DEFINING THE CURVATURE OF A STATISTICAL PROBLEM
(WITH APPLICATIONS TO SECOND ORDER EFFICIENCY)

BY

B. EFRON

TECHNICAL REPORT NO. 54
MARCH 12, 1974

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-30711X-2

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
DEFINING THE CURVATURE OF A STATISTICAL PROBLEM
(WITH APPLICATIONS TO SECOND ORDER EFFICIENCY)

by

B. EFRON

TECHNICAL REPORT NO. 54
MARCH 12, 1974

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-30711X-2

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
DEFINING THE CURVATURE OF A STATISTICAL PROBLEM
(WITH APPLICATIONS TO SECOND ORDER EFFICIENCY)

Bradley Efron
Stanford University

Summary

Statisticians know that one-parameter exponential families have very nice properties for estimation, testing, and general inference problems. In many ways this is because they can be considered to be "straight lines" through the space of all possible probability distributions on the sample space. We consider arbitrary one-parameter families $\gamma$ and try to quantify how nearly "exponential" they are. A quantity called "the statistical curvature of $\gamma$" is introduced. Statistical curvature is identically zero for exponential families, positive for non-exponential families. Our purpose is to show that families with small curvature enjoy the good properties of exponential families. Large curvature indicates a breakdown of these properties. Statistical curvature turns out to be closely related to Fisher and Rao's theory of second order efficiency.

American Mathematical Society 1970 subject classification: 62.05

Key words: curvature, exponential families, Cramér-Rao lower bound, locally most powerful tests, Fisher information, second order efficiency, deficiency, maximum likelihood estimation.
DEFINING THE CURVATURE OF A STATISTICAL PROBLEM
(WITH APPLICATIONS TO SECOND ORDER EFFICIENCY)

B. Efron

1. Introduction

Suppose we have a statistical problem involving a one-parameter family of probability density functions \( \mathcal{F} = \{f_\theta(x)\} \). Statisticians know that if \( \mathcal{F} \) is an exponential family then standard linear methods will usually solve the problem in neat fashion. For example, the locally most powerful test of \( \theta = \theta_0 \) versus \( \theta > \theta_0 \) is uniformly most powerful in an exponential family. The maximum likelihood estimator for \( \theta \) is a sufficient statistic in an exponential family, and achieves the Cramer-Rao lower bound if we have chosen the right function of \( \theta \) to estimate.

In this paper we consider arbitrary one-parameter families \( \mathcal{F} \) and try to quantify how nearly "exponential" they are. A quantity \( \gamma_\theta \), called "the statistical curvature of \( \mathcal{F} \) at \( \theta \)" is introduced such that \( \gamma_\theta \) is identically zero if \( \mathcal{F} \) is exponential and greater than zero, for at least some \( \theta \) values, otherwise.

Our purpose is to show that families with small curvature enjoy the good statistical properties of exponential families. Large curvature indicates a breakdown of this favorable situation. For example, if \( \gamma_{\theta_0} \) is large, the locally most powerful test of \( \theta = \theta_0 \) versus \( \theta > \theta_0 \) can be expected to have poor operating characteristics. Similarly the variance of the maximum likelihood estimator (MLE) exceeds the Cramer-Rao lower bound in approximate proportion to \( \gamma_\theta^2 \). (See Sections 8 and 10.)
For non-exponential families the MLE is not, in general, a sufficient statistic. How much information does it lose, compared with all the data \( x \)? The answer can be expressed in terms of \( \gamma_\theta^2 \). This theory goes back to Fisher [7] and Rao [10,11,12] who attempted to show that if \( \mathcal{F} \) is a one-parameter subset of the \( k \)-category multinomial distributions, indexed say by the vector of probabilities \( f_\theta(x) = P_\theta(X \in \text{category } x), x = 1,2,\ldots,k \), the following result holds: let \( i_\theta \) be the Fisher information in an independent sample of size \( n \) from \( f_\theta \), \( i_\theta \) the Fisher information in the maximum likelihood estimator \( \hat{\theta}(x_1,x_2,\ldots,x_n) \) based on that sample, and \( i_\theta \) the Fisher information in a sample of size one (so \( i_\theta = ni_\theta \)). Then

