Y(\theta))^2)$ denote the Fisher information in the sample Y. For a statistic T, let $i^T(\theta) = E_\theta((D_\theta l_T(\theta))^2)$ denote the information supplied by T, where $l_t(\theta) = \log(p(t \mid \theta))$, $p(t \mid \theta)$ being the density of the statistic T. Letting T and A be statistics derived from Y, the properties of Fisher information are as follows.

(i) $0 \le i^T(\theta) \le i^Y(\theta)$.
(ii) $i^T(\theta) = i^Y(\theta)$ for all θ if and only if T is sufficient.
(iii) $i^A(\theta) = 0$ for all θ if and only if A is distribution-constant (that is, its distribution does not depend on θ).
(iv) $i^{(T,A)}(\theta) = i^T(\theta) + i^A(\theta)$ if T and A are independent.

Writing $i^{T|A}(\theta) = E_\theta((D_\theta \log(p(T \mid A, \theta)))^2)$, with the expectation taken with respect to Y, (iv) may be strengthened to

(v) $i^{(T,A)}(\theta) = i^{T|A}(\theta) + i^A(\theta)$.

These elementary properties lead immediately to concepts of information loss and recovery. In the first place, the information lost in using an insufficient estimator T in place of the sample Y may be quantified by $i^T(\theta)/i^Y(\theta)$ or $i^Y(\theta) - i^T(\theta)$. Secondly, if (T, A) is sufficient and A is distribution-constant then $i^{T|A}(\theta) = i^Y(\theta)$. Thus, the information lost by an insufficient estimator T can be recovered by conditioning on an appropriate statistic A: if (T, A) is sufficient and A is distribution-constant, then the conditional distribution of T given A supplies all of the information in the sample.

It is plausible that these properties might characterize Fisher information, but I do not know of any results on this. Properties (i)–(iv) were specifically mentioned by Fisher (1935, page 47). Property (v) was used implicitly. Generalization to the multiparameter case is immediate.

3.1.2 Information Loss and Recovery

The information lost by an estimator could be quantified in terms of $i^T(\theta)/ni(\theta)$ or by $ni(\theta) - i^T(\theta)$, where $i(\theta)$ is the information per observation in a sample of size n, but it is analytically easier to use the limiting values of these quantities. The limiting value of the ratio form is 1 for efficient estimators. The limiting difference then provides a way of distinguishing among these estimators. Fisher defined the *information loss* of T as $\lim ni(\theta) - i^T(\theta)$. Note that this quantity will typically be infinite for inefficient estimators, which leads back to consideration of the ratio in that case. Fisher introduced this definition in his 1925 paper saying that in his 1922 paper the "discrimination among statistics within the efficient group, a discrimination which is essential to the advance of the theory of small samples, was left in much obscurity." In Section 12 of that paper, Fisher claimed that among all consistent, efficient estimators T, the MLE minimizes the information loss.

Fisher did not provide any justification for the statement, but geometrical analysis greatly clarifies the situation. His calculation of information loss for the MLE yielded the expression

$$i(\theta)\{i(\theta)^{-2}[\mu_{02} - 2\mu_{21} + \mu_{40}] - 1 - i(\theta)^{-3}[\mu_{11}^2 + \mu_{30}^2 - 2\mu_{11}\mu_{30}]\} \tag{3.1}$$

where, with $p_\theta = p(y \mid \theta)$ representing the density for each observation,

$$\mu_{jk} = E_\theta([D_\theta p]/p_\theta)^j([D_\theta^2 p]/p_\theta)^k.$$

Fisher's second claim was that the information lost by the MLE could be recovered using what is now (apparently since Efron and Hinkley, 1978) called observed information, $I_y(\hat\theta) = -D^2 l_y(\hat\theta)$. He said (1925, page 724), "With the aid of such an ancillary statistic [as $I_y(\hat\theta)$] the loss of accuracy tends to zero for large samples." This statement involves two facts. The first is that the statistic $(\hat\theta, I_y(\hat\theta))$ has zero information loss in the limiting sense defined above. The geometrical argument will be given in Section 3.3. The second is that a suitably normalized version A of the statistic $I_y(\hat\theta)$ is approximately ancillary, so that $i^A(\theta)$ tends to zero (Efron and Hinkley, 1978; Amari, 1982b; Skovgaard, 1985). From property (v) we will then obtain zero limiting information loss for the MLE after conditioning on A.

3.2 Curved Exponential Families

3.2.1 Imbedded Subfamilies

One-parameter subfamilies of exponential families of order k (for $k > 1$) may or may not themselves be exponential families of order 1. The Hardy–Weinberg model (defined in Section 1.2.3), for example, is itself a Binomial(2; θ) family and, therefore, a one-dimensional regular exponential family. (The parameter θ was β in Section 1.2.3.) The Trinomial(n; p_1, p_2) distribution is a two-dimensional regular full exponential family with natural parameter components $\eta_j = \log(p_j/p_3)$ for $j = 1, 2$. Substituting the expressions for p_1 and p_2 into the definitions of η_j, we obtain $\eta_1(\theta) = 2\log(\theta/(1-\theta))$ and $\eta_2(\theta) = \log(\theta/(1-\theta)) + \log 2$. Thus, the multinomial natural parameter restricted by this model becomes

$$\eta = \eta_A + \xi(\theta) \cdot \eta_B \tag{3.2}$$

where $\eta_A^T = (0, \log 2)$, $\eta_B^T = (2, 1)$ and $\xi(\theta) = \log(\theta/(1-\theta))$. The specification of the Hardy–Weinberg model within the trinomial, according to (3.2), restricts η to an affine subspace of the natural parameter space N. It may be verified that the ability to write $\eta(\theta)$ in the form (3.2) is necessary and sufficient for a one-parameter subfamily to be itself an exponential family of order 1.

To formalize the notion of a one-parameter subfamily of an exponential family, let $\mathscr{E}_0$ be the subfamily of $\mathscr{E}$ defined by the restriction of η to a subset N_0 of N, where N_0 is the image of a curve in N, i.e., of a mapping $\eta\colon \Theta \to N$, with Θ being an open interval. If $\mathscr{E}_0$ is an imbedded subfamily of $\mathscr{E}$, I will call it a one-parameter *imbedded exponential subfamily* or *curved exponential family*. As explained in Section 1.2.3, this means that the mapping $\eta(\cdot)$ is one-to-one, smooth (infinitely differentiable), and of rank one (its first derivative is not the zero vector), and in addition $\eta(\cdot)$ is a homeomorphism onto N_0 with the inherited topology (that is, for sequences $\{\theta_j\}$, $\eta(\theta_j) \to \eta(\theta)$ if and only if $\theta_j \to \theta$). For some purposes, it is preferable to represent $\mathscr{E}_0$ in the mean-value parameter space M as the image M_0 of $\mu\colon \Theta \to M$. Since (as noted in Section 1.2.5) the mapping from N to M is a diffeomorphism, it follows that N_0 is imbedded in N if and only if M_0 is imbedded in M.

The definition given here is slightly different than that of Efron (1975), who required only that $\eta(\cdot)$ be twice continuously differentiable rather than a smooth imbedding. With regard to the smoothness condition, it would be possible to count the required number of derivatives for each result, but there seems to be little point in doing so. If $\eta(\cdot)$ is one-to-one, smooth and of rank 1, but not necessarily a homeomorphism, $\mathscr{E}_0$ may be called an *immersed exponential subfamily*. I am requiring the additional topological condition to avoid nuisances such as the inconsistency of MLE's described in Section 2.1.5. It is easy to verify that if $\mathscr{E}_0$ is an immersed exponential subfamily, then it satisfies Cramér's conditions; but these only guarantee the existence of a well-behaved sequence of roots of the likelihood equation. Of course, since imbedded exponential subfamilies are also immersed, they too satisfy Cramér's conditions.

An important special case is that of a one-parameter subfamily of a Multinomial(n; $p_1, \cdots, p_{k+1}$) family. As mentioned in my introductory comments, this case was routinely analyzed by Fisher. In addition, in a series of papers, Rao (1961, 1962, 1963) studied more formally the problem of estimation in such families. The basic problem is to choose among the many proposed estimators. To name just a few, in addition to the MLE of θ, there is the minimum chi-square estimator found by minimizing $\sum_{j=1}^{k+1} (\hat{p}_j - p_j(\theta))^2/p_j(\theta)$ where $\hat{p}_j$ is the observed proportion for the jth sample value, the minimum Hellinger distance estimator found by minimizing $\sum_{j=1}^{k+1} (\hat{p}_j^{1/2} - p_j(\theta)^{1/2})^2$, and the minimum Kullback–Leibler number estimator found by minimizing $\sum_{j=1}^{k+1} p_j(\theta)\log(p_j(\theta)/\hat{p}_j)$. Note that the MLE is found by minimizing the Kullback–Leibler number with $p_j(\theta)$ and $\hat{p}_j$ interchanged. Least-squares and weighted least-squares estimators might also be considered. For example, after transforming to $\eta_j = \log(p_j/p_{k+1})$, $\sum_{j=1}^{k} w_j(\hat{\eta}_j - \eta_j(\theta))^2$ might be minimized, where $\hat{\eta}_j = \log(\hat{p}_j/\hat{p}_{k+1})$ and $\{w_j : j = 1, \cdots, k\}$ is some set of weights, which may depend on the data.

Each of these estimators involves solving some *estimating equation*, defined by a minimization problem. The solution may then be considered a mapping from the space of observations, that is, the sample space of the natural sufficient statistic for the Multinomial family, to the parameter space Θ. In the discussion here, it is convenient to restrict estimators to be mappings from M into Θ. This avoids nonexistence problems for the MLE, and for asymptotics it is irrelevant insofar as we will consider the mean $\bar{Y}$ of n iid observations from the curved exponential family and $\bar{Y}$, with probability one, will fall within M for sufficiently large samples.

When an estimator T is continuous, $T(\bar{Y}) \to T(\mu(\theta))$ with probability one. Thus, a continuous estimator T will be strongly consistent if and only if $T(\mu(\theta)) = \theta$. The latter condition may be considered a representation of Fisher consistency: T produces the correct value θ when applied to $\mu(\theta)$ rather than an observation $\bar{y}$. In addition to being continuous, it is convenient to assume that an estimator is smooth and of full rank. This will allow the decomposition used in Section 3.2.3. An estimator $T\colon M \to \Theta$, where θ is the parameter of a curved exponential family, will be called *regular* if on some neighborhood V of $M_0 = \mu(\Theta)$

in M, T is smooth and of rank 1 and for all $\theta \in \Theta$, $T(\mu(\theta)) = \theta$. We may now note that well-behaved estimating equations produce regular estimators. Let W be a neighborhood of M_0 in M. If $f: \Theta \times W \to R$ is a smooth function of rank 1 and $f(\theta, \mu(\theta)) = 0$ for all $\theta \in \Theta$, then it follows from the Implicit Function Theorem (e.g., Spivak, 1979) that there exists an open neighborhood V of M_0 in W on which a regular estimator T is uniquely defined by the estimating equation $f(T(y), y) = 0$.

In the case of the likelihood equation f is given by

$$f(\theta, y) = (y - \mu(\theta))^T D_\theta \eta[\theta].$$

Again, this function uniquely defines the MLE in a neighborhood of the curved exponential family. Furthermore, for a curved exponential family, the MLE is regular.

An example that is important in both theoretical and applied statistics is nonlinear regression. Let $Y_j = \eta_j + \varepsilon_j$, $j = 1, \cdots, k$, where the ε_j's are iid $N(0, \sigma^2)$ and $\eta_j = \eta_j(\theta) = h(\theta, x_j)$, with σ positive and known, Θ an open interval, and h a real function of $\Theta \times C$, where C is a subset of the real line. These models are one-parameter subfamilies of the Normal location model $Y \sim N_k(\eta, \sigma^2 I_k)$ with $\eta \in N = R^k$, and σ a known positive number. As a first special case, if $h(\theta, x) = \alpha + \theta x$ we obtain a simple linear regression model, with fixed intercept α, which defines an exponential family of order 1. As another special case, if $h(\theta, x) = \exp(-\theta x)$, we obtain the one-parameter exponential nonlinear regression model.

