The Annals of Statistics
1978, Vol. 6, No. 2, 362-376

THE GEOMETRY OF EXPONENTIAL FAMILIES

BY BRADLEY EFRON

Stanford University

There are two important spaces connected with every multivariate exponential family, the natural parameter space and the expectation parameter space. We describe some geometric results relating the two. (In the simplest case, that of a normal translation family, the two spaces coincide and the geometry is the familiar Euclidean one.) Maximum likelihood estimation, within one-parameter curved subfamilies of the multivariate family, has two simple and useful geometric interpretations. The geometry also relates to the Fisherian question: to what extent can the Fisher information be replaced by $-\partial^2/\partial\theta^2[\log f_\theta(x)]|_{\theta=\hat{\theta}}$ in the variance bound for $\hat{\theta}$, the maximum likelihood estimator?

1. Introduction. The most interesting and challenging one-parameter statistical problems are those for which there does not exist a one dimensional sufficient statistic. "Curved exponential families," which are the vehicle for the results we present, offer a simple paradigm for such problems, see Efron [5]. Using curved exponential families allows us to graphically picture the maximum likelihood estimation process, and important attendant concepts such as partial sufficiency and ancillarity. The author has found these pictures, which are the main topic of this paper, helpful in understanding finite sample size phenomena, avoiding the all too familiar complete dependence upon asymptotic formulae.

There is an intimate relationship between ordinary Euclidean geometry and the multivariate normal distribution. Analysis of variance and multivariate analysis depend upon this relationship. Exponential families, of which the normal is an example, do not in general enjoy simple Euclidean geometry. A secondary purpose of this article, which is partly expository in nature, is to point out that some Euclidean-like results, particulary those relating to maximum likelihood estimation, hold in all exponential families. Sections 3—7 review results from Barndorff-Nielsen [1], Cobb [2], Efron [5], Fisher [7, 8], Kullback [11], Hoeffding [10], and Simon [15], as well as some new material, including work with David Hinkley [6].

Section 6, based on [6], uses the geometrical approach to discuss an important statistical question: to what extent can the Fisher information be replaced by $-\partial^2/\partial\theta^2[\log f_\theta(x)]|_{\theta=\hat{\theta}}$ in the variance bound for $\hat{\theta}$, the maximum likelihood estimator? A very simple geometric model, based on the "curvature" considerations introduced in [5], and also on earlier work of Fisher [9], gives insight into the

Received September 1976; revised June 1977.

AMS 1970 *subject classification.* 62F10.

Key words and phrases. Curvature, maximum likelihood estimation, Kullback-Leibler distance, duality.

answer. A Monte Carlo study of the Cauchy translation family supports the predictions of this model.

Throughout the paper proofs and special examples are clearly marked as such, inviting omission by readers not interested in the technical details.

2. Exponential families. This section reviews basic facts about exponential families. An exponential family $\mathscr{G}$ is a family of density functions

$$g_\alpha(x) \equiv e^{\alpha' x - \psi(\alpha)}, \quad \alpha \in A \tag{2.1}$$

with respect to some *carrier measure* $\nu(x)$ on the *sample space* $\mathscr{X}$. The *natural parameter space* A consists of all α having the normalizing function

$$e^{\psi(\alpha)} \equiv \int e^{\alpha' x}\, d\nu(x) \tag{2.2}$$

less than infinity. The *sufficient statistic* x takes values in $\mathscr{X}$. It will be convenient here for both A and $\mathscr{X}$ to be subsets of the plane R^2. All of our results, with one exception noted later, extend easily to higher dimensions. To avoid trivialities, we assume that $\mathscr{G}$ does not assign probability one to any line in R^2, i.e., that ν is genuinely two dimensional.

The expectation and covariance of the random vector x,

$$\beta \equiv E_\alpha x, \qquad \Sigma_\alpha \equiv \mathrm{Cov}_\alpha x, \tag{2.3}$$

exist finitely in the interior of the convex set A, and can be obtained by differentiation of ψ: $\beta(i) = \partial/\partial\alpha(i)[\psi(\alpha)]$, $\Sigma_\alpha(i, j) = \partial^2/\partial\alpha(i)\,\partial\alpha(j)[\psi(\alpha)]$. The vector β is the *expectation parameter*. The mapping from α to β is one to one, taking the interior of A into the expectation parameter space B (not necessarily convex, as the example at the end of this section shows). Locally this mapping is expressed by the differential relation

$$d\beta = \Sigma_\alpha\, d\alpha. \tag{2.4}$$

The matrix Σ_α is assumed to be of full rank for all α, which is equivalent to requiring that $\mathscr{G}$ not be reducible to a one-parameter exponential family. Lehmann [11], Chapter 2, is a good reference for these definitions and results.

The expectation parameter β indexes $\mathscr{G}$ just as well as does α. We will write g_β, Σ_β, $\psi(\beta)$, etc. when it is convenient to work in B rather than A. Further notational shortcuts such as $g_\theta \equiv g_{\alpha_\theta}$, $\Sigma_0 \equiv \Sigma_{\beta_0}$, $\psi \equiv \psi(\alpha) \equiv \psi(\beta)$, will be used in unambiguous contexts.

Let ∇_α be the gradient operator $(\partial/\partial\alpha(1), \partial/\partial\alpha(2))'$, so that the differentiation results following (2.3) can be written as

$$\nabla_\alpha \psi = \beta \qquad \text{and} \qquad \nabla_\alpha \beta' = \nabla_\alpha \psi \nabla_\alpha{}' = \Sigma_\alpha. \tag{2.5}$$

(The last of these gives (2.4).) Similarly if $\nabla_\beta \equiv (\partial/\partial\beta(1), \partial/\partial\beta(2))'$ then $\nabla_\beta \alpha' = \Sigma_\alpha{}^{-1}$, and for any function h, $\nabla_\beta h = \Sigma_\beta{}^{-1}\nabla_\alpha h$, h being thought of as defined on both A and B. In particular

$$\nabla_\beta \psi = \Sigma_\alpha{}^{-1}\beta. \tag{2.6}$$

The Kullback–Leibler "distance" between two members of $\mathscr{G}$ is defined to be

$$I(\alpha_1, \alpha_2) \equiv E_{\alpha_1} \log \frac{g_{\alpha_1}(x)}{g_{\alpha_2}(x)} = (\alpha_1 - \alpha_2)'\beta_1 - [\psi(\alpha_1) - \psi(\alpha_2)] , \tag{2.7}$$

which we will also denote $I(\beta_1, \beta_2)$ as convenient. From (2.5)—(2.7) it is easy to derive

$$\nabla_\beta I(\beta_0, \beta) = \dot{\Sigma}_\beta^{-1}(\beta - \beta_0) . \tag{2.8}$$

The role of the Kullback–Leibler distance is clearer in the framework of convex duality, as introduced into exponential family theory by Barndorff-Nielsen [1], see Section 7.