\[
\lim_{n \to \infty} \left( i_\theta - \frac{\hat{\theta}}{i_\theta} \right) = 4 \left( \frac{\mu_{02} - 2\mu_{11} + \mu_{40}}{i^2} - 1 \right) \left( \frac{\mu_{11}^2 + \mu_{40}^2 - 2\mu_{11} \mu_{40}}{i_\theta^3} \right) \tag{1.1}
\]

where

\[
\mu_{ij} = E_\theta \left( \frac{f_\theta(x)}{\hat{f}_\theta(x)} \right)^i \left( \frac{f_\theta(x)}{\hat{f}_\theta(x)} \right)^j \tag{1.2}
\]

the dot indicating differentiation with respect to \( \theta \). Moreover, for any other consistent, efficient estimator \( T(x_1,x_2,\ldots,x_n) \) the asymptotic loss of information \( \lim_{n \to \infty} \left( i_\theta - i_\theta \right) \) is equal or greater than the right side of (1.1). Rao has coined the term "second order efficiency" for this property of the MLE which gives it a preferred place in the class of "first order efficient" estimators \( T \), those which satisfy the weaker condition

\[
\lim_{n \to \infty} i_\theta \cdot i_\theta = 1.
\]
It turns out that the unpleasant looking bracketed term in (1.1) equals $\gamma_0$. This leads to a straightforward geometrical "proof" of (1.1). The quotes are necessary here since, as the counter-example of Section 9 shows, the result is actually not true for multinomial families. However, the difficulty arises only because of the discrete nature of the multinomial, and can be overcome by dealing with less lumpy distributions. More importantly, a similar result of Rao's for squared error estimation risk holds even for the multinomial, as discussed in Section 10.

Under our definition an exponential family has zero curvature everywhere so in some sense it is a "straight line through the space of possible probability distributions." (This is intuitively plausible since linear methods, to the log likelihood function, that is methods based on linear approximations/tend to work perfectly in exponential families. The fact that locally most powerful tests are uniformly most powerful is an example of this.) We will make this notion precise by considering families $\mathcal{F}$ which are subsets of multi-parameter exponential families. If the subset is a straight line in the natural parameter space of the bigger family then $\mathcal{F}$ is a one-parameter exponential family. If the subset is a curved line through the natural parameter space then $\mathcal{F}$ is not exponential, and it turns out that the statistical curvature exactly equals the ordinary geometric curvature of the line, the rate of change of direction with respect to arc-length. For the sake of exposition we actually start with this latter definition in Section 3 and show in Section 5 how it leads to a sensible definition of statistical curvature in the general case. First we give a brief review of the notion of the geometrical curvature of a line.
2. Curvature

If \( Y = Y(X) \) defines a curved line \( \mathcal{L} \) in the \((X,Y)\) plane then

\[
\gamma_X = \frac{(Y'')^2}{\sqrt{1 + (Y')^2}} \tag{2.1}
\]

is defined to be the curvature of \( \mathcal{L} \) at \( X \), where \( Y' = \frac{dY}{dX} \), \( Y'' = \frac{d^2Y}{dX^2} \) are assumed to exist continuously in a neighborhood of the value \( X \) where the curvature is being evaluated. In particular if \( Y' = 0 \) then \( \gamma_X = |Y''| \). An exercise in differential calculus shows that \( \gamma_X \) is the rate of change of direction of \( \mathcal{L} \) with respect to arc-length along the curve. \( \rho_X = \frac{1}{\gamma_X} \), the "radius of curvature", is the radius of the circle tangent to \( \mathcal{L} \) at \((X,Y)\) whose Taylor expansion about \((X,Y)\) agrees up to the quadratic term with that of \( \mathcal{L} \). Struik [13] is a good elementary reference for curvature and related concepts.