Binary regression also furnishes a class of examples. Let Z_i be independently Binomial(n_i, p_i), with $p_i \in (0, 1)$ for $i = 1, \cdots, k$. If the p_i's are not further restricted, then the Z_i's form a product of k exponential families, which is itself an exponential family of order k. The natural parameter for this family has components $\eta_i = \log(p_i/(1 - p_i))$. Now suppose h is a real function on $\Theta \times C$, where Θ is an open interval and C is a subset of the real line. Then, as in the nonlinear regression setting, $\eta_i = \eta_i(\theta) = h(\theta, x_i)$ defines a one-parameter subfamily of the Binomial product model. A first special case $h(\theta, x) = \alpha + \theta x$ yields a simple logistic regression model with fixed intercept α, which again is an exponential family of order 1. As a second case, letting F be a continuous distribution function, take $h(\theta, x) = \log(p/(1 - p))$ with $p = F(\alpha + \theta x)$. When F is the Normal(0, 1) distribution function, we obtain the probit regression model with fixed intercept α, which is not an exponential family of order 1.

3.2.2 The Auxiliary Space Associated with an Estimator

At the heart of Fisher's analysis of an estimator T is a replacement of the data $\bar{Y}$ by a sufficient statistic (T, A). From $i^{(T,A)}(\theta) = i^T(\theta) + i^{A|T}(\theta)$, which is property (v) of Section 3.1.1 (with the roles of T and A reversed), the information not contained in T must instead be contained in the conditional distribution of A given T; thus, good estimators are those for which there is little information about θ in A given T.

In curved exponential families, the situation is simplified and it is easy to construct a suitable statistic A. If T is a regular estimator, then for each $\theta_0 \in \Theta$ there exists a neighborhood U of $\mu(\theta_0)$ in M and a diffeomorphism (T_U, A) of U onto an open subset of R^k such that the intersection of U with M_0 is $\{\mu \in U: A(\mu) = 0\}$ and T_U is the restriction of T to U. (This follows from the Implicit Function Theorem.) This allows a local decomposition of M near $\mu(\theta_0)$ which is identified by (T, A) with a product subset of $\Theta \times R^{k-1}$. As in Figure 5, a point μ in U becomes identified with $(T(\mu), A(\mu)) = (\theta, \alpha)$. The data $\bar{Y}$, with probability one, will fall in U for all sufficiently large n. Large deviations results (as in, for instance, Brown, 1986, Chapter 7) may be used to show that replacing the sample space of $\bar{Y}$ by U does not affect the local asymptotic calculations given below, in Section 3.3. Making this replacement and restricting the family to $\{\theta: \mu(\theta) \in U\}$, $(T(\bar{Y}), A(\bar{Y}))$ becomes sufficient, and the amount of information lost by T is the amount contained in the distribution of $A(\bar{Y})$ given T. The set $A_t = \{\mu \in U: T(\mu) = t\}$ is called the *auxiliary subspace* associated with the estimator T at t. In Section 3.3, we will see that the geometry of the decomposition, specifically involving both M_0 and the auxiliary subspaces, determines the loss of information of T. In general, an open set together with a diffeomorphism onto an open subset of R^k is called a local *coordinate system* and is a basic element in the construction of a smooth manifold. In statistics, these become local parameterizations (see Section 2.2). Here, (θ, α) is a local parameterization of $\mathscr{E}$ on the domain U.

I would like to draw attention to an important but potentially confusing point. If the MLE of (θ, α) as

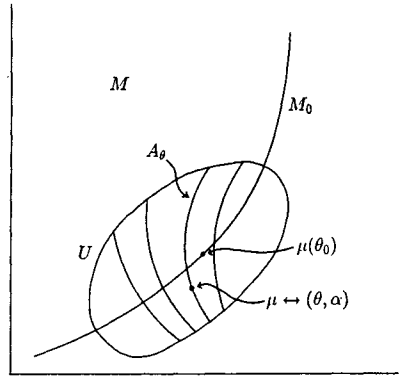

FIG. 5. *The local decomposition of M near $\mu(\theta_0)$. Points μ within the neighborhood U become identified by coordinates (θ, α).*

a parameterization of U is denoted by (θ^*, α^*) then $(T, A)(\tilde{y}) = (\theta^*, \alpha^*)$ for $\tilde{y} \in U$ and, in particular, $T(\tilde{y}) = \theta^*$. Thus, every regular estimator T may be viewed the first component of the MLE of (θ, α) for some α. T need not be the MLE of θ as a parameter of Q_0, that is, θ^* need not equal $\hat{\theta}$. The definition of a component of a parameter, here θ, depends on the manner in which other components are used in identifying the model: when θ is viewed as the first component of (θ, α), θ^* results from ML estimation, whereas when it is viewed as the only parameter of $\mathscr{E}_0$, $\hat{\theta}$ results. This is a general phenomenon most familiar in the regression context: the meaning of a regression coefficient for a particular explanatory variable depends on the span of the other explanatory variables in the model.

3.2.3 The Information Inner Product and Exponential and Mixture Curvatures

The *information inner product* on the natural parameter space N at η is defined by

$$\langle , \rangle_\eta : N \times N \to R,$$
$$\langle v, w \rangle_\eta = v^T i(\eta) w.$$

Since $i(\eta)$ is positive definite for a regular exponential family, the inner product is well defined. The information inner product on the mean-value parameter space M is also needed. It is defined by

$$\langle , \rangle_\mu : M \times M \to R,$$
$$\langle v, w \rangle_\mu = v^T i(\mu) w,$$

where $i(\mu)$ is the information matrix in terms of μ, i.e., $i(\mu)_{jk} = E_\mu(\partial_j l_Y[\mu] \partial_k l_Y[\mu])$ with $l_y(\mu)$ being the loglikelihood function on μ. Note that there is the change-of-variables formula $i(\mu) = (D_\mu \eta)^T i(\eta)(D_\mu \eta)$.

Suppose $\eta^{(1)}$ and $\eta^{(2)}$ are two one-to-one, smooth, rank 1 curves in N and let $\mu^{(1)}$ and $\mu^{(2)}$ be the corresponding curves $\mu^{(i)} = \mu(\eta^{(i)})$ in M. (These would be representations of two curved exponential families.) Suppose further that the curves intersect at a point η_0 in N and $\mu_0 = \mu(\eta_0)$ in M. Writing $D\eta^{(1)}$ for the derivative of $\eta^{(1)}$ at η_0, etc.,

$$\langle D\mu^{(1)}, D\mu^{(2)} \rangle_{\mu_0} = \langle D\eta^{(1)}, D\eta^{(2)} \rangle_{\eta_0}. \tag{3.3}$$

Thus, as one consequence, the angle between two curves with respect to the information inner product does not depend on whether it is measured in M or N. It may be noted that the same argument could also be applied using any alternative parameterization β: $\mathscr{E} \to B$, with the analogous definition of the information inner product on B, as long as the transformation from B to N is a diffeomorphism. Another formula that will be used below is

$$\langle D\mu^{(1)}, v \rangle_{\mu_0} = (D\eta^{(1)})^T v$$

for all v in M. This follows from the chain rule $D\eta^{(1)} = D_\eta \mu \cdot D\mu^{(1)} = i(\mu_0) D\mu^{(1)}$.

Now suppose $\mathscr{E}_0$ is an imbedded exponential subfamily defined by $\eta : \Theta \to N$. Letting $i(\theta)$ be the Fisher information for $\mathscr{E}_0$,

$$i(\theta) = \| D_\theta \eta \|^2_{\eta(\theta)},$$

as is easily verified from the definition of $i(\theta)$, together with the chain rule. The quantity

$$\gamma = \gamma(\theta) = \| (D^2_\theta \eta)_N \|_{\eta(\theta)} \cdot \| D_\theta \eta \|^{-2}_{\eta(\theta)}$$

where the normal component $(D^2_\theta \eta)_N$ and the norms are computed with respect to the $\langle , \rangle_{\eta(\theta)}$-inner product, is called the *exponential curvature* or the *statistical curvature* of $\mathscr{E}_0$ at θ. The analogous quantity

$$\beta = \beta(\theta) = \| (D^2_\theta \mu)_N \|_{\mu(\theta)} \cdot \| D_\theta \mu \|^{-2}_{\mu(\theta)}$$

calculated with respect to the $\langle , \rangle_{\mu(\theta)}$-inner product, is called the *mixture curvature* of $\mathscr{E}_0$ at θ. Note that each of these is a direct generalization of ordinary curvature, as in equation (1.2) of Section 1.2.3. The exponential curvature measures the departure of $\mathscr{E}_0$ from being an exponential family. The mixture curvature measures the departure of $\mathscr{E}_0$ from being a mixture family.

3.3 Geometrical Interpretation of Fisher's Principles

3.3.1 Efficiency

Using the local parameterization (θ, α), described in Section 3.2.2, a simple geometrical expression for the asymptotic variance of the estimator T may be derived. Consider first an imbedded Normal nonlinear regression model. As shown in Figure 6 (for $k = 2$), the angle between N_0 and A_θ is, by definition, the angle ϕ between $D_\theta \eta = D_\theta \eta[\theta]$ and the tangent hyperplane to A_θ. This, in turn, is the angle between $D_\theta \eta$ and the vector closest to it that lies in the tangent hyperspace (when $k \geq 2$). Writing $Z \equiv D_\theta \eta$ and, regarding η as a function of both θ and α on U, $X \equiv (D_\alpha \eta(\theta, \alpha))$ so that X is a $k \times (k-1)$ matrix spanning the tangent hyperplane, there is the familiar

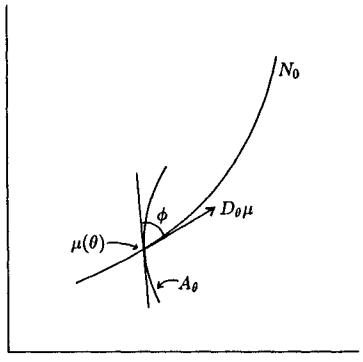

FIG. 6. *The angle between N_0 and A_θ is the angle ϕ.*

expression

$$Z^TZ \cdot \cos^2\phi = Z^TX(X^TX)^{-1}X^TZ. \tag{3.4}$$

Now, to obtain the asymptotic variance of T, the information matrix for (θ, α) may be used, with the formula above substituted for simplification. Viewing η first of all as providing a Euclidean coordinate system on N, the parameter space for the k-dimensional Normal$(\eta, \sigma^2 I_k)$ family, the information matrix $i(\eta)$ is σ^{-2} times the identity. Changing coordinates to (θ, α) using the general formula

$$i(\gamma) = (D_\gamma\beta)^T i(\beta)(D_\gamma\beta)$$

where $(D_\gamma\beta)$ is the k by k matrix with $(D_\gamma\beta)_{ij} = \partial_j\beta_i$ and $\beta = \beta(\gamma)$, yields the partitioned matrix

$$i(\theta, \alpha) = \sigma^{-2}\begin{Bmatrix} Z^TZ & Z^TX \\ X^TZ & X^TX \end{Bmatrix}$$

with X and Z as defined above, and then the upper left block of the partitioned inverse is

$$(i(\theta, \alpha)^{-1})_{11} = \sigma^2\{Z^TZ - Z^TX(X^TX)^{-1}X^TZ\}^{-1}.$$

Substitution of (3.4) into this expression yields the desired asymptotic variance formula. We find that if T is a regular estimator for a one-dimensional imbedded Normal nonlinear regression model $\mathscr{E}_0$ with variance σ^2, and $\bar{Y}_n$ is the mean of n iid observations from $Q_{\eta(\theta)} \in \mathscr{E}_0$, then the asymptotic variance of T is given by

$$\text{Avar}_\theta(n^{1/2}T(\bar{Y}_n)) = \sigma^2\{\sin^2\phi \cdot \| D_\theta\eta[\theta]\|^2\}^{-1}$$

where ϕ is the angle between N_0 and A_θ.