The simplest example of a two-dimensional exponential family is the normal translation family $x \sim \mathscr{N}_2(\beta, \Sigma)$, Σ fixed and known. The usual density function can be written as

$$e^{\beta'\Sigma^{-1}x - \frac{1}{2}\beta'\Sigma^{-1}\beta}\left[\frac{e^{-\frac{1}{2}x'\Sigma^{-1}x}}{2\pi|\Sigma|^{\frac{1}{2}}}\right], \tag{2.9}$$

showing that we can take $\alpha = \Sigma^{-1}\beta$, $\psi = \frac{1}{2}\beta'\Sigma^{-1}\beta$, $\nu \sim \mathscr{N}_2(0, \Sigma)$. In this case, and only in this case, the local linear mapping (2.4) is globally valid, $B = \Sigma A$, the crucial fact being that $\dot{\Sigma}_\alpha$ does not depend on α. By transforming to $\tilde{x} \equiv \Sigma^{-\frac{1}{2}}x$ we make $\Sigma = I$, in which case A and B exactly coincide.

Example of B not convex. Define $\mathscr{L}_1 \equiv \{(x_1, 0) : -\infty < x_1 < \infty\}$, $\mathscr{L}_2 \equiv \{(0, x_2) : 0 \leqq x_2 < \infty\}$, and let $\nu(x)$ be the probability measure putting probability one on $\mathscr{L}_1 \cup \mathscr{L}_2$ as follows: $\nu(\mathscr{L}_1) = \nu(\mathscr{L}_2) = \frac{1}{2}$, the conditional density of x_2 given x on $\mathscr{L}_2$ is e^{-x_2}, while the conditional density of x_1 given x on $\mathscr{L}_1$ is

$$h(x_1) \equiv c_0 \frac{e^{-|x_1|}}{1 + x_1^4} , \qquad -\infty < x_1 < \infty , \tag{2.10}$$

c_0 a normalizing constant. The exponential family (2.1) with respect to this carrier ν has $A = \{(\alpha_1, \alpha_2) : -1 \leqq \alpha_1 \leqq 1, \alpha_2 \leqq 1\}$. The expectation space B is not convex. Its closure includes $\mathscr{L}_2$ and a finite subinterval of $\mathscr{L}_1$, say $(-c_1, c_1)$ where $c_1 \equiv \int_{-\infty}^{\infty} x_1 e^{x_1} h(x_1)\, dx_1$. It is easy to verify that for a fixed value of $\beta_2 > 0$ the set of points (β_1, β_2) in B is the interval

$$|\beta_1| \leqq c_1\left[1 - \frac{(\beta_2^2 + 4\beta_2 k_1)^{\frac{1}{2}} - \beta_2}{2k_1}\right] \tag{2.11}$$

where $k_1 \equiv \int_{-\infty}^{\infty} e^{x_1} h(x_1)\, dx_1$. B looks like an infinitely tall concave Christmas tree.

3. Two pictures of the MLE. Consider a one-parameter subfamily $\mathscr{F}$ of the two-parameter exponential family $\mathscr{G}$, represented, say, by the densities

$$f_\theta(x) = e^{\alpha_\theta' x - \psi_\theta} , \quad \theta \in \Theta , \tag{3.1}$$

with respect to the carrier ν for $\mathscr{G}$. Here Θ is an interval of the real line, α_θ a continuously twice differentiable function from Θ into A, and $\psi_\theta \equiv \psi(\alpha_\theta)$. We are interested in the maximum likelihood estimation of θ from the data x. Two

complementary pictures of the estimation process will be presented, one due to Fisher [7], and the other to Hoeffding [10].

$\mathscr{F}$ is a *curved exponential family* in the terminology of Efron [5]. It is represented by the curve $\mathscr{F}_A \equiv \{\alpha_\theta : \theta \in \Theta\}$ in A, and equally well by $\mathscr{F}_B \equiv \{\beta_\theta = \beta(\alpha_\theta) : \theta \in \Theta\}$ in B. If $\mathscr{F}_A$ is a straight line segment then $\mathscr{F}$ is contained in a one-parameter exponential family. (This does not imply that $\mathscr{F}_B$ is straight.) In the more challenging cases $\mathscr{F}_A$ is curved. An example is repeated sampling from a normal distribution with fixed coefficient of variation, i.e., $y_1, y_2, \cdots, y_n$ are independent and $\mathscr{N}(\theta, c\theta^2)$, $c > 0$ known. In this case x is the two dimensional sufficient statistic $(\sum_1^n y_i/n, \sum_1^n y_i^2/n)'$.

In a certain sense curved exponential families are a paradigm for all smoothly defined statistical problems in which the minimal sufficient statistic has higher dimension than the parameter space. Fisher particularly liked to use curved multinomial families to represent general statistical problems. See Efron [5] for more background and properties.

Let

$$l_\theta(x) \equiv \log f_\theta(x) = \alpha_\theta' x - \psi_\theta \tag{3.2}$$

so that

$$\dot{l}_\theta(x) = \dot{\alpha}_\theta'(x - \beta_\theta)\,, \tag{3.3}$$

the dot indicating differentiation with respect to θ. Here we have used $\dot{\psi}_\theta = \dot{\alpha}_\theta' \beta_\theta$, derived from the first line of (2.5). The maximum likelihood estimate $\hat{\theta}$ satisfies

$$0 = \dot{l}_{\hat{\theta}}(x) = \dot{\alpha}_{\hat{\theta}}'(x - \beta_{\hat{\theta}})\,, \tag{3.4}$$

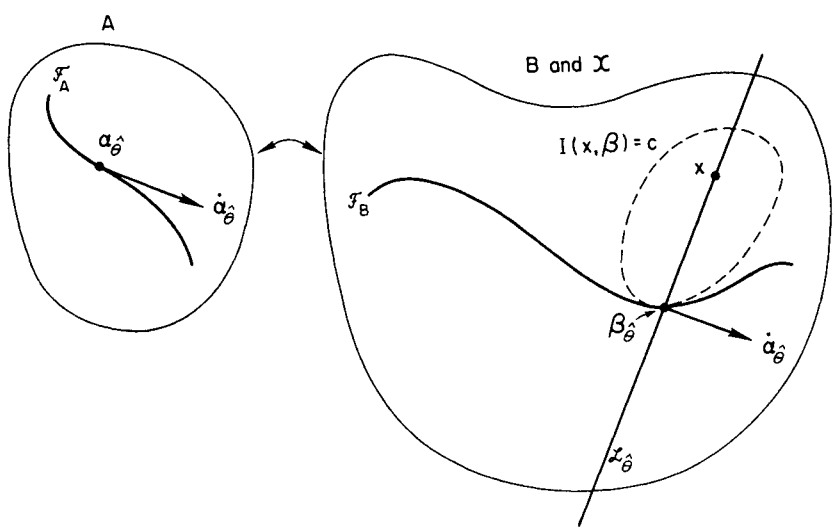

FIG. 1. The curved exponential family $\mathscr{F}$ is represented by the curve $\mathscr{F}_A \equiv \{\alpha_\theta : \theta \in \Theta\}$ is A and also by $\mathscr{F}_B \equiv \{\beta_\theta : \theta \in \Theta\}$ in B. The MLE $\beta_{\hat{\theta}}$ is obtained by "projecting" the data point x onto $\mathscr{F}_B$ orthogonally to $\dot{\alpha}_{\hat{\theta}}$. It is also obtained by finding the nearest point to x on $\mathscr{F}_B$ in terms of the Kullback-Leibler distance $I(x, \beta)$.