The concept of curvature extends to curved lines in Euclidean k-space, \( E^k \), say \( \mathcal{L} = \{\eta_{\theta}, \theta \in \Theta\} \), where \( \Theta \) is an interval of the real line. For each \( \theta \), \( \eta_{\theta} \) is a vector in \( E^k \) whose componentwise derivatives with respect to \( \theta \) we denote \( \dot{\eta}_{\theta} = \frac{\partial}{\partial \theta} \eta_{\theta}, \ddot{\eta}_{\theta} = \frac{\partial^2}{\partial \theta^2} \eta_{\theta} \).

Figure 1. The curvature of \( \mathcal{L} \) at \( \theta_0 \) is \( \frac{d\eta_\theta}{d\theta} \bigg|_{\theta=\theta_0} \).
These derivatives are assumed to exist continuously in a neighborhood of a value of \( \theta \) where we wish to define the curvature. Suppose also that a \( k \times k \) non-negative definite matrix \( \mathcal{L}_\theta \) is defined continuously in \( \theta \).

Let \( M_\theta \) be the \( 2 \times 2 \) matrix, with entries denoted \( v_{20}(\theta), v_{11}(\theta), v_{02}(\theta) \) as shown, defined by

\[
M_\theta = \begin{pmatrix} v_{20}(\theta) & v_{11}(\theta) \\ v_{11}(\theta) & v_{02}(\theta) \end{pmatrix} = \begin{pmatrix} \eta_\theta^2 \xi_\theta \hat{n}_\theta & \eta_\theta \xi_\theta \hat{n}_\theta \\ \eta_\theta \xi_\theta \hat{n}_\theta & \eta_\theta^2 \xi_\theta \hat{n}_\theta \end{pmatrix}
\tag{2.2}
\]

and let

\[
\gamma_\theta = \sqrt{|M_\theta|/v_{20}^2(\theta)}
\tag{2.3}
\]

Then \( \gamma_\theta \) is "the curvature of \( \mathcal{L} \) at \( \theta \) with respect to the inner product \( \mathcal{L}_\theta \)". (If we take \( k = 2 \), \( \theta = x \), \( \eta_\theta = (x,y(x)) \), and \( \mathcal{L}_\theta = \mathcal{L} \), then (2.3) reduces to (2.1).)

Again it can be shown that \( \gamma_\theta \) is the rate of change of direction of \( \eta_\theta \) with respect to arc-length along \( \mathcal{L} \). The relevant quantities are illustrated in Figure 1, where the arc-length from a given point \( \eta_{\theta_0} \) to \( \eta_\theta \) is called "\( s_\theta \)" and the angle between \( \hat{n}_\theta \) and \( \eta_\theta \) called "\( a_\theta \)". Then

\[
\gamma_\theta_0 = \left. \frac{d s_\theta}{d \theta} \right|_{\theta_0}
\tag{2.4}
\]

or equivalently \( \gamma_{\theta_0} = d \sin a_\theta/d \theta \big|_{\theta_0} \). Both \( s_\theta \) and \( a_\theta \) are defined relative to the inner product \( \mathcal{L}_\theta \).
\[
\frac{d\theta}{d\hat{\theta}} = \sqrt{\frac{\eta_\theta \hat{z}_\theta \hat{\eta}_\theta}{(\hat{\eta}_\theta' \hat{z}_\theta \hat{\eta}_\theta')^2}} (2.5)
\]
and
\[
\sin \hat{\theta}_0 = \sqrt{1 - \frac{(\hat{\eta}_\theta' \hat{z}_\theta \hat{\eta}_\theta')^2}{(\hat{\eta}_\theta' \hat{z}_\theta \hat{\eta}_\theta')(\hat{\eta}_\theta' \hat{z}_\theta \hat{\eta}_\theta')}} (2.6)
\]

(\hat{z}_\theta can be replaced by \hat{\gamma}_\theta anywhere in (2.6).) As Figure 1 indicates, for the purpose of evaluating \gamma_{\theta_0}, the k-dimensional curve \hat{z}_\theta can be considered locally as a two-dimensional curve in the plane through \eta_{\theta_0} spanned by \hat{\eta}_{\theta_0} and \hat{\eta}_{\theta_0}.