The interpretation of the formula is clear: note that $\| D_\theta\eta[\theta]\|^2 = i(\theta)$; the asymptotic efficiency of the estimator T is $\sin^2\phi$. Together, T and A contain all of the information about θ; when they are highly asymptotically correlated, A will contain much information about θ and the use of T alone results in substantial loss of information. The coefficient $\sin^2\phi$ depends only on the angle ϕ between N_0 and A_θ. When the estimator T has its auxiliary space A_θ orthogonal to N_0, it becomes efficient.

The corresponding result for general one-dimensional curved exponential families is obtained by repeating the argument, while noting that the information matrix $i(\eta)$ is no longer equal to a multiple of the identity. If T is a regular estimator for a one-dimensional curved exponential family $\mathscr{E}_0$ and $\bar{Y}_n$ is the mean of n iid observations from $Q_{\eta(\theta)} \in \mathscr{E}_0$, then the asymptotic variance of T is given by

$$\text{Avar}_\theta(n^{1/2}T(\bar{Y}_n)) = \{\sin^2\phi \cdot \| D_\theta\eta[\theta]\|^2_{\eta(\theta)}\}^{-1}$$

where $\| \cdot \|_{\eta(\theta)}$ is the norm, and ϕ is the angle between N_0 and A_θ, with respect to the information inner product on N at $\eta(\theta)$. The interpretation here is analogous to that in the Normal case and was discussed by Efron (1982, Section 6).

This result may be applied to estimators that are defined by estimating equations based on smooth, full-rank functions. Suppose $f(\cdot, \cdot)$ is such a function and let D_1f and D_2f be its partial derivatives with respect to its two arguments. Note that D_1f is a scalar while D_2f is the gradient of $f(\theta, \cdot)$ and, therefore, is normal to the contour $A_\theta = \{\mu: f(\theta, \mu) = 0\}$, with respect to the usual Euclidean inner product. Letting $n_\theta = i(\mu(\theta))^{-1}D_2f(\theta, \mu(\theta))$, there is then for any vector v tangent to A_θ at $\mu(\theta)$,

$$0 = v^T(D_2f) = \langle v, n_\theta \rangle_{\mu(\theta)}$$

where $(D_2f) = D_2f(\theta, \mu(\theta))$. Thus, n_θ is normal to A_θ at $\mu(\theta)$, with respect to the $\langle , \rangle_{\mu(\theta)}$-inner product. The angle between n_θ and M_0 thus determines the efficiency of the estimator defined by $f(\cdot, \cdot)$.

In the case of the MLE, from the likelihood equation

$$0 = f(\theta, \bar{y}) = (\bar{y} - \mu(\theta))^T(D_\theta\eta)$$

there is

$$D_2f = D_\theta\eta = D_\mu\eta \cdot D_\theta\mu = i(\mu(\theta))D_\theta\mu.$$

Thus, $n_\theta = D_\theta\mu$ so that the angle between M_0 and A_θ is $\pi/2$. This maximizes $\sin^2\phi$ and thereby minimizes the asymptotic variance. Thus, we have a geometrical argument that among regular estimators for a one-dimensional curved exponential family, the MLE minimizes the asymptotic variance.

It is easy to check whether an estimator is efficient by following the preceding argument. For instance, for the least-squares estimator minimizing $\Sigma(\bar{y}_i - \mu_i(\theta))^2$ there is $D_2f = D_\theta\mu$. When $i(\mu(\theta))$ differs from a multiple of the identity, as it does when the family is non-Normal, n_θ is not proportional to $D_\theta\mu$, so the least-squares estimator is inefficient. The weighted least-squares estimator is also easily shown by this method to be efficient when the weight-matrix $i(\hat{\mu})$ is used where $\hat{\mu} = \bar{y}$.

3.3.2 Information Loss

Fisher was interested in the quantity $\lim ni(\theta) - i^T(\theta)$, which has been called the (limiting) *information loss*. Unlike the efficiency measure $i^T(\theta)/ni(\theta)$ (or its limit), the quantity $ni(\theta) - i^T(\theta)$, is not invariant with respect to reparameterization. Thus, it is in this respect desirable to consider the (limiting) *relative information loss*, $\lim i(\theta)^{-1}[ni(\theta) - i^T(\theta)]$. This quantity was interpreted by Fisher (1925, page 720) as the number of additional observations required when using T in place of the whole sample.

There are two issues involved with the information loss of efficient estimators: how much information must be lost, and which estimators lose the least

possible. As far as the first question is concerned, a quick appreciation of the role of the model and the simplicity of the results may be gained from the following version of a heuristic argument used by Fisher, here presented in the context of Normal nonlinear regression.

To calculate the difference $ni(\theta) - i^T(\theta)$, it is most convenient to use a result that is due to Fisher: for a regular estimator T of θ in a curved exponential family,

$$D_\theta l_t(\theta) = E_\theta(D_\theta l_Y[\theta] \mid T = t)$$

where $l_t(\theta) = \log p(t \mid \theta)$ is the loglikelihood based on the estimator T at t. Using $V(Dl_Y) = E(V(Dl_Y \mid T)) + V(E(Dl_Y \mid T))$ together with this result,

$$ni(\theta) - i^T(\theta) = E_\theta(V_\theta(D_\theta l_Y \mid T)). \tag{3.5}$$

To get the limiting relative information loss, a new parameterization may be chosen freely; for the nonlinear regression model it is most convenient to use arclength. For this parameterization,

$$\begin{aligned} D_s^2 l_y[s] &= -n\sigma^{-2}[(D_s^2\eta)^T(\bar{y} - \eta(s)) - (D_s\eta)^T(D_s\eta)] \\ &= -n\sigma^{-2}[\kappa u^T(\bar{y} - \eta(s)) - 1] \end{aligned}$$

where $u = (D_s^2\eta)/\| D_s^2\eta \|$ is a unit vector and $\kappa = \| D_s^2\eta \|$ is the curvature of $\eta(\cdot)$ at s. Now, putting the above in the approximation

$$\begin{aligned} D_s l_y[s] &= D_s^2 l_y[\hat{s}](s - \hat{s}) + R_1 \\ &= -D_s^2 l_y[s](\hat{s} - s) + R_1 + R_2 \end{aligned}$$

where R_1 and R_2 are lower-order remainder terms yields

$$D_s l_y[s] = n\sigma^{-2}[\kappa u^T(\bar{y} - \eta(s)) - 1](\hat{s} - s) + R_1 + R_2.$$

The variable $n^{1/2}u^T(\bar{Y} - \eta(s))$ is distributed as $N(0, \sigma^2)$ (under the true value s) so that when the remainders are ignored there is, with T being the MLE of s,

$$V_s(D_s l[s] \mid T = \hat{s}) = n\sigma^{-2}\kappa^2(\hat{s} - s)^2.$$

Finally, since $i(s) = \sigma^{-2}(D_s\eta)^T(D_s\eta) = \sigma^{-2}$, the limiting relative information loss of the MLE is found by taking the expectation and passing to the limit while noticing that $n\sigma^{-2}(\hat{s} - s)^2$ is asymptotically χ_1^2:

$$\lim i(\theta)^{-1}[ni(\theta) - i^T(\theta)] = \sigma^2\kappa^2$$

where T is the MLE.

This argument could be made rigorous, but the general result for an arbitrary regular efficient estimator, a result which involves statistical and mixture curvatures rather than ordinary curvature, has been established by a somewhat different derivation. Assuming Fisher's assertion that the limiting information loss is minimized by the MLE was correct (and, of course, it was), the "result" above already shows how the model determines the least limiting information loss. For linear models the MLE is sufficient and there is zero loss of information, and also zero curvature. As the model becomes more curved, however, the MLE is less able to summarize the sample. Also, the factor σ^2 appears in the expression: as σ decreases, the probability increases that the observation $\bar{Y}$ will be near $\eta(\theta)$, and in terms of the limiting information loss, the curve effectively becomes more nearly linear.

The basis of the alternative derivation is an expansion of the score function using the local reparameterization (θ, α). From $ni(\theta) = i^{(T,A)}(\theta)$ and $i^{(T,A)}(\theta) = i^{(A \mid T)}(\theta) + i^T(\theta)$ it is apparent that, as in the analysis of efficiency, the loss of information will depend on the amount of information in A given T. This is the average information in A given $T = t$, averaged over T, and the set $T = t$ is A_t. Thus, the shape of the auxiliary space will once again play a major role. The simplest case of the general result is the following.

Suppose T is a regular efficient estimator for a one-dimensional imbedded Normal nonlinear regression model where the ambient full Normal exponential family has dimension $k = 2$; then the limiting relative information loss at $\theta \in \Theta$ is

$$\lim i(\theta)^{-1}[ni(\theta) - i^T(\theta)] = \sigma^2\kappa^2 + (\tfrac{1}{2})\sigma^2\kappa_a^2$$

where κ is the curvature of the curve $\eta(\cdot)$ at θ and κ_a is the curvature of A_θ (i.e., A_t with $t = \theta$) at $\alpha = 0$. This says that there is a remarkably simple decomposition of the relative information loss into two components. The first is inherent to the model and does not depend on the estimator used, while the second is determined by the estimator. Both are products of σ^2 and squared curvatures: as σ decreases, the statistical effect of curvature diminishes, the curves effectively straightening out in the local statistical sense of the result. For the MLE, $A_{\hat{\theta}}$ is the set of possible observed values that would lead, via the likelihood equations (least-squares, in this case), to $\hat{\theta}$, i.e., $A_{\hat{\theta}} = \{\eta: (\eta - \eta(\hat{\theta}))^T D\eta(\hat{\theta}) = 0\}$. Thus, $A_{\hat{\theta}}$ is a line segment in N, so that $\kappa_a = 0$, which verifies Fisher's claim in this restricted context: the MLE minimizes the information loss.

The generalization to curved exponential families requires an additional regularity condition: the Edgeworth density approximation must converge uniformly to the true density of the sample mean $\bar{Y}$. A sufficient condition (Feller, 1966, Section 16.2) is that for all $\eta \in N$, $E_\eta(\exp(it^T Y))^p$ be an integrable function of t for some $p \geq 1$. This rules out discrete distributions, and a counterexample was given by Efron (1975) in the multinomial case. (Although the existence of such counterexamples must have been known to earlier workers, it is very nice to have a clear prescription

in the literature: it is a reminder that moments of asymptotic distributions can differ from limits of corresponding sample moments, and it also furnishes a counterexample to the Koopman-Pitman-Darmois theorem in the discrete case.) The remainder of the discussion will be carried out assuming this condition is satisfied.

Suppose $\mathscr{E}_0$ is a one-dimensional imbedded exponential subfamily of a two-dimensional regular full exponential family $\mathscr{E}$ and T is a regular efficient estimator for $\theta \in \Theta$. Then the limiting relative information loss of T based on iid sampling from $\mathscr{E}_{\eta(\theta)}$ is

$$\lim i(\theta)^{-1}[ni(\theta) - i^T(\theta)] = \gamma^2 + (\tfrac{1}{2})\beta^2 \tag{3.6}$$

where γ is the exponential curvature of $\mathscr{E}_0$ at θ and β is the mixture curvature of A_θ at $\alpha = 0$. Note that in the nonlinear regression context we have $\gamma = \sigma\kappa$ and $\beta = \sigma\kappa_a$. The exponential and mixture curvatures measure curvature in a statistical sense, adjusting for standard deviation. From this result, the exponential curvature may, in general, be considered a measure of the insufficiency of the MLE, while the mixture curvature of the auxiliary space quantifies the deficiency of other estimators.

In the general case, in which the dimension of $\mathscr{E}$ is $k \geq 2$, the expression for relative information loss remains the same with

$$\beta^2 = \sum_{i,j,k,l} g^{ik} g^{jl} \langle (\partial_{ij}\mu)_N, (\partial_{kl}\mu)_N \rangle \tag{3.7}$$

where the indices correspond to components of α, g^{ik} is the (i, k)-element of the asymptotic covariance matrix of the auxiliary statistic A (based on Fisher information for (θ, α) at $\alpha = 0$), the derivatives are taken with respect to components of the auxiliary coordinates and evaluated at $\alpha = 0$ and the normal components and norms are based on the information inner product at (θ, α) for $\alpha = 0$ (with normal meaning normal to the auxiliary space A_θ). Reduction of this expression to the previous one when $k = 2$ is immediate, since then the mixture curvature is $\| D_\alpha \mu \|^{-2} \cdot \| (D_\alpha^2 \mu)_N \|$ and the asymptotic variance of A becomes $\| D_\alpha \mu \|^{-2}$.