assuming, as we shall, that the maximum is not achieved at an endpoint of $\mathscr{F}_B$. In Figure 1 the sample space $\mathscr{X}$ is superimposed on B. (See Section 7 for a discussion of B's relation to $\mathscr{X}$.) For a given value of $\hat{\theta}$, the set of x points having $\hat{\theta}$ as a solution to the MLE equation (3.4), say

$$\mathscr{L}_{\hat{\theta}} \equiv \{x : l_{\hat{\theta}}(x) = 0\} \tag{3.5}$$

is a straight line orthogonal to $\dot{\alpha}_{\hat{\theta}}$ and intersecting $\mathscr{F}_B$ at $\beta_{\hat{\theta}}$. This is Fisher's [7] picture, as illustrated in Figure 2. If it should happen that x falls on $\mathscr{F}_B$, say $x = \beta_\theta$, then the MLE equals θ. In other words, the MLE is "Fisher consistent."

Hoeffding's [10] picture relies on the relationship

$$g_\beta(x) = g_x(x)e^{-I(x,\beta)}, \tag{3.6}$$

g_x indicating the density function of x when $\beta = x$. From (2.1) we get

$$\log \frac{g_\beta(x)}{g_x(x)} = (\alpha - \alpha_x)'x - [\psi(\alpha) - \psi(\alpha_x)], \tag{3.7}$$

where α and α_x are the points in A corresponding to β and x respectively. Comparison of (3.7) with (2.7) gives (3.6).

The "circle" of β values $\{\beta : I(x, \beta) = c\}$ grows larger as the constant c is increased. The smallest value of c for which such a circle touches $\mathscr{F}_B$ gives the maximum likelihood value $\beta_{\hat{\theta}}$ at the point of contact, by (3.6). Figure 1 illustrates Hoeffding's picture, showing $\beta_{\hat{\theta}}$ as the nearest point to x on $\mathscr{F}_B$ in terms of the Kullback–Leibler distance function $I(x, \beta)$. Hoeffding's approach, unlike Fisher's, seems to require $x \in B$, but that this is not actually the case is shown in Section 7.

Of course (3.4) is just a differential condition for $\hat{\theta}$ to locally minimize $I(x, \beta_\theta)$, so that the two pictures are simple variations on the same theme. Both are interesting nevertheless, and in different situations can be individually useful for visualizing the maximum likelihood process. See for example Dempster [4], and Hoeffding [10].

The local curvature of $\mathscr{F}_A$ at $\alpha_{\hat{\theta}}$ plays an important role in the subsequent theory, both geometrically and statistically. Letting $\mathfrak{V}_{\hat{\theta}} \equiv \mathfrak{V}_{\alpha_{\hat{\theta}}}$ and $\nu_{20} \equiv \dot{\alpha}_\theta' \mathfrak{V}_{\hat{\theta}} \dot{\alpha}_{\hat{\theta}}$, $\nu_{11} \equiv \dot{\alpha}_{\hat{\theta}}' \mathfrak{V}_{\hat{\theta}} \ddot{\alpha}_{\hat{\theta}}$, $\nu_{02} \equiv \ddot{\alpha}_{\hat{\theta}}' \mathfrak{V}_{\hat{\theta}} \ddot{\alpha}_{\hat{\theta}}$, this curvature is defined to be

$$\gamma_{\hat{\theta}} \equiv \left[\frac{\nu_{02}}{\nu_{20}^2} - \frac{\nu_{11}^2}{\nu_{20}^3}\right]^{\frac{1}{2}}. \tag{3.8}$$

The quantity $\gamma_{\hat{\theta}}$, called the *statistical curvature of* $\mathscr{F}$ *at* $\hat{\theta}$ in Efron [5], is invariant under any smooth reparametrization of $\mathscr{F}$. The *radius of curvature* is the inverse quantity

$$\rho_{\hat{\theta}} \equiv 1/\gamma_{\hat{\theta}}. \tag{3.9}$$

If $\mathscr{F}$ is a one-parameter exponential family, uncurved, then $\gamma_\theta = 0$, $\rho_\theta = \infty$ for all θ.

4. Constant information "circles." The equivalence of Fisher's and Hoeffding's characterizations of the MLE results in curves of constant Kullback–Leibler information distance having some circle-like properties. For a fixed β_0 in B and $c \geqq 0$ define

$$\mathscr{C}_B \equiv \{\beta : I(\beta_0, \beta) = c\}\,, \tag{4.1}$$

it being assumed that $\mathscr{C}_B$ is contained in the interior of B. The corresponding curve in A is

$$\mathscr{C}_A \equiv \{\alpha : I(\alpha_0, \alpha) = c\}\,, \tag{4.2}$$

obtained by mapping β_0 to α_0 and every point β on $\mathscr{C}_B$ to a point α on $\mathscr{C}_A$. Figure 2 illustrates the situation.

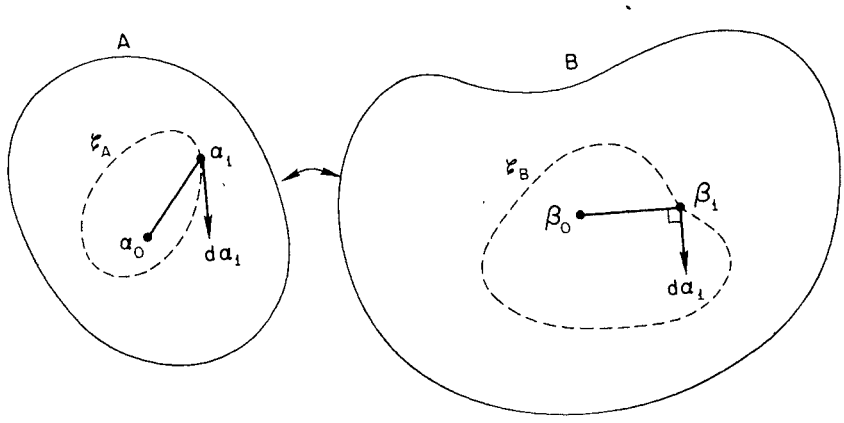

FIG. 2. $\mathscr{C}_B$ is the set of points β in B satisfying $I(\beta_0, \beta) = c$. The corresponding set in A is $\mathscr{C}_A$. The radius of $\mathscr{C}_B$ from β_0 to β_1 is perpendicular to the tangent of $\mathscr{C}_A$ at α_1.