3. Curved Exponential Families.

In this section we define statistical curvature for one parameter families \mathcal{H} which are curved subsets of a larger k-parameter exponential family, "curved exponential families" for short. Denote the multiparameter family by

\[
\mathcal{g}_\eta(x) = g(x) e^{\eta' x - \psi(\eta)} (3.1)
\]
a family of densities with respect to some given measure \( m(\cdot) \), possibly discrete, on Euclidean k-space \( E^k \). Here \( \eta \in \mathcal{H} \), the subset of \( E^k \) for which \( \int_{E^k} g(x) e^{\eta' x} m(x) < \infty \). \( \mathcal{H} \), a convex set, is called the natural parameter space of the exponential family. If we define

\[
\lambda(\eta) = E_{\eta} x (3.2)
\]
the components of $\lambda$ can be obtained by differentiation of $\psi$,

$$\lambda_i(\eta) = \frac{\partial}{\partial \eta_i} \psi(\eta).$$

Moreover the covariance matrix $\Xi(\eta)$ of $x$ under $g_\eta$ has $i,j$th element equal to $\frac{\partial^2 \psi(\eta)}{\partial \eta_i \partial \eta_j}$. We denote by $\Lambda$ the set of all mean vectors $\lambda$,

$$\Lambda = \{ \lambda(\eta): \eta \in \mathcal{H} \}$$  \hspace{1cm} (3.3)

The mapping (3.2) from $\mathcal{H}$ to $\Lambda$ is one-to-one, and we will often write $\lambda$ instead of $\lambda(\eta)$, recognizing that $\lambda$ indexes the exponential family as well as $\eta$ does. $\Xi(\eta)$ has the same rank $r$ for all $\eta$, and we will assume $r \geq 2$ to avoid trivialities.

Now suppose that

$$\mathcal{L} = \{ \eta_\theta: \theta \in \Theta \}$$  \hspace{1cm} (3.4)

is a one-parameter subset in the interior of $\mathcal{H}$, where $\eta_\theta$ is a continuously twice differentiable function of $\theta \in \Theta$, an interval of the real line. Define the density $f_\theta$ to be

$$f_\theta(x) = g_{\eta_\theta}(x) e^{\eta_\theta^t x - \psi_\theta},$$  \hspace{1cm} (3.5)

where $\psi_\theta \equiv \psi(\eta_\theta)$. (Likewise $\lambda_\theta \equiv \lambda(\eta_\theta)$, $\Xi_\theta \equiv \Xi(\eta_\theta)$.) It is easy to verify that

$$\dot{\lambda}_\theta = \Xi_\theta \dot{\eta}_\theta,$$

$$\dot{\psi}_\theta = \dot{\eta}_\theta^t \lambda_\theta = E_\theta \dot{\eta}_\theta^t x$$

(3.6)

$\mathcal{F}$ will stand for the family of densities $\{f_\theta(x): \theta \in \Theta\}$, our curved exponential family.
**Definition.** \( y_\theta \), the statistical curvature of \( \mathcal{I} \) at \( \theta \), is the geometrical curvature of \( \mathcal{L} = \{ y_\theta : \theta \in \Theta \} \) at \( \theta \) with respect to the covariance inner product \( \mathcal{E}_\theta \), as defined in (2.2), (2.3).

**Example 1.** Bivariate normal. \( x \) is a bivariate normal random vector with covariance matrix \( \mathcal{I} \) and mean vector \( \eta_\theta = (\theta, (\gamma_0/2) \theta^2) \), \( \theta \in \Theta = (-\infty, \infty) \),

\[
x \sim \mathcal{N}(\eta_\theta, \mathcal{I}) .
\] (3.7)

Then \( \eta_\theta = (1, \gamma_0 \theta)^t \), \( \eta_\theta = (0, \gamma_0)^t \), and

\[
\Gamma_\theta = \begin{pmatrix}
1 + \gamma_0^2 \theta^2 & \gamma_0 \theta^2 \\
\gamma_0 \theta^2 & \gamma_0^2
\end{pmatrix}
\] (3.8)

so

\[
\gamma_0^2 = \frac{\gamma_0^2}{(1 + \gamma_0^2 \theta^2)^3} .
\] (3.9)

In particular \( \gamma_0^2 = \gamma_0^2 \), justifying the notation.