Equation (3.6) is a one-parameter version of a multiparameter expression derived by Amari (1982b). I will comment further on it in Section 3.5. It is clear, though, that the result provides verification of Fisher's claim in the case of one-dimensional curved exponential families: the limiting information loss is minimized by the MLE since $A_\theta = \{\mu \in U: (\mu - \mu(\theta))^T D_\theta \eta(\theta) = 0\}$ is flat in M, so that the derivatives $\partial_{ij}\mu$ in β^2 vanish and $\beta^2 = 0$ in this case.

To conclude this section, I note that the argument used (by Amari) to derive the tensorial form of (3.6) involves an expansion of the score function,

$$\begin{aligned} n^{-1} D_\theta l_y[\theta_0] = {} & (D_\theta \eta)^T (D_\theta \mu)(\theta^* - \theta_0) \\ & + \sum_i (D_\theta \eta)^T (D_{\alpha_i} \mu)\alpha_i \\ & + \frac{1}{2}(D_\theta \eta)^T (D_\theta^2 \mu)(\theta^* - \theta_0)^2 \\ & + \sum_i (D_\theta \eta)^T (D_{\theta\alpha_i} \mu)(\theta^* - \theta_0)\alpha_i \\ & + \frac{1}{2}\sum_{i,j} (D_\theta \eta)^T (D_{\alpha_i\alpha_j} \mu)\alpha_i\alpha_j \\ & + R(\theta^*, \alpha) \end{aligned} \tag{3.8}$$

where $T(y) = \theta^*$. The asymptotic expectation of the conditional variance, conditional on T, gives the desired result according to (3.5). Omitting details, the outline is clear. First of all, after conditioning on T, the variances of the first and third terms (and cross-product terms involving these) vanish. The second term vanishes because of the orthogonality relation,

$$0 = (D_\theta \eta)^T (D_{\alpha_i} \mu)$$

which is equivalent to $\langle D_\theta \mu, D_{\alpha_i} \mu \rangle_{\mu(\theta)} = 0$, and is required for efficiency of the estimator (as in Section 3.3.1). Differentiation of this equation produces

$$(D_\theta \eta)^T (D_{\theta\alpha_i} \mu) = -(D_\theta^2 \eta)^T (D_{\alpha_i} \mu) \tag{3.9}$$

which allows conversion of the fourth term in (3.8) to a summation involving $D_\theta^2 \eta$, and this leads to the curvature γ^2. The fifth term leads to β^2.

3.3.3 Information Recovery

The quantity β^2 appearing in the expression for information loss is invariant in the sense that it does not depend on the choice of coordinates for the auxiliary space A_θ. In this subsection, A will be transformed to simplify the fourth term of (3.8), so that it will be apparent geometrically that a one-dimensional statistic (the first component of A) may be used to recover the information lost by the MLE. Then, an explicit expression for this statistic will verify Fisher's second claim (see Section 3.1.2).

We first transform A so that $D_\alpha \mu$ is made orthogonal, and then rotate the result so that the first column is in the direction of $(D_\theta^2 \eta)_N$. Using (3.9), we then have

$$(D_\theta \eta)^T (D_{\theta\alpha_i} \mu) = -(D_\theta^2 \eta)^T (D_{\alpha_1} \mu) \cdot \delta_{i1}$$

where δ_{i1} is 1 if $i = 1$ and 0 otherwise, and the fourth term in the score function expansion (3.8) becomes

$$\begin{aligned} & \Sigma (D_\theta \eta)^T (D_{\theta\alpha_i} \mu)(\theta^* - \theta_0)\alpha_i \\ & \quad = -(D_\theta^2 \eta)(D_{\alpha_1} \mu)(\theta^* - \theta_0)\alpha_1. \end{aligned} \tag{3.10}$$

This is especially interesting: suppose T is the MLE so that the normal components of $D_{\alpha_i\alpha_j}\mu$ vanish. Then, from the discussion following (3.8), the only term contributing to the limiting information loss is (3.10). It is apparent, however, that conditioning on (T, A_1) rather than T alone will make the conditional variance of this term vanish. Thus, the limiting information loss of (T, A_1) will be zero.

To obtain A_1 explicitly, assume that $T = \hat{\theta}$ is the MLE. Note that A_θ may be regarded as a vector space of vectors with tails at $\mu(\theta)$, i.e., $A_\theta = \mu(\theta) + \{\mu - \mu(\theta)\colon \langle \mu - \mu(\theta), D_\theta\mu\rangle_{\mu(\theta)} = 0\}$. Pick an orthonormal basis $v_1(\theta), \cdots, v_{k-1}(\theta)$ for this space (i.e., for $A_\theta - \mu(\theta)$) in such a way that the $v_i(\theta)$'s vary smoothly with θ. For instance, Gram–Schmidt orthogonalization could be used with the existing basis vectors $D_{\alpha_i}\mu$. The resulting coordinate system will then be $(\theta, \alpha_1, \cdots, \alpha_{k-1})$ with the α_i's defined by

$$\mu - \mu(\theta) = \Sigma\alpha_i v_i(\theta)$$

and, writing $\mu = \mu(\theta) + \Sigma\alpha_i v_i(\theta)$, this implies

$$D_{\alpha_1}\mu = v_1$$

where $v_1 = v_1(\theta)$. Decomposing $D_\theta^2\eta$ into its tangential and normal components $D_\theta^2\eta = (D_\theta^2\eta)_T + (D_\theta^2\eta)_N$, to require $D_{\alpha_1}\eta$ to be in the direction of $(D_\theta^2\eta)_N$ is to require

$$D_{\alpha_1}\eta = b(\theta)\cdot(D_\theta^2\eta)_N$$

where $b(\theta)$ is a number. Now, from the orthogonality of the basis,

$$\begin{aligned}\alpha_1 &= \langle v_1, \mu - \mu(\theta)\rangle_{\mu(\theta)}\\ &= \langle D_{\alpha_1}\mu, \mu - \mu(\theta)\rangle_{\mu(\theta)}\\ &= (D_{\alpha_1}\eta)^T(\mu - \mu(\theta))\\ &= b(\theta)(D_\theta^2\eta)_N^T(\mu - \mu(\theta)),\end{aligned}$$

but for any vector w tangent to N_0 at $\eta(\theta)$ there is $w^T(\mu - \mu(\theta)) = 0$ and, in particular,

$$(D_\theta^2\eta)_T^T(\mu - \mu(\theta)) = 0$$

so that

$$\alpha_1 = b(\theta)(D_\theta^2\eta)^T(\mu - \mu(\theta)).$$

For definiteness, and for subsequent use, we choose $b(\theta)$ so that $\|D_{\alpha_1}\eta\|_{\eta(\theta)} = 1$, which yields $b(\theta) = [i(\theta)\gamma(\theta)]^{-1}$. Now, to define the corresponding statistic we let $A_1 = \alpha_1(\bar{y})$ so that $\theta = \hat{\theta} = T(\bar{y})$ and we obtain the explicit form

$$A_1 = [i(\hat{\theta})\gamma(\hat{\theta})]^{-1}(D_\theta^2\eta[\hat{\theta}])^T(\bar{y} - \mu(\hat{\theta})).$$

To get the interpretation of A_1 in terms of observed information, notice that the second derivative of the loglikelihood is

$$D_\theta^2 l_y[\theta] = n(\bar{y} - \mu(\theta))^T(D_\theta^2\eta) - n(D_\theta\eta)^T i(\eta)(D_\theta\eta),$$

the second term of which is $ni(\theta)$, so that

$$A_1 = -[ni(\hat{\theta})\gamma(\hat{\theta})]^{-1}[I_y(\hat{\theta}) - ni(\hat{\theta})]. \tag{3.11}$$

This verifies Fisher's claim that observed information may be used to recover the information lost by the MLE. In addition, the choice of $b(\theta) = [i(\theta)\gamma(\theta)]^{-1}$ to normalize $D_{\alpha_1}\eta$, so that the asymptotic variance of $n^{1/2}A_1$ is 1, has two further consequences. First, since the asymptotic density of $n^{1/2}A_1$ is standard Normal, the score function based on $n^{1/2}A_1$ is of order $O(n^{-1/2})$ and, therefore, the Fisher information $i^{A_1}(\theta)$ is of order $O(n^{-1})$. Using $\lim ni(\theta) - i^{(T,A_1)}(\theta) = 0$, which was noted earlier in this subsection, together with property (v) of Fisher information (from Section 3.1.3), we obtain

$$\lim ni(\theta) - i^{T|A_1}(\theta) = 0$$

which is a fuller substantiation of Fisher's claim (see Amari, 1985; Skovgaard, 1985). Second, from the expression (3.11) we get another interpretation of the statistical curvature γ. In addition to its representing the insufficiency of the MLE, it is also the asymptotic coefficient of variation of the observed information. (This was suggested by Pierce in his discussion of Efron (1975) and was proved by Efron and Hinkley (1978).)

3.4 A Few More Facts about Curvature

3.4.1 Efron's General Formula

A formula for γ that holds outside of curved exponential families is available. Within curved exponential families, using

$$\begin{aligned}D_\theta l_y &= (y - \mu(\theta))^T D_\theta\eta,\\ D_\theta^2 l_y &= (y - \mu(\theta))^T D_\theta^2\eta - (D_\theta\eta)^T i(\eta(\theta))(D_\theta\eta)\end{aligned}$$

there are

$$\begin{aligned}V_\theta(D_\theta l_Y) &= \langle D_\theta\eta, D_\theta\eta\rangle_{\eta(\theta)},\\ \mathrm{Cov}_\theta(D_\theta l_Y, D_\theta^2 l_Y) &= \langle D_\theta\eta, D_\theta^2\eta\rangle_{\eta(\theta)},\\ V_\theta(D_\theta^2 l_Y) &= \langle D_\theta^2\eta, D_\theta^2\eta\rangle_{\eta(\theta)},\end{aligned}$$

so that

$$\begin{aligned}\gamma(\theta)^2 &= \langle (D_\theta^2\eta)_N, (D_\theta^2\eta)_N\rangle_{\eta(\theta)}(\langle D_\theta\eta, D_\theta\eta\rangle_{\eta(\theta)})^{-2}\\ &= i(\theta)^{-2}\langle D_\theta^2\eta - (D_\theta^2\eta)_T, D_\theta^2\eta - (D_\theta^2\eta)_T\rangle_{\eta(\theta)}\\ &= i(\theta)^{-2}[\langle D_\theta^2\eta, D_\theta^2\eta\rangle_{\eta(\theta)}\\ &\qquad - \langle (D_\theta^2\eta)_T, (D_\theta^2\eta)_T\rangle_{\eta(\theta)}] \qquad (3.12)\\ &= i(\theta)^{-2}[V_\theta(D_\theta^2 l_Y) - i(\theta)^{-1}\mathrm{Cov}_\theta(D_\theta l_Y, D_\theta^2 l_Y)].\end{aligned}$$

The last formula may now be applied whenever the quantities are defined. Of course, statistical curvature remains invariant with respect to reparameterization for general families.