Let $d\alpha_1$ represent a tangent vector to $\mathscr{C}_A$ at point α_1 corresponding to point β_1 on $\mathscr{C}_B$. Let ρ_1 be the radius of curvature of $\mathscr{C}_A$ at α_1 as defined at (3.8) (using any parametrization of $\mathscr{C}_A$).

THEOREM 1.

$$\text{(i)}\quad (d\alpha_1)'(\beta_1 - \beta_0) = 0\,. \tag{4.3}$$

$$\text{(ii)}\quad [(\beta_1 - \beta_0)'\dot{\Sigma}_{\beta_1}^{-1}(\beta_1 - \beta_0)]^{\frac{1}{2}} = \rho_1\,. \tag{4.4}$$

For the normal translation family $x \sim \mathscr{N}_2(\beta, I)$, A and B are identical, and $\mathscr{C}_A$ and $\mathscr{C}_B$ are circles of radius $(2c)^{\frac{1}{2}}$. In this case (4.3) says that a circle's radius is orthogonal to the tangent where it touches, while (4.4) expresses the fact, almost a definition, that the radius of curvature of a circle equals its radius. Theorem 1 generalizes these statements, it once again being necessary to consider spaces A and B simultaneously in general exponential families.

PROOF OF (4.3). From (2.4) we see that $d\beta_1 \equiv \dot{\Sigma}_{\alpha_1} d\alpha_1$ is tangent to $\mathscr{C}_B$ at β_1. Since $\mathscr{C}_B$ is a level curve of $I(\beta_0, \beta)$, (2.8) gives

$$0 = (d\beta_1)'\dot{\Sigma}_{\beta_1}^{-1}(\beta_1 - \beta_0) = (d\alpha_1)'(\beta_1 - \beta_0)\,. \tag{4.5}$$

The proof of (4.4) is deferred to Section 5. The generalization of (4.4) to dimensions greater than 2 is not straightforward, unlike all of our other results.

Since $I(\alpha_0, \alpha)$ is convex in α, $(\nabla_\alpha I(\alpha_0, \alpha)\nabla_\alpha' = \Sigma_\alpha$ by (2.5)—(2.8)), $\mathscr{C}_A$ is the boundary of a convex set. Counterexamples can be constructed to show that the region bounded by $\mathscr{C}_B$ is not necessarily convex.

The gradient relation (2.8) which leads to (4.3) has an important statistical implication. Suppose $\mathscr{F}$ is a genuine one-parameter exponential family (uncurved) contained in $\mathscr{G}$, corresponding say to the straight line segment of α vectors $\alpha_0 + \theta\alpha_1$ for some interval $\Theta \subset R^1$. Let x be the data point, $\hat{\theta}$ the MLE, which is assumed to be in the interior of Θ, and θ any point in Θ. Then we have the Pythagorean relationship

$$I(x, \beta_\theta) = I(x, \beta_{\hat{\theta}}) + I(\beta_{\hat{\theta}}, \beta_\theta)\,, \tag{4.6}$$

a result first given in full generality by Simon [15]. See also Efron [5]. In normal families (4.6) is a typical additivity result for sums of squares in linear models.

PROOF OF (4.6). $\dot{\alpha}_\theta = \alpha_1$ so $\dot{\beta}_\theta = \Sigma_{\beta\theta}\alpha_1$ by (2.4). Combining this with (2.8) gives

$$\frac{\partial[I(x, \beta_\theta) - I(\beta_{\hat{\theta}}, \beta_\theta)]}{\partial\theta} = \alpha_1'(\beta_{\hat{\theta}} - x) = 0\,, \tag{4.7}$$

the last equality following from (3.4). This says that $I(x, \beta_\theta) - I(\beta_{\hat{\theta}}, \beta_\theta)$ is a constant, which is seen to equal $I(x, \beta_{\hat{\theta}})$ upon substituting $\theta = \hat{\theta}$.

5. Second derivative of the log likelihood. In this section we consider the variation of $\ddot{l}_{\hat{\theta}}(x) \equiv \partial^2/\partial\theta^2[\log f_\theta(x)]|_{\theta=\hat{\theta}}$ along lines $\mathscr{L}_{\hat{\theta}}$ of constant maximum likelihood estimate, for curved exponential families. See Figure 2 and the definitions at the beginning of Section 3. From (3.3) we see that the Fisher information $i_\theta \equiv E[\dot{l}_\theta(x)]^2$ is given by

$$i_\theta = \dot{\alpha}_\theta'\Sigma_\theta\dot{\alpha}_\theta\,, \tag{5.1}$$

$\Sigma_\theta \equiv \Sigma_{\alpha_\theta} \equiv \Sigma_{\beta_\theta}$. Differentiating (3.3) gives

$$\ddot{l}_\theta(x) = \ddot{\alpha}_\theta'(x - \beta_\theta) - i_\theta\,, \tag{5.2}$$

$\ddot{\alpha}_\theta \equiv \partial^2/\partial\theta^2[\alpha_\theta]$, where we have used

$$\dot{\beta}_\theta = \Sigma_\theta\dot{\alpha}_\theta \tag{5.3}$$

derived from (2.4).

Equation (5.2) shows that $\ddot{l}_{\hat{\theta}}(x)$ varies linearly in x along the line $\mathscr{L}_{\hat{\theta}}$. This means that there is a critical point $c_{\hat{\theta}}$ on $\mathscr{L}_{\hat{\theta}}$ such that $\ddot{l}_{\hat{\theta}}(c_{\hat{\theta}}) = 0$. The location of $c_{\hat{\theta}}$ is "above" $\mathscr{L}_B$ (i.e., in the direction along $\mathscr{L}_{\hat{\theta}}$ most closely aligned with $\ddot{\alpha}_{\hat{\theta}}$, as in Figure 3) at a position determined by $\rho_{\hat{\theta}}$, the radius of curvature (3.9).