**Example 2.** Poisson Regression. \( x_1, x_2, \ldots, x_k \) are independent Poisson random variables, \( x_i \) having mean \( a + \theta b_i \), \( b_1, b_2, \ldots, b_k \) and \( a \) being known parameters. \( \Theta \) is the interval of \( \theta \) values such that \( a + \theta b_i > 0 \) for \( i = 1, 2, \ldots, k \). Since \( x = (x_1, \ldots, x_k)^t \) has a k parameter exponential family of distributions if the k means are unconstrained, we apply definition (2.2) to get the elements of \( M_\theta \).
\[ \nu_{20}(\theta) = \frac{k}{\sum_{i=1}^{k} \frac{b_i^2}{a + \theta b_i}} \quad \nu_{11}(\theta) = -\frac{k}{\sum_{i=1}^{k} \frac{b_i^3}{(a + \theta b_i)^2}} \quad \nu_{02}(\theta) = \frac{k}{\sum_{i=1}^{k} \frac{b_i^4}{(a + \theta b_i)^3}} \] (3.10)

The formula (2.3) for \( \nu_{\theta}^2 \) simplifies at \( \theta = 0 \) to

\[ \nu_{\theta}^2 = \frac{1}{a} \left[ \frac{\sum_{i=1}^{k} b_i^4}{(\sum_{i=1}^{k} b_i^2)^2} - \frac{(\sum_{i=1}^{k} b_i^3)^2}{(\sum_{i=1}^{k} b_i^2)^3} \right] \] (3.11)

The fact that the entries of \( M_\theta \) are summations follows from the independence of \( x_1, x_2, \ldots, x_k \), as mentioned in Section 6. A very similar formula holds for the analogous binomial regression model.

The Neyman-Davies model, \( x_1', x_2', \ldots, x_k' \) independent scaled \( \chi_1^2 \) random variables, \( x_i \) independent \( \chi_1^2 \), \( \delta_1, \delta_2, \ldots, \delta_k \) known constants, has the same structure. (Davies [4] uses this model, which originates in an application due to Neyman, to investigate the power of the locally most powerful test of \( \theta = 0 \) versus \( \theta > 0 \). We compare our results with his in Section 8.) By direct calculation or by the remark at the end of Section 6 we get that \( M_0 \) has elements
\[ \nu_{00}(0) = \frac{1}{k} \sum_{i=1}^{k} \sigma_{i}^2, \quad \nu_{11}(0) = -\sum_{i=1}^{k} \sigma_{i}^3, \quad \nu_{02}(0) = 2 \sum_{i=1}^{k} \sigma_{i}^4 \] (3.12)

and so

\[ \gamma^2 = 8 \left[ \frac{\sum_{i=1}^{k} \sigma_{i}^4}{(\sum_{i=1}^{k} \sigma_{i}^2)^2} - \frac{(\sum_{i=1}^{k} \sigma_{i}^3)^2}{(\sum_{i=1}^{k} \sigma_{i}^2)^3} \right] \] (3.13)

**Example 3. Autoregressive Process.** \( y_0, y_1, \ldots, y_T \) are observations of the autoregressive process \( y_0 = u_0, \ y_{t+1} = \theta y_t + \sqrt{1-\theta^2} u_{t+1} \), \( t = 1,2,\ldots,T \). Here \( u_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \), \( t = 0,1,\ldots,T \) and \( \Theta = (-1,1) \).