Efron also pointed out that any regular parametric family can be approximated at a point θ_0 by a curved exponential family having the same loglikelihood derivatives at θ_0. Specifically, in terms of a loglikelihood $l_y(\theta) = \log(p(y \mid \theta))$,

$$\tilde{p}(y \mid \eta) = \exp[l_y(\theta_0) + \eta_1 \cdot Dl_y(\theta_0) + \cdots + \eta_k \cdot D^k l_y(\theta_0) - \psi(\eta)]$$

with $\exp[\psi(\eta)]$ being a normalizing constant, defines a density in an exponential family of order k (where the derivatives $D^j l_y(\theta_0)$ are the components of the sufficient statistic). The specification $\eta(\theta) = (\theta - \theta_0, (1/2)(\theta - \theta_0)^2, \cdots, (1/k!)(\theta - \theta_0)^k)$ then defines a one-parameter curved exponential subfamily. The loglikelihood derivatives satisfy $D^j(\log \tilde{p}(y \mid \eta(\cdot)))(\theta_0) = D^j l_y(\theta_0)$ and, since $\tilde{p}(y \mid \eta(\theta_0)) = p(y \mid \eta(\theta_0))$, their moments at θ_0 agree with those for the original family. Thus, the calculation of statistical curvature using the geometrical formula of Section 3.2.3 for the approximating family agrees with that using the general formula (3.12) for the original family.

An interesting example is the t_ν location family, for which Efron (1975) calculated $\gamma^2 = 6(3\nu^2 + 18\nu + 19)/[\nu(\nu + 1)(\nu + 5)(\nu + 7)]$. For the Cauchy, $\gamma^2 = 5/2$. Here is one of my favorite pieces of trivia. Suppose we ask which distribution in the t_ν family is half way between Normal and Cauchy on the statistical curvature scale (the scale of sufficiency loss of the MLE). For Normal $\gamma = 0$, while for Cauchy $\gamma = (5/2)^{1/2}$. Thus we seek ν such that $\gamma = (5/2)^{1/2}/2$. There is no reason, a priori, that ν should turn out to be an integer: it merely has to be a number greater than 1 ($\nu = 1$ for Cauchy, $\nu = \infty$ for Normal). The answer is $\nu = 3$. Thus, in the sense of insufficiency of the MLE, as measured by statistical curvature, the t on 3 degrees of freedom is halfway between Normal and Cauchy. (This supports the intuition behind the choice of t_3 tails in many robustness studies, such as Rogers and Tukey (1972).)

3.4.2 Other Applications

Statistical curvature appears in many second-order formulas. Here I give four additional examples of its occurrence.

Second-order efficiency. Rao (1961) introduced a definition of second-order efficiency of an estimator T based on the ability of a quadratic in $T - \theta$ to approximate the score function. Rao (1963) then introduced a second definition based on asymptotic risk for squared-error loss (see also Rao, 1962; Hodges and Lehmann, 1970; Ghosh and Subramanyam, 1974; Efron, 1975; Ghosh, Sinha and Wieand, 1980; Lindsay, 1988). Suppose the bias of T may be expanded and, with $\lim n \cdot E_\theta(T - \theta) = b_T(\theta)$, define the "bias-corrected" version of T by

$$\tilde{T} = T - n^{-1} \cdot b_T(T).$$

As a loss function for use in asymptotics consider $L(T, \theta) = i(\theta)(T - \theta)^2$. Then, if T is efficient we have $n \cdot E_\theta(L(\tilde{T}, \theta)) \to 1$ and

$$\lim n^2 \cdot E_\theta\left(L(\tilde{T}, \theta) - \frac{1}{n}\right) = \gamma^2 + \beta^2 + \omega^2 \tag{3.13}$$

where

$$\omega^2 = \| D_\theta \mu \|^{-2} \cdot \| (D_\theta^2 \mu)_T \| \tag{3.14}$$

with respect to the information inner product on μ. Note that ω is analogous to γ except that μ is substituted for η and the tangential component is substituted for the normal component. Equation (3.13) is a special case of (3.19) below. I should note that Rao did not include the factor $i(\theta)$ in the loss function (nor have subsequent authors in interpreting his result).

Deficiency. As an alternative measure of insufficiency, Le Cam (1964) defined the *deficiency* of an estimator T in terms of the distance between the joint distribution of the sample Y and the best reconstruction of it based on T followed by a randomization. Letting P_θ and $\tilde{P}_\theta$ denote these two distributions, Skovgaard (1985) derived the inequality

$$n \cdot \sup(P_\theta(B) - \tilde{P}_\theta(B))^2 \le \gamma^2$$

where the supremum is taken over all Borel sets B in the sample space.

Large deviations. Fu (1982) demonstrated a role for statistical curvature in the study of estimation via large deviations (see also Fu and Kass, 1984). Define

$$\beta(T, \varepsilon) = \operatorname{llm} n^{-1} \cdot \log P\{i(\theta)(T - \theta)^2 > \varepsilon\},$$

$$B(\theta, \varepsilon) = \inf\{K(\theta^*, \theta): i(\theta)(\theta^* - \theta)^2 > \varepsilon\},$$

where the infimum is over θ^* and $K(\theta^*, \theta)$ is the Kullback–Leibler number. The quantity $B(\theta, \varepsilon)$, which is often called the Bahadur bound, satisfies $B(\theta, \varepsilon) \ge \beta(T, \varepsilon)$ and thus bounds the exponential rate of convergence of consistent estimators. The difference between B and β may be expanded in a Taylor series in ε. Carrying out the expansion, Fu (1982) found that for efficient estimators the first three terms (involving ε, ε^2, and ε^3) vanish and

$$\lim_{\varepsilon \to 0} \varepsilon^{-4}(B(\theta, \varepsilon) - \beta(T, \varepsilon)) \ge \frac{1}{8}\gamma^2$$

with equality holding when T is the MLE. (Fu actually did not use $i(\theta)$ in the definitions of β and B, so $i(\theta)$ appeared instead in his result.)

The Fisher scoring algorithm. The Fisher scoring algorithm is an iterative procedure for computing the MLE when an explicit analytical expression is unavailable. Its sequence of iterates has the form

$$\theta_{k+1} = \theta_k + G(\theta_k)^{-1} D_\theta l_y[\theta_k] \tag{3.15}$$

with the particular choice $G(\theta_k) = ni(\theta_k)$. In general, an algorithm producing a sequence defined by (3.15) for some G is called Newton-like (or pseudo-Newton or sometimes quasi-Newton) because when $G(\theta) = -D_\theta^2 l_y[\theta]$, (3.15) defines Newton's method.

Newton-like alternatives to Newton's method are used to avoid computation of a second derivative or to improve global behavior. In doing so, they may fail to provide local convergence (though in practice this can usually be fixed) and they sacrifice some finishing speed: both convergence and limiting rate of convergence are governed by the quantity

$$\lambda = |\, G(\hat\theta)^{-1}(I_y(\hat\theta) - G(\hat\theta))\,|.$$

When $0 < \lambda < 1$, the sequence (3.15) is locally convergent (i.e., converges to a root of the likelihood equations for any sufficiently close starting value) with limiting rate of convergence

$$\lim_{k\to\infty} \frac{|\,\theta_{k+1} - \hat\theta\,|}{|\,\theta_k - \hat\theta\,|} = \lambda.$$

When $\lambda \geq 1$, however, (3.15) is not locally convergent (see Ortega and Rheinboldt, 1970).

In the case of the scoring algorithm, λ takes an interesting form (Kass, 1983; Smyth, 1987). From (3.11), for a curved exponential family we have

$$\lambda = \gamma(\hat\theta) \cdot |\,(\bar y - \mu(\hat\theta))^T \hat n\,|$$

where $\hat n$ is the unit normal vector in the plane spanned by $D\eta$ and $D^2\eta$ at $\eta(\hat\theta)$. Whether or not the scoring algorithm converges and, if so, how fast, thus depends on two quantities: the statistical curvature of the model and the component of the residual in the first and second derivative plane (the osculating plane).

3.5 Generalizations

To generalize the results, multiparameter curved exponential families, together with auxiliary spaces associated with estimators, are needed and then the score function expansion may be carried out as before. If a regular k-dimensional exponential family $\mathscr{E}$ is given the structure of a smooth manifold, then a subfamily $\mathscr{E}_0$ is an *imbedded exponential subfamily* or a *curved exponential family* if it forms an imbedded submanifold of the full family. This can also be described in terms of a parameter space Θ as in Section 1.2.3. Estimators now must map into Θ in R^m and to be regular they must be of full rank m (in addition to being smooth) on some neighborhood of $\mu(\theta_0)$ in the family's representation M_0 within the mean-value parameter space M. The existence of an auxiliary space, and a coordinate decomposition (θ, α), in a neighborhood of $\mu(\theta_0)$ is again ensured.

3.5.1 Scalar Curvatures and Information Loss

The presentation of results on information loss and recovery in Section 3.3 used limiting *relative* information loss, since this quantity is invariant to reparameterization. As a multidimensional analog, I will consider the limiting *scalar relative information loss* of an estimator T, defined to be the limit of the trace of the matrix $i(\theta)^{-1}[ni(\theta) - i^T(\theta)]$. Since both $i(\theta)$ and $i^T(\theta)$ transform according to the change of variables formula (e.g., $i(\xi) = J^T i(\theta) J$, where J is the Jacobian matrix $D_\xi\theta$), the scalar relative information loss is also invariant.

I will not bother to write out the score function expansion in terms of (θ, α), which generalizes (3.8), but some explanation of notation is needed: g^{ij} is the (i, j)-component of the inverse of the information matrix for (θ, α) at $\alpha = 0$; following Amari, I will let Latin indices a, b, c, d refer to components of θ, while Greek indices $\tau, \delta, \kappa, \lambda$ will refer to components of α. The generalization of (3.6) then becomes

$$\lim \operatorname{tr}(i(\theta)^{-1}[ni(\theta) - i^T(\theta)]) = \gamma^2 + (\tfrac{1}{2})\beta^2 \tag{3.16}$$

in which

$$\gamma^2 = \sum_{a,b,c,d} g^{ac} g^{bd} \cdot \langle (\partial_{ab}\eta)_N, (\partial_{cd}\eta)_N \rangle_\eta, \tag{3.17}$$

$$\beta^2 = \sum_{\tau,\delta,\kappa,\lambda} g^{\tau\kappa} g^{\delta\lambda} \cdot \langle (\partial_{\tau\delta}\mu)_N, (\partial_{\kappa\lambda}\mu)_N \rangle_\mu \tag{3.18}$$

where β^2 is exactly as in (3.7) and γ^2 is analogous, with the normal components $(\partial_{ab}\eta)_N$ and $(\partial_{cd}\eta)_N$ and their inner product being taken with respect to the information inner product at (θ, α) for $\alpha = 0$.

The statistical interpretation of the result is as before: the scalar information loss is minimized for any estimator having zero normal component of $\partial_{\tau\delta}\mu$ for all τ and δ; in particular, it is minimized for the MLE. The quantity γ^2, which may again be called the *statistical curvature*, is a measure of the insufficiency of the MLE.

3.5.2 Asymptotic Risk and Bias

Other results generalize similarly. As a loss function for the multidimensional problem, it is simplest to use $L(T, \theta) = (T - \theta)^T i(\theta)(T - \theta)$. The limiting risk then satisfies

$$\lim n \cdot E(L(T, \theta)) = m$$

whenever T is efficient, and then, with $\tilde{T}$ being the bias-corrected version of T, we have

$$\lim n^2 \cdot \left[E(L(\tilde{T}, \theta)) - \frac{m}{n}\right] = \gamma^2 + \beta^2 + \omega^2 \tag{3.19}$$

where

$$\omega^2 = \sum_{a,b,c,d} g^{ac} g^{cd} \langle (\partial_{ab}\mu)_T, (\partial_{cd}\mu)_T \rangle_{\mu(\theta)}. \tag{3.20}$$

The right-hand side of (3.19) is also equal to the limiting trace (minus m) of the covariance matrix of $\tilde{T}$ relative to the first-order covariance $[ni(\theta)]^{-1}$, i.e., $\lim n \cdot [\mathrm{tr}(ni(\theta) \cdot \mathrm{Var}(\tilde{T})) - m]$.

The quantity ω^2 is defined analogously to γ^2 with the tangential component of $\partial_{ab}\mu$ replacing the normal component of $\partial_{ab}\eta$. It is invariant to affine transformations of the parameter space but, unlike γ^2, ω^2 is not invariant to more general changes of parameterization.