THEOREM 2. *The critical point $c_{\hat{\theta}}$ is Mahalanobis distance $\rho_{\hat{\theta}}$ above $\mathscr{F}_B$,*

$$\rho_{\hat{\theta}} = ((c_{\hat{\theta}} - \beta_{\hat{\theta}})'\Sigma_{\hat{\theta}}^{-1}(c_{\hat{\theta}} - \beta_{\hat{\theta}}))^{\frac{1}{2}}\,. \tag{5.4}$$

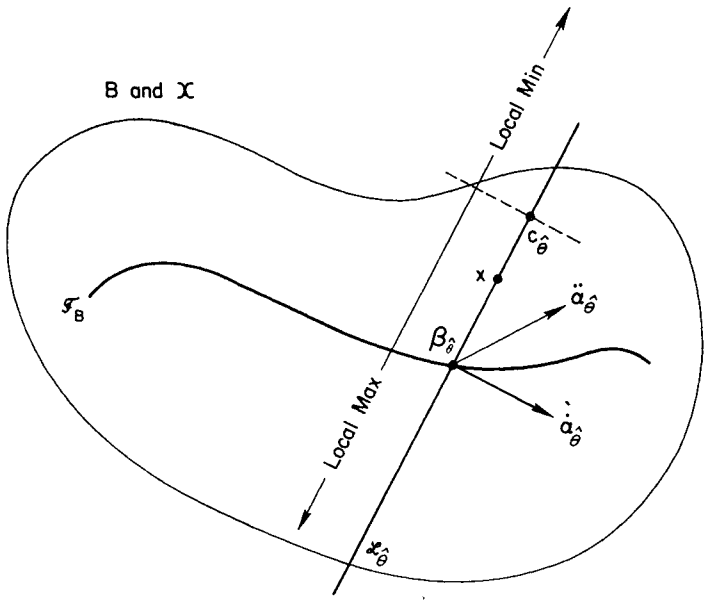

FIG. 3. If x is a proportion R of the way from the critical point $c_{\hat\theta}$ to $\beta_{\hat\theta}$ then $-\ddot{l}_{\hat\theta}(x) = Ri_{\hat\theta}$. $\hat\theta$ is a local maximum of the likelihood function for $R > 0$ and a local minimum for $R < 0$. The critical point is on the side of $\mathscr{F}_B$ determined by the direction of $\ddot{\alpha}_{\hat\theta}$, at Mahalanobis distance $\rho_{\hat\theta}$ from $\beta_{\hat\theta}$.

If x on $\mathscr{L}_{\hat\theta}$ is a proportion R of the distance from $c_{\hat\theta}$ to $\beta_{\hat\theta}$ then

$$-\ddot{l}_{\hat\theta}(x) = Ri_{\hat\theta}\,. \tag{5.5}$$

Several comments precede the proof of Theorem 2.

(1) Why should we be interested in the variation of $\ddot{l}_{\hat\theta}(x)$ along $\mathscr{L}_{\hat\theta}$? Fisher [8] thought of this as a (primary) explanation for the insufficiency of the MLE when $\mathscr{F}$ is not an exponential family. The importance of Fisher's considerations is illustrated by the two examples in Section 6.

(2) If $\mathscr{F}$ is an exponential family then $\rho_\theta = \infty$ for all θ, since $\mathscr{F}_A$ is straight, and $-\ddot{l}_{\hat\theta}(x) = i_{\hat\theta}$ for all x on $\mathscr{L}_{\hat\theta}$ in accordance with (5.5). In this case $\hat\theta$ is sufficient, of course, so the entire likelihood function up to constant multiples, and not just $\ddot{l}_{\hat\theta}(x)$, is a function of x only through $\hat\theta$.

(3) If $\mathscr{F}$ is a curved exponential family and $x_1, x_2, \cdots, x_n$ a random sample from $f_\theta(x)$ in $\mathscr{F}$, it is possible to show that the coefficient of variation of $R \equiv -\ddot{l}_{\hat\theta}(x_1, \cdots, x_n)/ni_{\hat\theta}$ goes to zero as $(n^{\frac{1}{2}}\rho_\theta)^{-1}$ as n goes to infinity. The proof, which depends on Theorem 2 and Efron [5] will not be given here. The quantity $(n^{\frac{1}{2}}\rho_\theta)^{-1}$ is the statistical curvature based on all the available data, which suggests that this last quantity be used to approximate the coefficient of variation of R, as in the examples of Section 6.

(4) The proportion R referred to in Theorem 2 does not depend on which metric (Mahalanobis or ordinary) is used to measure distance, since $\mathscr{L}_{\hat\theta}$ is one dimensional.

(5) If $x = \beta_{\hat\theta}$ then $R = 1$ implying $-\ddot{l}_{\hat\theta}(\beta_{\hat\theta}) = i_{\hat\theta}$. We see that $-\ddot{l}_{\hat\theta}(x)$ is Fisher consistent for $i_\theta = E_\theta\{-\ddot{l}_{\hat\theta}(x)\}$, and hence consistent in the usual sense, though it is not in general unbiased. (Of course, $-\ddot{l}_{\hat\theta}$ is consistent for i_θ in more general contexts than curved exponential families.)

(6) As x moves along $\mathscr{L}_{\hat\theta}$ away from $\mathscr{F}_B$, $\hat\theta$ changes from a local *maximum* of the likelihood function to a local *minimum* at $x = c_{\hat\theta}$.

(7) For i.i.d. sampling as in Remark 3, Theorem 2 becomes $-\ddot{l}_{\hat\theta}(x_1, \cdots, x_n) = Rni_{\hat\theta}$, where R is as pictured in Figure 3 with "x" replaced by $\bar{x} \equiv \sum_1^n x_n/n$. All of the other quantities in Figure 3 remain exactly as defined before, see Efron [5], Section 6.

(8) If $\mathscr{G}$ is a k dimensional exponential family and $\mathscr{F}$ a one-parameter curved subfamily then $\mathscr{L}_{\hat\theta}$ is a $k - 1$ dimensional hyperplane orthogonal to $\dot\alpha_{\hat\theta}$, passing through $\beta_{\hat\theta}$. The critical "point" $c_{\hat\theta}$ is now a $k - 2$ dimensional hyperplane contained in $\mathscr{L}_{\hat\theta}$ and perpendicular to the projection of $\ddot\alpha_{\hat\theta}$ into $\mathscr{L}_{\hat\theta}$; (5.4) takes the form

$$\min_{y \in c_{\hat\theta}} [(y - \beta_{\hat\theta})'\Sigma_{\hat\theta}^{-1}(y - \beta_{\hat\theta})]^{\frac{1}{2}} = \rho_{\hat\theta}\,. \tag{5.6}$$

If x in $\mathscr{L}_{\hat\theta}$ lies on the parallel to $c_{\hat\theta}$ proportion R of the way from $c_{\hat\theta}$ to $\beta_{\hat\theta}$ then (5.5) holds.

PROOF OF THEOREM 2. There is no loss of generality in assuming that $\mathscr{F}$ is in "standard form at $\theta = \hat\theta$," Section 4, Efron [5],

$$\alpha_{\hat\theta} = \beta_{\hat\theta} = 0\,, \qquad \Sigma_{\hat\theta} = I \tag{5.7}$$

and

$$\dot\alpha_{\hat\theta} = \dot\beta_{\hat\theta} = i_{\hat\theta}^{\frac{1}{2}} e_1\,, \qquad \ddot\alpha_{\hat\theta} = a(\hat\theta)e_1 + (i_{\hat\theta}/\rho_{\hat\theta})e_2\,. \tag{5.8}$$

Here e_1 and e_2 are the coordinate axes in R^2, and $a(\theta) \equiv \dot\alpha_\theta'\Sigma_\theta\ddot\alpha_\theta/i_\theta^{\frac{1}{2}}$.

Comparing (3.4) with (5.7)—(5.8) shows that in the standard form coordinates, x must lie along e_2, $x = (0, x_2)'$. Then (5.2) yields

$$\ddot{l}_{\hat\theta}(x) = (i_{\hat\theta}/\rho_{\hat\theta})x_2 - i_{\hat\theta} = i_{\hat\theta}(x_2/\rho_{\hat\theta} - 1)\,. \tag{5.9}$$

The condition $\ddot{l}_{\hat\theta}(x) = 0$ is equivalent to $x_2 = \rho_{\hat\theta}$, $x = (0, \rho_{\hat\theta})'$, which implies (5.4). (The transformation to standard form preserves Mahalanobis distance.) The special case where x lies exactly on $\mathscr{F}_B$ has $x_2 = 0$, so by (5.9) $-\ddot{l}_{\hat\theta}(x) = i_{\hat\theta}$. Finally (5.5) follows by the linearity of $\ddot{l}_{\hat\theta}(x)$ along $\mathscr{L}_{\hat\theta}$.

PROOF OF (4.4). If we parametrize $\mathscr{C}_B$ by some parameter θ, and take x equal to β_0 in (3.6), then it is obvious from definition (4.1) that $-\ddot{l}_\theta(\beta_0) = 0$ for every θ. In other words, $c_{\hat\theta} = \beta_0$ for every θ. Then (5.3) implies (4.4).

6. A simple example. A particularly simple curved exponential family due to Fisher [9] is discussed next, for which the relationships of Section 5 can be explicitly displayed. Some of Fisher's ideas on ancillarity and conditional inference are illustrated using this example, and a connection made to the Cauchy translation family.

The model for $\mathscr{F}$ is bivariate normal with identity covariance matrix and mean vector on the circle of radius ρ centered at $(0, \rho)'$,

$$x \sim \mathscr{N}_2(\beta_\theta, I), \qquad \beta_\theta = \rho \begin{pmatrix} \sin\theta \\ 1 - \cos\theta \end{pmatrix} \quad \text{for} \quad -\pi < \theta \leqq \pi, \tag{6.1}$$

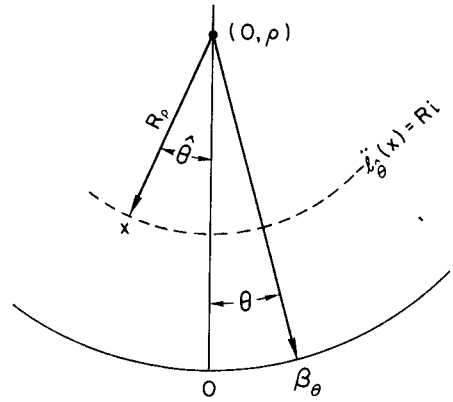

FIG. 4. Fisher's circle model. The random variable x is bivariate normal $\mathscr{N}_2(\beta_\theta, I)$, where β_θ lies on a circle of radius ρ. In the model the curvature ρ and information i do not depend on θ. The statistic R is ancillary, as is $-\ddot{l}_{\hat{\theta}}(x) = Ri$.

as shown in Figure 4. This is a curved exponential family in which $\alpha \equiv \beta$, $A \equiv B$. The following facts are easily verified:

(i) $\rho_\theta = \rho$ for all θ, $i_\theta \equiv i = \rho^2$ for all θ.

(ii) $\mathscr{L}_{\hat{\theta}}$ is the line through $\beta_{\hat{\theta}}$ and $(0, \rho)'$.

(iii) The critical point $c_{\hat{\theta}}$ is $(0, \rho)'$ for all $\hat{\theta}$.

(vi) If x lies on the circle of radius $R\rho$ centered at $(0, \rho)'$, then $-\ddot{l}_{\hat{\theta}}(x) = Ri$.

(v) The conditional density of $\hat{\theta}$ given x on the circle of radius R described in (iv) is proportional to $\exp[R\rho^2 \cos(\hat{\theta} - \theta)]$. This is the von Mises distribution [9].

(vi) The statistic R is ancillary (distribution not depending on θ) as is $-\ddot{l}_{\hat{\theta}}(x) = Ri$.

Fisher argued forcefully that all statistical inferences should be made conditional on ancillary statistics, his 1934 paper [8] being particularly persuasive. In the present context one might try to follow Fisher's dictum by making the conditional variance approximation

$$\mathrm{Var}_\theta\{\hat{\theta} \mid -\ddot{l}_{\hat{\theta}}(x) = Ri\} \approx \frac{1}{Ri} \tag{6.2}$$

rather than the familiar unconditional approximation

$$\mathrm{Var}\{\hat{\theta}\} \approx \frac{1}{i}. \tag{6.3}$$

In more complicated contexts, for example Cox [3], approximation (6.2) can be a good deal easier to compute than (6.3). It also gives a cozy feeling of being "closer to the data." But how accurate is it?

Figure 5 displays the answer for the circle model with $\rho = 8^{\frac{1}{2}}$. For comparison with the Cauchy translation problem considered next, the parameter of interest is taken to be

$$\phi \equiv \theta/(1.25)^{\frac{1}{2}} \tag{6.4}$$

rather than θ itself. When parameterized by ϕ, $\mathscr{T}$ has constant curvature ρ and constant information $i_\phi = 1.25\,\rho^2 = 10$. The approximation $\mathrm{Var}_\phi\{\hat{\phi} \mid -\ddot{l}_{\hat{\phi}}(x) = Ri_\phi\} \approx 1/(Ri_\phi)$ is seen to be quite accurate when compared to the exact conditional variance computed for the von Mises distribution.