By permuting the indices in the inverse information matrices appearing in (3.17), we obtain an alternative reduction of the three-way array $(\partial_{ab}\eta)_N$,

$$m^2\bar{\gamma}^2 = \sum_{a,b,c,d} g^{ab} g^{cd} \langle (\partial_{ab}\eta)_N, (\partial_{cd}\eta)_N \rangle$$

which will be interpreted in Section 3.5.5. A quantity analogous to $\bar{\gamma}^2$ may also be defined,

$$m^2\bar{\omega}^2 = \sum_{a,b,c,d} g^{ab} g^{cd} \langle (\partial_{ab}\mu)_T, (\partial_{cd}\mu)_T \rangle_{\mu(\theta)},$$

and this is directly related to the asymptotic relative bias of the MLE (or any other estimator having minimal information loss),

$$\begin{aligned} &\lim (E(\hat{\theta} - \theta))^T (ni(\theta)) (E(\hat{\theta} - \theta)) \\ &\quad = (\tfrac{1}{4}) m^2 \bar{\omega}^2. \end{aligned} \tag{3.21}$$

Equations (3.19) and (3.21) follow from the matrix (tensorial) form of the results given by Amari (1985, equations (5.4), (5.11), (7.11)).

Finally, I should note that Kumon and Amari (1988) have found that γ^2 and $\bar{\gamma}^2$ play a role in determining the loss of power of efficient tests. Readers should be aware, however, that those authors use slightly different notation and terminology.

3.5.3 Statistical Curvature in General Families

It is also straightforward to generalize Efron's formula for curvature in a general parametric family. Writing l for the loglikelihood from a single observation and

$$\kappa_{a,bc} = \mathrm{Cov}(\partial_a l, \partial_{bc} l),$$

$$\kappa_{ab,cd} = \mathrm{Cov}(\partial_{ab} l, \partial_{cd} l)$$

and continuing to use g^{ab} to denote the (a, b)-component of the inverse of the Fisher information matrix, we have

$$\gamma^2 = \sum_{a,b,c,d} g^{ac} g^{bd} \left(\kappa_{ab,cd} - \sum_{i,j} g^{ij} \kappa_{ab,i} \kappa_{cd,j} \right).$$

Meanwhile, Amari (1987) has given a very nice construction (of an appropriate "Hilbert bundle") to show how curved exponential families "approximate" general families, thereby formalizing a suggestion of Efron (1975) described here in Section 3.4.1.

3.5.4 The Use of Scalar Curvatures

Information loss could be studied in its matrix (tensor) form $\lim [ni(\theta) - i^T(\theta)]$, as could asymptotic variance. This is the way Amari, following Reeds (1975) and Madsen (1979), has chosen to work. There is, of course, more detail in the matrix than in its invariant trace, so results in matrix form produce those in scalar form as corollaries. Though the tensorial form opens the door to investigation of directions of information loss, I have not seen any research along these lines. (The "directions" would not correspond to interesting combinations of parameters, but rather alternative one-parameter subfamilies; directional loss functions would seem more promising.) I have preferred to state results in terms of scalars mainly because I find them simpler, the invariance of a number with respect to reparameterization being more immediate than the remark that an array defines a tensor. In addition, the scalar curvatures appearing above will be compared with nonlinearity diagnostics in Section 3.5.6.

3.5.5 Interpretation in Nonlinear Regression

The geometrical interpretation of γ is partly apparent already, but more can be said. As in the one-parameter case (and by definition), all of the curvature information is contained in the normal components of the second derivatives. The question is how this three-way array may be usefully summarized.

Consider the special case of nonlinear regression, that is, $Y \sim N(\eta, \sigma^2 I_k)$ with $\eta = \eta(\theta) \in N_0$, where I_k is the k-dimensional identity and N_0 is an m-dimensional imbedded submanifold of $N = R^k$. Here, $N = M$ and the information inner product is the usual Euclidean inner product except for a factor of σ^{-2}. An especially simple form for γ^2 at a point $\eta_0 \in N_0$ occurs if we assume that at η_0 the parameterization θ of N_0 is orthonormal with respect to the Euclidean inner product on the space to N_0 at η_0; this means that $g_{ab} = \langle \partial_a \eta, \partial_b \eta \rangle_{\eta_0} = \sigma^{-2} \langle \partial_a \eta, \partial_b \eta \rangle^E = \sigma^{-2} \cdot \delta_{ab}$ (σ^{-2} if $a = b$, 0 otherwise) where $\langle \cdot, \cdot \rangle^E$ is the Euclidean inner product. Then, letting h^j_{ab} be the jth component of the vector $(\partial_{ab}\eta)_N$ we have

$$\gamma^2 = \sigma^2 \cdot \sum_{a,b} \sum_j h^j_{ab} h^j_{ab}. \tag{3.22}$$

In this form, it is apparent that γ^2 is σ^2 times a squared magnitude of the three-way array of normal components of second derivatives. If h were a vector having components h^i with respect to the usual basis, its squared Euclidean magnitude would be $\sum_i h^i h^i$; if h were a symmetric matrix with components h^j_i, a commonly-used squared magnitude would be $\text{tr}(hh^T) = \sum_{i,j} h^j_i h^j_i$. Furthermore, since the normal components of the tangent vectors $\partial_a \eta$ are zero, $(\partial_{ab}\eta)_N$ satisfies the same transformation expression as the information matrix when a change of parameters is made. Thus, the expression (3.17) for γ^2 is invariant with respect to reparameterization; it provides the general expression for this "squared magnitude" just as $\sum g_{ij} h^i h^j$ provides the general expression for the squared length of a vector.

There is another invariant reduction of the normal second-derivative array, which plays an important role, along with γ^2, in the geometry of surfaces. Just as γ^2 was considered above to be a generalization of σ^2 times $\text{tr}(hh^T)$ when h was a matrix, this other scalar quantity is a generalization of σ^2 times $\text{tr}(h)^2$. In the orthogonal case of (3.22), it becomes

$$(3.23) \qquad m^2\bar{\gamma}^2 = \sigma^2 \cdot \sum_j \left(\sum_a h^j_{aa}\right)\left(\sum_a h^j_{aa}\right)$$

which is a special case of $m^2\bar{\gamma}^2$ defined in Section 3.5.2. Here, $\bar{\gamma}^2$ is the squared length of the *mean curvature vector* of the surface. The difference between these two quantities

$$(3.24) \qquad r = m^2\bar{\gamma}^2 - \gamma^2$$

is the Riemannian *scalar curvature* of the surface. Both $\bar{\gamma}$ and r are fundamental in the study of submanifolds in Riemannian geometry.

Thus, in the nonlinear regression context, the statistical curvature γ^2 may be interpreted as a squared magnitude of the array of normal components of the second derivative of η (with respect to the information metric or, equivalently, σ^2 times the squared magnitude with respect to the Euclidean metric); it is one of two basic scalar invariant summaries of this array.

3.5.6 Curvature Measures in Data Analysis

The discussion of curvature throughout Sections 3.3 through 3.5 has involved statistical theory rather than methodology. In nonlinear regression, however, various measures of curvature have been proposed to help diagnose poor performance of asymptotic approximations due to the nonlinearity of the regression surface $\eta(\Theta)$. Following Beale (1960) and Bates and Watts (1980), these are usually based on summaries of the second derivative array $D^2\eta(\hat{\theta})$, which would have all entries zero in the linear case. There arises then the question of how these measures of nonlinearity are related to the scalar summaries discussed earlier.

The nonlinearity measures of both Beale and Bates and Watts are based on normal and tangential components of the second derivatives of curves $c_v(t) = \eta(\hat{\theta} + tv)$ at $c(0) = \eta(\hat{\theta})$, where v is a vector in R^m, which are sometimes called "lifted lines." By analogy with ordinary curvature, as in (1.2), the normalized length

$$\kappa_N(v) = \| c_v'(0) \|^{-2} \cdot \| (c_v''(0))_N \|$$

is then taken as a measure of curvature of the surface at $\eta(\hat{\theta})$ in the direction of the tangent vector $c_v'(0) = D\eta(\hat{\theta})v$. Here the inner product is Euclidean, but note that with respect to the information inner product, the squared length of v is $\langle v, v \rangle_{\eta(\hat{\theta})} = \sigma^{-2} \cdot \| c_v'(0) \|^2$.

Bates and Watts used the maximal value of $(\kappa_N(v))^2$, over vectors v, while Beale used the integrated average of $(\kappa_N(v))^2$, over the sphere $\{v: \| c_v'(0) \| = 1\}$. (In each case, these numbers were multiplied by a convenient constant.) These measures are invariant with respect to the parameterization θ used to calculate them.

The quantity

$$\kappa_T(v) = \| c_v'(0) \|^{-2} \cdot \| (c_v''(0))_T \|,$$

analogous to $\kappa_N(v)$, is used in the same way (by finding the maximum or by averaging) by these authors. Its properties make it suitable for defining the effects of parameterization: measures of nonlinearity based on $\kappa_T(\cdot)$ are affine-invariant, that is, invariant with respect to affine transformations of the parameter space, but not generally invariant. Affine invariance is desirable geometrically because affine transformations preserve linearity (or nonlinearity) and it is desirable statistically because they preserve Normality.

Authors subsequent to Bates and Watts (1980) have used their notation and terminology for the second-derivative array after suitable affine transformation of the coordinates. Let ξ coordinates be obtained from the usual Euclidean coordinates for R^k by first translating so that the new origin is at $\eta(\hat{\theta})$, then rotating so that the first m coordinates and last $k - m$ coordinates are, respectively, tangent to and normal to the surface $\eta(\Theta)$ at $\eta(\hat{\theta})$; ϕ-coordinates for the surface are then obtained from a linear transformation of θ, so that the first derivatives of ξ (or η) with respect to ϕ are orthonormal tangent vectors. Then

$$a_{ijk} = s \cdot m^{1/2} \cdot \frac{\partial \xi_k}{\partial \phi_i \partial \phi_j}$$

where s is the sample standard deviation. The $m \times m \times m$ array of tangential components is A^T and the $m \times m \times (k - m)$ array of normal components is A^N. Bates and Watts called the associated curvatures "parameter-effects" and "intrinsic" curvatures. I prefer

here to call the latter "imbedding" curvatures, because the conventional differential geometry terminology reserves the label "intrinsic" for properties that do not depend on the imbedding in R^k.

Using the A^N array, the statistical curvature in (3.17) evaluated at the least-squares estimate becomes

$$\gamma^2 = c\sigma^2 \cdot \sum_{b,c,\lambda} (a_{bc\lambda})^2$$

where $c^{-1/2} = s \cdot m^{1/2}$ is the constant scaling factor used by Bates and Watts, and λ is summed over the $k - m$ normal-component indices. Meanwhile, as noted in Section 3.5.5, the squared length $\bar{\gamma}^2$ of the mean curvature at $\hat{\theta}$ is

$$m^2\bar{\gamma}^2 = c\sigma^2 \cdot \sum_{\lambda} \left(\sum_{b} a_{bb\lambda} \right)^2$$

where λ is again summed over the normal-component indices. The two measures could both be used to assess nonlinearity and, in fact, the spherical mean squared curvature of Beale (in terms of his quantity N_ϕ) is

$$4 \cdot N_\phi = m^2\bar{\gamma}^2 + 2\gamma^2$$

where $\bar{\gamma}$ and γ are evaluated at the least-squares estimate. This follows immediately from the definitions of $\bar{\gamma}^2$ and γ^2 together with equation (2.29) of Bates and Watts (1980).

By analogy, as in Section 3.5.2, the tangential components may be summarized by ω^2 or $\bar{\omega}^2$, using the A^T array,

$$\omega^2 = c\sigma^2 \cdot \sum_{b,c,d} (a_{bcd})^2,$$

$$m^2\bar{\omega}^2 = c\sigma^2 \cdot \sum_{d} \left(\sum_{b} a_{bbd} \right)^2$$

where d is summed over the m tangential-component indices. The tangential spherical mean squared curvature of Beale (using his symbols N_θ and N_ϕ) becomes

$$4 \cdot (N_\theta - N_\phi) = m^2\bar{\omega}^2 + 2\omega^2$$

where $\bar{\omega}$ and ω are evaluated at the least-squares estimate.