As a point of comparison with a statistically more interesting model, consider the maximum likelihood estimation of the unknown center ϕ of a standard Cauchy translation family, based on a random sample of size 20. Although this is not a curved exponential family it can be approximated by such a family having constant radius of curvature $\rho = 8^{\frac{1}{2}}$ and information $i_\phi = 10$. See Efron [5], Section 10. The results of a Monte Carlo trial of 14048 such random samples of size 20, $\phi = 0$, are recorded in Figure 5. The ancillary statistic $1/(-\ddot{l}_{\hat{\phi}})$, grouped over small intervals, is plotted versus the observed second

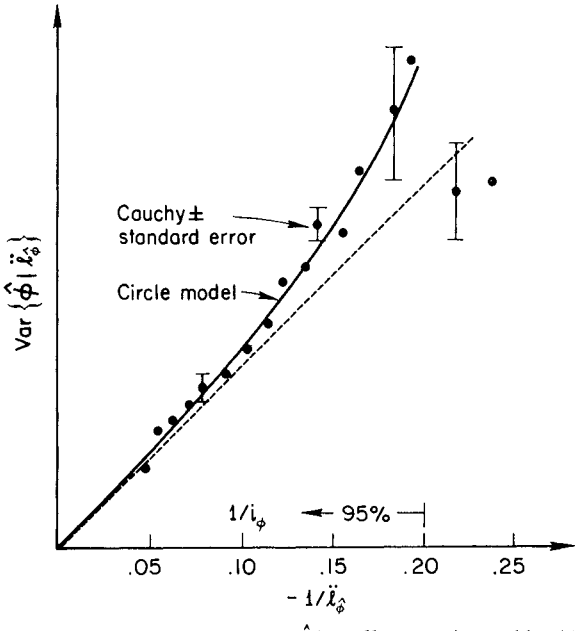

FIG. 5. The conditional variance of $\hat{\phi}$ is well approximated by $1/(-\ddot{l}_{\hat{\phi}}(x))$ in the circle model. The case illustrated has $\rho = 8^{\frac{1}{2}}$, $i_\phi = 10$, which is the same as for a Cauchy translation problem with $n = 20$. 14048 Monte Carlo replications for the Cauchy problem showed a similar pattern. (Based on work with D. Hinkley [6].)

moment of the corresponding $\hat{\phi}$ values. For example, 224 of the 14048 trials had $1/(-\ddot{l}_{\hat{\phi}}(x))$ in the range .170—.180, averaging .175, and the 224 values of $\hat{\phi}^2$ had mean .201, standard error .023.

The implications of Figure 5 are considerable. A Cauchy or circle model sample with $-\ddot{l}_{\hat{\phi}} = 15$ gives conditional 95% confidence interval approximately $\hat{\phi} \pm 1.95/15^{\frac{1}{2}}$, compared to the unconditional interval $\hat{\phi} \pm 1.95/10^{\frac{1}{2}}$ based on (6.3). David Hinkley and the author [6] have investigated (6.2) in contexts more general than translation families, which Fisher [8] showed to have particularly simple ancillary structure. (See the discussion following Efron [5], particularly the remarks of Cox and Pierce. Cobb [2] gives an especially interesting discussion of ancillarity in constant curvature families.)

Remark 3 following Theorem 2 suggests that $-\ddot{l}_{\hat{\phi}}$ should have coefficient of variation approximately $1/\rho = .35$ in the circle and Cauchy cases (since $i_{\hat{\phi}}$ is a constant). The actual value is .32 for both cases. Table 1 shows that $-\ddot{l}_{\hat{\phi}}$ has almost exactly the same distribution for both problems. The similarity is explained to a large extent, but not completely, by the identical values for ρ and i_{ϕ}.

TABLE 1

The percentiles of $-\ddot{l}_{\hat{\phi}}$ for the circle model (6.1), (6.4), $\rho = 8^{\frac{1}{2}}$, *and for the Cauchy translation problem,* $n = 20$. *The coefficient of variation of $-\ddot{l}_{\hat{\phi}}$ is* .32 *in both cases, compared to the approximation* $1/\rho = .35$ *suggested in Section* 5. (*The Cauchy results are those from the Monte Carlo study with* 14048 *replications.*)

Prob $\{-\ddot{l}_{\hat{\phi}} < c\}$		5%	10%	30%	50%	70%	90%	95%
c for	circle model	5.1	6.3	8.8	10.6	12.4	15.0	16.3
	Cauchy	5.1	6.3	8.8	10.7	12.6	15.2	16.6

7. Duality. The spaces A and B play dissimilar roles in Figure 1, and also in Figure 2. There is a dual theory in which these roles are reversed. "Dual" is the appropriate word here since the reversed results follow most easily from our previous ones by means of convex duality. Barndorff-Nielson [1] has pioneered the application of this elegant theory to exponential families. Only a brief glimpse of the theory will be given here.

Figure 6 shows the point α_x in A corresponding to the data point x in B. An estimator $\tilde{\theta}(x)$ analogous to the MLE can be defined as that value in Θ minimizing $I(\alpha_\theta, \alpha_x)$, as shown in Figure 6. The "circle" $\tilde{\mathscr{C}}_A \equiv \{\alpha : I(\alpha, \alpha_x) = c\}$ is tangent to $\mathscr{T}_A$ at $\alpha_{\tilde{\theta}}$, where c is the minimum value for which contact is made. It then turns out that the line $\tilde{\mathscr{L}}_{\tilde{\theta}}$ passing through $\alpha_{\tilde{\theta}}$ orthogonal to $\dot{\beta}_{\tilde{\theta}}$ goes through α_x.

The same relationship exists between $\tilde{\mathscr{C}}_A$ and its image $\tilde{\mathscr{C}}_B$ in B as did between $\mathscr{C}_B$ and $\mathscr{C}_A$. Theorem 1, as illustrated in Figure 2, remains valid with the places of α and β and of A and B everywhere reversed. $\tilde{\mathscr{C}}_B$ is the boundary of a convex set, but not necessarily $\tilde{\mathscr{C}}_A$.

Kullback [11] and others have made extensive use of the $\tilde{\theta}$ estimation process.

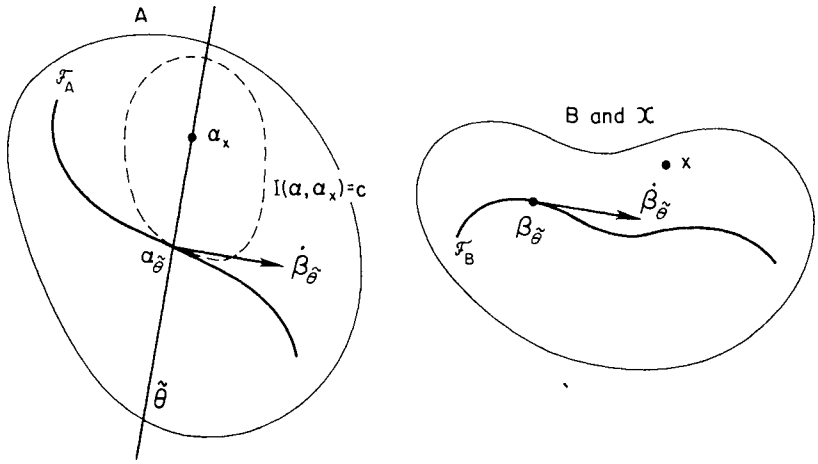

FIG. 6. A theory of estimation dual to maximum likelihood. This picture should be compared with Figure 1. The estimate $\alpha_{\tilde{\theta}}$ is the projection of α_x onto $\mathcal{F}_A$, orthogonally to $\dot{\beta}_{\tilde{\theta}}$. It is also the nearest point α to α_x on $\mathcal{F}_A$, measured in terms of $I(\alpha, \alpha_x)$.