By virtue of the results in Sections 3.3.2, 3.4.2 and 3.5.2, these simple relationships may be considered additional motivation for the curvature measures proposed by Beale (and, less directly, for those of Bates and Watts). Other connections to inferential quantities were provided by Hougaard (1985), who gave bounds on the magnitude of the bias and skewness of the least-squares estimator, and on the magnitude of the expected third derivative of the loglikelihood, in terms of parameter-effect curvatures. (See also Cook, Tsai and Wei, 1986, for further discussion of bias.) A correction factor for likelihood-based confidence regions based on imbedding curvature was proposed by Beale (1960). Further work in this vein was undertaken by Hamilton, Watts and Bates (1982), while McCullagh and Cox (1986) gave some geometrical interpretation of Bartlett's correction, showing that it involves γ^2 and $\bar{\gamma}^2$, and contrasting it with Beale's.

The malady that is supposed to be diagnosed by curvature measures, poor performance of asymptotic approximations, is not limited to nonlinear regression. One might hope that analogous measures could be developed for other problems. One possibility would be to use the α-connections, mentioned in Section 3.6, to define more general measures.

I must comment, however, that for an investigator who takes a likelihood or Bayesian approach to inference, these curvature measures have little direct relevance to the analysis of a given set of data. If we assume a quadratic approximation to the loglikelihood function will form the basis for inference, then, for a diagnostic to be convincing in its data analytical use, a decrease in its value would have to indicate an improvement in inferences based on the quadratic approximation to the loglikelihood for the data at hand. Summaries of the third derivatives evaluated at $\hat{\theta}$ become conspicuous candidates for this purpose. Letting $\tilde{g}^{ab}$ be the (a, b)-component of the inverse of observed information and evaluating the third derivatives at $\hat{\theta}$, the quantities

$$B^2 = \sum_{a,b,c,d,e,f} \tilde{g}^{ad}\tilde{g}^{be}\tilde{g}^{cf} \cdot \partial_{abc}l \cdot \partial_{def}l,$$

$$m^2\bar{B}^2 = \sum_{a,b,c,d,e,f} \tilde{g}^{ab}\tilde{g}^{de}\tilde{g}^{cf} \cdot \partial_{abc}l \cdot \partial_{def}l$$

are invariant with respect to affine transformations. This makes them suitable for comparing alternative parameterizations. In the Bayesian approach, the log posterior $\tilde{l}$ may be substituted for the loglikelihood, and the posterior mode $\tilde{\theta}$ may be used as the point at which the derivatives are evaluated, with $\tilde{g}^{ab}$ becoming the (a, b)-component of $-D^2\tilde{l}(\tilde{\theta})^{-1}$. The measures B and $\bar{B}$ would then be used to diagnose inaccuracy of the "modal approximation" to the posterior, which is the Normal distribution centered at $\tilde{\theta}$ with variance $-D^2\tilde{l}(\tilde{\theta})^{-1}$. (That is, they diagnose inaccuracy of the quadratic approximation to $\tilde{l}$.) B and $\bar{B}$ are applicable to general models.

A little manipulation shows that the quantities B and $\bar{B}$ bear a formal resemblance to ω and $\bar{\omega}$, with $\partial_{abc}l$ replacing $((\partial_{ab}\mu)_T)_c$ and $\tilde{g}^{ab}$ replacing g^{ab}. In fact, there is an interesting Bayesian analog to the expression for the relative bias of the MLE given in (3.21). Letting $\bar{\theta} = E(\theta \mid y)$ be the posterior mean of θ, I will call $\tilde{\theta} - \bar{\theta}$ the *posterior bias* of the mode $\tilde{\theta}$. For diagnostic purposes, it is a Bayesian counterpart of the bias of the MLE $\hat{\theta} - \theta$ in the sense that it indicates inaccuracy in the centering of the Normal distribution used for approximate inference. From an order $O(n^{-2})$

expansion of $\bar{\theta}$, as in equation (2.7) of Kass, Tierney and Kadane (1988), the relative posterior bias of the mode (relative to the posterior information matrix $-D^2\tilde{l}(\tilde{\theta})$) is

$$(\tilde{\theta} - \bar{\theta})^T(-D^2\tilde{l}(\tilde{\theta}))(\tilde{\theta} - \bar{\theta}) \doteq (1/4)m^2\bar{B}^2$$

with an error of order $O(n^{-2})$. That is, the role of $\bar{B}^2$ in the leading term of the asymptotic expansion of the relative posterior bias of the mode is exactly analogous to the role of $\tilde{\omega}^2$ in the leading term of the asymptotic expansion of the relative bias of the MLE. Although B and $\bar{B}$ have not, to my knowledge, been studied, a related diagnostic was discussed by Jennings (1986) in the context of logistic regression.

3.6 The Role of Affine Connections

I have alluded to the non-Riemannian geometrical foundation for the curvature measures γ and β. The reason such a structure is required may be seen by considering the trinomial model, while recalling that the statistical curvature of a one-parameter subfamily $\mathscr{C}_0$ is $\| D_\theta\eta \|^{-2} \cdot \| (D_\theta^2\eta)_N \|$ where the information inner product defines the normal component and the norm. The quantity $\| D_\theta\eta \|$ is easily understood using the information metric: $D_\theta\eta$ is a tangent vector written in terms of η; the choice of η is irrelevant in the sense that for any other parameterization ξ of the trinomial we have $\| D_\theta\xi \|_{\xi_0} = \| D_\theta\eta \|_{\eta_0}$, where $\xi_0 = \xi(\theta_0)$ and $\eta_0 = \eta(\theta_0)$ (see (3.3)). In contrast, the length of the normal component of the second derivative depends on the parameterization of the ambient space; for instance, $\| D_\theta\mu \|^{-2} \cdot \| (D_\theta^2\mu)_N \|$ is a mixture curvature, which is distinct from the statistical curvature. Furthermore, if we were to define curvature of curves relative to the information geometry, then curves with zero curvature would have to correspond to great circles on the sphere representing that geometry, given by (2.2). That would be inconsistent with statistical curvature: the Hardy–Weinberg model is an exponential family, having zero statistical curvature, but does not form a great circle on the sphere (2.2). This is an example of the way statistical curvature uses a derivative calculation that is not naturally compatible with the information metric.

The appropriate foundation for the curvature calculations given here involves a pair of *affine connections*, the *exponential connection*, denoted by $\overset{e}{\nabla}$, and the *mixture connection*, denoted by $\overset{m}{\nabla}$. In general, affine connections determine the calculation of derivatives (of vector and tensor fields). Within exponential families, curvatures calculated using $\overset{e}{\nabla}$ correspond to the use of the natural parameter space, while curvatures calculated using $\overset{m}{\nabla}$ correspond to the use of the mean-value parameter space. Thus, the formulas of Section 3.5 did not explicitly use these connections, but alternative versions would (e.g., as in Amari's work).

The exponential and mixture connections were identified by Dawid in his discussion of Efron's 1975 paper. A one-parameter family of connections based on these, called the α-connections was introduced by Amari (1982a). They had been discussed previously by Centsov (1972) in the discrete case and have been studied in some detail by Amari (1985, 1987) and Lauritzen (1987); see also Kass (1984) for a comment on their interpretation in terms of various parameterizations. The resulting curvature formulas for submanifolds have been derived by Vos (1989); these show, in particular, that the relationship of γ and $\bar{\gamma}$ to the Riemannian scalar curvature, mentioned for nonlinear regression models in Section 3.5.5, requires modification for general curved exponential families. This is one of many ways that the α-connection geometries are subtle.

3.7 Related Work

In his 1975 paper, Efron noted a role for statistical curvature in hypothesis testing: the locally most powerful test tends to perform more poorly as the curvature increases (see also the comments to that paper by Ghosh, 1975, and by Pfanzagl, 1975). This subject was analyzed in greater depth by Amari (1985) and Kumon and Amari (1983, 1988).

Kumon and Amari (1984) studied problems involving infinitely many nuisance parameters (see also Amari 1985, 1987a). The same authors (Amari and Kumon, 1983; Amari, 1985) also developed the use of tensorial Hermite polynomials in Edgeworth series for the distribution of (T, A), and related distributions, where A is an auxiliary statistic associated with the estimator T (as described in Section 3.3). Related, though rather different, is work on expansions by Barndorff-Nielsen and his colleagues, which begins (Barndorff-Nielsen, 1986a, 1987a) with the introduction of a Riemannian metric based on observed information rather than expected information, and has led to a generalization of tensors called *strings* (Barndorff-Nielsen, 1986b; Murray, 1988).

Eguchi (1983) used geometrical analysis to show that many of the good asymptotic properties of the MLE are shared by certain minimum contrast estimators. He also showed (Eguchi, 1984) that, in curved exponential families, estimators are second-order efficient if and only if they minimize the expected length of the residual vector (with respect to the information inner product). Working within the setting of curved exponential families, Moolgavkar and Venzon (1987a, b) proposed and studied confidence intervals defined in terms of the Riemannian geometry of the family inherited from the ambient exponential family, based on the information inner product at the MLE. Vos (1988) used geometrical methods to discuss influential observations in exponential-family regression models.

4. CLOSING COMMENTS

In this paper, the geometry of asymptotic inference has been motivated, primarily, as an attempt to better understand Jeffreys and Fisher. There is an obvious historical interest in doing so, and it is worth continuing to examine Jeffreys' and Fisher's theories for aspects that remain relevant to current views of the discipline. The development here, however, should not be construed as an accurate account of the evolution of ideas in the literature. In particular, relatively little of the published work on information metric Riemannian geometry evolved from Jeffreys' use of it; most appears to have roots in the paper by Rao (1945). Meanwhile, the "main result" of Efron (1975) seems widely considered to have been the geometrical interpretation not of Fisher's definition of information loss, but rather of second-order efficiency, as defined by Rao (1963). (This was apparently the point of view of Efron, and was the chief concern in the generalizations by Reeds (1975) and Madsen (1979); Amari has given considerable attention to information loss, however.) In addition, my narrow focus has surely not done justice to the depth and breadth of the research in this area. I recommend Barndorff-Nielsen, Cox and Reid (1986) and the recent IMS monograph (Amari, Barndorff-Nielsen, Kass, Lauritzen and Rao, 1987) as starting points for further reading. A rather different line of interesting recent research using geometrical methods in inference is reviewed by Johansen and Johnstone (1989).

I have also emphasized invariance. We often think of geometrical arguments as being driven by pictorial representations. The relations among objects that may be represented in pictures, however, are those that do not require coordinate description. Although it is not easy to say precisely what constitutes geometry within statistics (indeed Chern (1979) has noted that it is not easy to say what constitutes geometry within mathematics), in the most fundamental sense used here geometrical results are those that do not necessitate a choice of parameterization. Their geometrical nature is most apparent when they are stated in terms of invariants as in Section 3.5, or with the use of coordinate-free symbolism, in the manner indicated in parts of Sections 2.3 and 3.6.

In presenting the material here, I have freely added a number of critical remarks and made both implicit and explicit suggestions for future research. As should be apparent, I find the geometrical arguments reviewed here to be of interest first and foremost as an aid in understanding basic statistical theory, and many of them can be readily appreciated without detailed knowledge of differential geometry. While differential geometry provides a helpful perspective on asymptotic theory for regular, parametric families, I do not yet perceive a basis for a claim that differential geometrical research has made inroads into a large class of problems that is otherwise unreachable. On the other hand, I believe that the methods are so powerful, and the connections with statistics so plausible, that some further developments, of great methodological importance, might well occur. Exposure to the ideas outlined in this paper, and available in the literature, may accelerate that process.

ACKNOWLEDGMENTS

My work on the revision of this paper was supported by the National Science Foundation, under Grant DMS-87-05646, while I was a visiting scholar in the Department of Statistics, Harvard University. I thank the referees and the former Executive Editor, Morris DeGroot, for their many helpful comments and my hosts at Harvard for their hospitality.

REFERENCES

AMARI, S.-I. (1982a). Differential geometry of curved exponential families—curvature and information loss. *Ann. Statist.* **10** 357–387.

AMARI, S.-I. (1982b). Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika* **69** 1–17.