If $\mathcal{F}_B$ is linear, rather than $\mathcal{F}_A$ as for exponential families, then $\tilde{\theta}$ is easier to work with than the MLE. Simon's additivity theorem (4.6) holds in this case, in the interchanged form $I(\alpha_\theta, \alpha_x) = I(\alpha_{\tilde{\theta}}, \alpha_x) + I(\alpha_\theta, \alpha_{\tilde{\theta}})$. On the other hand, this form of estimation may be outperformed by the MLE in small samples, see Rao [13] and Efron [5].

The theory of convex duality begins with a convex function $\psi(\alpha)$ defined on a convex set A. It is convenient to assume that ψ has positive definite second derivative matrix Σ_α. Letting the gradient of ψ be indicated by $\beta \equiv \nabla_\alpha \psi$, the dual function $\phi(\beta) \equiv \alpha'\beta - \psi(\alpha)$, thought of as a function of β, is convex with second derivative matrix Σ_β^{-1}. The dual of the dual ϕ is ψ and $\nabla_\beta \phi = \alpha$, the $\alpha - \beta$ mapping being one to one.

For α_1, α_2 in A, define the *tangent function* to ψ to be

$$I_A(\alpha_1, \alpha_2) \equiv \psi(\alpha_1) - [\psi(\alpha_2) + \beta_2'(\alpha_1 - \alpha_2)] . \tag{7.1}$$

($I_A(\alpha_1, \alpha_2)$ is the difference evaluated at α_1 between ψ and the plane tangent to ψ at α_2.) Then ϕ has tangent function $I_B(\beta_1, \beta_2)$, $\beta_1, \beta_2 \in B$, satisfying

$$I_B(\beta_1, \beta_2) = I_A(\alpha_2, \alpha_1) , \tag{7.2}$$

where $\beta_1 \equiv \nabla_\alpha \psi|_{\alpha_1}$, $\beta_2 \equiv \nabla_\alpha \psi|_{\alpha_2}$. See Barndorff-Nielson [1] or Rockafellar [14] for a thorough treatment of the theory.

In exponential families, ψ is the normalizing function (2.1), α the natural parameter, β the expectation parameter, Σ_α the covariance matrix, and $I_B(\beta_i, \beta_j)$ the Kullback–Leibler distance. The function ϕ is seen to be the *log maximum likelihood* for $x \in B$,

$$\phi(x) = \alpha_x'x - \psi(\alpha_x) = \max_{\alpha \in A} \log g_\alpha(x) \tag{7.3}$$

by (2.1), (3.5).

Differential relations such as (2.4), (2.6), and (2.8), which are the basis of our previous results for exponential families, are easily verified in the more general duality framework. Theorems 1 and 2 are valid in this wider context. Because of the symmetry of the duality theory we also obtain reversed results such as those illustrated in Figure 6.

We now consider the possibility that the sample point x falls outside of B. More generally let $\bar{\mathscr{X}}_{(n)}$ be the sample space of the sufficient statistic $\bar{x} = \sum_1^n x_i/n$, where $x_1, x_2, \cdots, x_n$ is a random sample from some member g_α of $\mathscr{G}$. The family of distributions of $\bar{x}$, say $\mathscr{G}_{(n)}$, is an exponential family with $\alpha_{(n)} = n\alpha_1$, $\beta_{(n)} \equiv \beta$, $\Sigma_{\alpha_{(n)}} \equiv \Sigma_\alpha/n$, and $\psi_{(n)}(\alpha_{(n)}) = n\psi(\alpha)$, as can be seen from the density

$$g_\alpha(x_1, x_2, \cdots, x_n) = e^{n[\alpha'\bar{x} - \psi(\alpha)]}. \tag{7.4}$$

A curved exponential family in $\mathscr{G}$, say $\mathscr{F}$, is also a curved exponential family in $\mathscr{G}_{(n)}$, and is represented by the same curve $\mathscr{F}_B$ through B. Maximum likelihood estimation remains exactly as pictured in Figure 1, except that x is replaced by $\bar{x}$. (The fact that A is magnified by a factor of n does not affect the direction of $\dot{\alpha}_{\hat{\theta}}$, which, as Figure 1 shows, is its only influence on $\hat{\theta}$.)

Letting $n \to \infty$, the sample spaces $\bar{\mathscr{X}}_{(n)}$ can be taken to be subsets of $\bar{\mathscr{X}}$, the smallest convex set containing $\mathscr{X}$ with probability one. The space $\bar{\mathscr{X}}$ contains B, and the important fact is that the maximum likelihood theory illustrated in Figure 1 can be extended to all points $\bar{x}$ in $\bar{\mathscr{X}}$, not just those in B.

To do so, the function ϕ is defined on $\bar{\mathscr{X}}$ by

$$\phi(\bar{x}) \equiv \sup_{\alpha \in A} \{\alpha'\bar{x} - \psi(\alpha)\}. \tag{7.5}$$

It is not difficult to show, as in Barndorff-Nielsen [1] Chapter 6, that ϕ is convex on $\bar{\mathscr{X}}$, and finite in the interior of $\bar{\mathscr{X}}$. The function $I(\bar{x}, \beta)$, $\bar{x} \in \bar{\mathscr{X}}$, $\beta \in B$, is then defined as the tangent function to ϕ. It is easy to verify $\nabla_\beta I(\bar{x}, \beta) = \Sigma_\beta^{-1}(\beta - \bar{x})$, as at (2.8), and that both Fisher's and Hoeffding's pictures remain valid for $\bar{x} \in \bar{\mathscr{X}}$.

REFERENCES

[1] Barndorff-Nielsen, O. (1970). *Exponential Families, Exact Theory*. Aarhus University, Various Publications Series, Vol. 19.

[2] Cobb, G. W. (1974). Generalization of a paradigm of R. A. Fisher. Harvard Univ., Dept. of Statistics, Research Report S-28.

[3] Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

[4] Dempster, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

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

[6] Efron, B. and Hinkley, D. (1977). The conditional variance of the maximum likelihood estimation. Stanford Univ., Dept. of Statistics, Technical Report No. 100.

[7] Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Trans.* **122** 700–725.

[8] FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Statist. Soc. Ser. A* **144** 285–307.
[9] FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.
[10] HOEFFDING. W. (1965). Asymptotically optimum tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–408.
[11] KULLBACK, S. L. (1959). *Information Theory and Statistics.* Wiley, New York.
[12] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.
[13] RAO, C. R. (1963). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. Ser. B* **24** 46–72.
[14] ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton Univ. Press.
[15] SIMON, G. (1973). Additivity of information in exponential family probability laws. *J. Amer. Statist. Assoc.* **68** 478–482.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305