AMARI, S.-I. (1985). *Differential-Geometrical Methods in Statistics. Lecture Notes in Statist.* **28**. Springer, New York.

AMARI, S.-I. (1987a). Differential geometrical theory of statistics. In *Differential Geometry in Statistical Inference* 19–94. IMS, Hayward, Calif.

AMARI, S.-I., BARNDORFF-NIELSEN, O. E., KASS, R. E., LAURITZEN, S. L. and RAO, C. R. (1987). *Differential Geometry in Statistical Inference*. IMS, Hayward, Calif.

AMARI, S. and KUMON, M. (1983). Differential geometry of Edgeworth expansions in curved exponential families. *Ann. Inst. Statist. Math.* **35** 1–24.

ATKINSON, C. and MITCHELL, A. F. S. (1981). Rao's distance measure. *Sankhyā Ser. A* **43** 345–365.

BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

BARNDORFF-NIELSEN, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

BARNDORFF-NIELSEN, O. E. (1986a). Likelihood and observed geometries. *Ann. Statist.* **14** 856–873.

BARNDORFF-NIELSEN, O. E. (1986b). Strings, tensorial combinants, and Bartlett adjustments. *Proc. Roy. Soc. London Ser. A* **406** 127–137.

BARNDORFF-NIELSEN, O. E. (1987a). Differential and integral geometry in statistical inference. In *Differential Geometry in Statistical Inference* 95–161. IMS, Hayward, Calif.

BARNDORFF-NIELSEN, O. E. and BLÆSILD, P. (1987a). Strings: Mathematical theory and statistical examples. *Proc. Roy. Soc. London Ser. A* **411** 155–176.

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

BATES, D. M. and WATTS, D. G. (1980). Relative curvature measures of nonlinearity (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 1–25.

BEALE, E. M. L. (1960). Confidence regions in non-linear estimation (with discussion). *J. Roy. Statist. Soc. Ser. B* **22** 41–88.

BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463.

BHATTACHARYYA, A. (1943). On discrimination and divergence. *Proc. 29th Indian Sci. Cong.*, Part III, 13.

BHATTACHARYYA, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā* **7** 401–406.

BIRCH, M. W. (1964). A new proof of the Pearson-Fisher theorem. *Ann. Math. Statist.* **35** 817–824.

BISHOP, Y., FIENBERG, S. E. and HOLLAND, P. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

BOOTHBY, W. M. (1975). *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic, New York.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decison Theory*. IMS, Hayward, Calif.

BURBEA, J. (1986). Informative geometry of probability spaces. *Exposition. Math.* **4** 347–378.

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

BURBEA, J. and RAO, C. R. (1982b). Differential metric in probability spaces. *Probab. Math. Statist.* **3** 115–132.

CENTSOV, N. N. (1972). *Statistical Decision Rules and Optimal Inference*. Nauka, Moscow. (In Russian.) English translation (1982) *Trans. Math. Monographs* **53**. Amer. Math. Soc., Providence, R.I.

CHERN, S.-S. (1979). From triangles to manifolds. *Amer. Math. Monthly* **86** 339–349.

COOK, R. D., TSAI, C.-L. and WEI, B. C. (1986). Bias in nonlinear regression. *Biometrika* **73** 615–623.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

DAWID, A. P. (1975). Discussion of "Defining the curvature of a statistical problem (with applications to second-order efficiency)," by B. Efron. *Ann. Statist.* **3** 1231–1234.

DAWID, A. P. (1977). Further comments on some comments on a paper by Bradley Efron. *Ann. Statist.* **5** 1249.

DUDLEY, R. M. (1979). On χ^2 tests of composite hypotheses. In *Probability Theory. Banach Center Publ.* **5** 75–87. PWN-Polish Scientific Publishers, Warsaw.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second-order efficiency) (with discussion). *Ann. Statist.* **3** 1189–1242.

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

EFRON, B. (1982). Maximum likelihood and decision theory. *Ann. Statist.* **10** 340–356.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–487.

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

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

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Royal Soc. London Ser. A* **222** 309–368.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. London Ser. A* **144** 285–307.

FISHER, R. A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.* **98** 39–54.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*, 3rd ed. Hafner, New York.

FU, J. C. (1982). Large sample point estimation: A large deviation theory approach. *Ann. Statist.* **10** 762–777.

FU, J. C. and KASS, R. E. (1984). A note on the interpretation of the Bahadur bound and the rate of convergence of the maximum likelihood estimator. *Statist. Probab. Lett.* **2** 269–273.

GHOSH, J. K. (1975). Discussion of "Defining the curvature of a statistical problem (with applications to second-order efficiency), by B. Efron. *Ann. Statist.* **3** 1224–1226.

GHOSH, J. K., SINHA, B. K. and WIEAND, H. S. (1980). Second order efficiency of mle with respect to any bounded bowl-shaped loss function. *Ann. Statist.* **8** 506–521.

GHOSH, J. K. and SUBRAMANYAM, K. (1974). Second-order efficiency of maximum likelihood estimators. *Sankhyā Ser. A.* **36** 325–358.

GOOD, I. J. (1969). What is the use of a distribution? In *Multivariate Analysis II*, (P. R. Krishnaiah, ed.) 183–203. Academic, New York.

HAMILTON, D. C., WATTS, D. G. and BATES, D. M. (1982). Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions. *Ann. Statist.* **10** 386–393.

HODGES, J. L., JR. and LEHMANN, E. L. (1970). Deficiency. *Ann. Math. Statist.* **41** 783–801.

HOLLAND, P. W. (1973). Covariance stabilizing transformations. *Ann. Statist.* **1** 84–92.

HOUGAARD, P. (1985). The appropriateness of the asymptotic distribution in a nonlinear regression model in relation to curvature. *J. Roy. Statist. Soc. Ser. B* **47** 103–114.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A* **196** 453–461.

JEFFREYS, H. (1955). The present position in probability theory. *Brit. J. Philos. Sci.* **5** 275–289.

JEFFREYS, H. (1957). *Scientific Inference*, 2nd ed. Cambridge Univ. Press, Cambridge.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press, Oxford.

JENNINGS, D. E. (1986). Judging inference adequacy in logistic regression. *J. Amer. Statist. Assoc.* **81** 471–476.

JOHANSEN, S. and JOHNSTONE, I. (1989). Hotelling's theorem on the volumes of tubes: Some illustrations in simultaneous inference and data analysis. To appear.

KASS, R. E. (1980). The Riemannian structure of model spaces. Ph.D. dissertation, Univ. Chicago.

KASS, R. E. (1983). The rate of convergence of the Fisher scoring and Gauss-Newton algorithms. Technical Report No. 284, Carnegie Mellon Univ.

KASS, R. E. (1984). Canonical parameterizations and zero parameter-effects curvature. *J. Roy. Statist. Soc. Ser. B* **46** 86–92.

KASS, R. E. (1987). Introduction. In *Differential Geometry in Statistical Inference* 1–17. IMS, Hayward, Calif.

KASS, R. E. (1988). Data translated likelihood and Jeffreys' rules. Technical Report No. 439, Carnegie Mellon Univ.

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1988). Asymptotics in Bayesian computations (with discussion). In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 263–278. Oxford Univ. Press, Oxford.

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1989). Approximate methods for inference and sensitivity in Bayesian analysis. *Biometrika* **76**. To appear.

KOBAYASHI, S. and NOMIZU, K. (1969). *Foundations of Differential Geometry*. Interscience, New York.

KUMON, M. and AMARI, S.-I. (1983). Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference. *Proc. Roy. Soc. London Ser. A* **387** 429–458.

KUMON, M. and AMARI, S.-I. (1984). Estimation of structural parameter in the presence of a large number of nuisance parameters. *Biometrika* **71** 445–459.

KUMON, M. and AMARI, S.-I. (1988). Differential geometry of testing hypothesis: A higher order asymptotic theory in multiparameter curved exponential family. *J. Fac. Engrg., Univ. Tokyo Ser. B* **39** 241–273.

LAURITZEN, S. L. (1987a). Statistical manifolds. In *Differential Geometry in Statistical Inference* 163–216. IMS, Hayward, Calif.

LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

LINDSAY, B. G. (1988). Blended chi-squared distance: A shortcut to second order efficiency. Technical Report, Dept. Statistics, Pennsylvania State Univ.

MADSEN, L. T. (1979). The geometry of statistical model—a generalization of curvature. Res. Report 79-1, Statist. Res. Unit, Danish Medical Res. Council.

MCCULLAGH, P. and COX, D. R. (1986). Invariants and likelihood ratio statistics. *Ann. Statist.* **14** 1419–1430.

MITCHELL, A. F. S. and KRZANOWSKI, W. J. (1985). The Mahalonobis distance and elliptic distributions. *Biometrika* **92** 464–467.

MOOLGAVKAR, S. H. and VENZON, D. J. (1987a). Confidence regions in curved exponential families: Application to matched case-control and survival studies with general relative risk functions. *Ann. Statist.* **15** 346–359.

MOOLGAVKAR, S. H. and VENZON, D. J. (1987b). Confidence regions for parameters of the proportional hazards model: A simulation study. *Scand. J. Statist.* **14** 43–56.

MURRAY, M. K. (1988). Coordinate systems and Taylor series in statistics. *Proc. Royal Soc. London Ser. A* **415** 445–452.

OLLER, J. M. and CUADRAS, C. M. (1985). Rao's distance for negative multinomial distributions. *Sankhyā Ser. A* **47** 75–83.

ORETGA, J. M. and RHEINBOLDT, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic, New York.

PERKS, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuar.* **73** 285–334.

PFANZAGL, J. (1975). Discussion of "Defining the curvature of a statistical problem (with applications to second-order efficiency)," by B. Efron. *Ann. Statist.* **3** 1226–1228.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

PIERCE, D. A. (1975). Discussion of "Defining the curvature of a statistical problem (with applications to second-order efficiency)," by B. Efron. *Ann. Statist.* **3** 1219–1221.

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 efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 531–545. Univ. California Press.

RAO, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **24** 46–72.

RAO, C. R. (1963). Criteria of estimation in large samples. *Sankhyā Ser. A* **25** 189–206.

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

REEDS, J. (1975). Discussion of "Defining the curvature of a statistical problem (with applications to second-order efficiency)," by B. Efron. *Ann. Statist.* **3** 1234–1238.

REID, N. (1988). Saddlepoint methods and statistical inference (with discussion). *Statist. Sci.* **3** 213–238.

ROGERS, W. H. and TUKEY, J. W. (1972). Understanding some long-tailed symmetrical distributions. *Statist. Neerlandica* **26** 211–226.

SKOVGAARD, I. M. (1985). A second-order investigation of asymptotic ancillarity. *Ann. Statist.* **13** 534–551.

SKOVGAARD, L. T. (1984). A Riemannian geometry of the multivariate normal model. *Scand. J. Statist.* **11** 211–223.

SMYTH, G. K. (1987). Curvature and convergence. Technical Report No. 33, Univ. California, Santa Barbara.

SPIVAK, M. (1979). *A Comprehensive Introduction to Differential Geometry*, 2nd ed. Publish or Perish, Boston.

STEIN, C. (1965). Approximation of improper prior measures by prior probability measures. In *Bernoulli, Bayes, Laplace* (J. Neyman and L. M. Le Cam, eds.) 217–240. Springer, New York.

STOKER, J. J. (1969). *Differential Geometry*. Wiley, New York.

VILLEGAS, C. (1971). On Haar priors. In *Foundations of Statistical Inference* (V. F. Godambe and D. A. Sprott, eds.) 409–414. Holt, Rinehart, and Winston, Toronto.

VILLEGAS, C. (1977). On the representation of ignorance. *J. Amer. Statist. Assoc.* **72** 651–654.

VOS, P. (1988). The geometry of exponential family regression with application to detecting influential cases. Unpublished manuscript.

VOS, P. (1989). Fundamental equations for statistical submanifolds with applications to the Bartlett correction. *Ann. Inst. Statist. Math.* To appear.

YOSHIZAWA, T. (1971). A geometrical interpretation of location and scale parameters. Memo TYH-2, Harvard Univ.