Writing out the likelihood function of \( (y_0, \ldots, y_T) \) shows that this is a curved exponential family with \( k = 3 \), the \( \eta \) vector being \( \eta_{\theta} = (-1 + \theta^2)/\alpha, \ \theta, \ (-1/2)/(1-\theta^2) \), with corresponding sufficient statistics \( x' = (\sum_{t=1}^{T} y_t^2, \sum_{t=1}^{T} y_t y_{t-1}, y_0^2 + y_T^2) \). For \( \theta = 0 \) the calculations are easy, yielding

\[ M_0 = \begin{pmatrix} T & 0 \\ 0 & 8T - 6 \end{pmatrix}, \quad \gamma^2_0 = \frac{8T - 6}{T^2} \] (3.18)

Much messier expressions are found for other values of \( \theta \). \( \gamma^2_\theta \) is of the form \( c_0/T + O(1/T^2) \) as \( T \to \infty \), with \( c_0 = 8, c_{.25} = 6.25, c_{.5} = 3.07, c_{.75} = .96 \). (For any \( T \), \( \gamma_{-\theta} = \gamma_\theta, \ i_{-\theta} = i_\theta \) since the
mapping \((y_0, y_1, y_2, \ldots) \rightarrow (y_0, -y_1, y_2, \ldots)\) takes \(\theta\) into \(-\theta\) while preserving the curvature and Fisher information. This family is least like a one-parameter exponential family at \(\theta = 0\).

If \(\mathcal{L}\) is a straight line through \(\gamma\), \(\eta_\theta = a + b\tau(\theta)\) real-valued function of \(\theta\), then \(\gamma_\theta = 0\) for all \(\theta\) since the curvature of a straight line is zero. In this case \(f_\theta(x) = (g(x)e^{b'x}) \exp[\tau(\theta)b'x - \psi_\theta]\) is a one-parameter exponential family with natural parameter \(\tau(\theta)\) and sufficient statistic \(b'x\). Under our definition all one-parameter exponential families \(\mathcal{F}\), and only such families, have statistical curvature everywhere equal to zero. This desirable property would still hold if we defined the curvature with respect to an inner product other than \(\mathcal{E}_\theta\), say \(\mathcal{E}_\theta^{-1}\) or \(\mathcal{I}\). The following discussion adds support to the choice \(\mathcal{E}_\theta\), as further shown in Section 4.

Let \(\ell_\theta(x)\) denote the logarithm of \(f_\theta(x)\),

\[
\ell_\theta(x) = \log f_\theta(x) \tag{3.15}
\]

and denote the first and second partial derivations with respect to \(\theta\) by

\[
\ell_\theta(x) = \frac{\partial}{\partial \theta} \ell_\theta(x), \quad \ell''_\theta(x) = \frac{\partial^2}{\partial \theta^2} \ell_\theta(x). \tag{3.16}
\]

The moment relationships

\[
E_\theta \ell_\theta = 0, \quad E_\theta \ell''_\theta = -E_\theta \ell'_\theta = i_\theta, \tag{3.17}
\]

11
where $i_\theta$ is Fisher's information, hold because the exponential family structure (3.1)-(3.5) allows us to differentiate under integral signs with impunity. (We will suppress the random element "x" in much of the subsequent notation.)

Notice that $\lambda_\theta(x) = \eta_\theta'x - \psi_\theta + \log g(x)$ so that

$$
\lambda_\theta(x) = \eta_\theta'(x-\lambda_\theta), \quad \lambda_\theta(x) = \eta_\theta'(x-\lambda_\theta) - \eta_\theta'\xi_\theta \eta_\theta',
$$

where we have made use of (3.6) in taking the derivatives. Remembering that $\xi_\theta$ is the covariance matrix of $x$, we see that (3.17) holds with

$$
i_\theta = \eta_\theta'\xi_\theta \eta_\theta
$$

As a matter of fact the covariance matrix of $(\lambda_\theta, \dot{\lambda}_\theta)$ is

$$
E_\theta \begin{pmatrix}
\dot{\lambda}_\theta \\
\ddot{\lambda}_\theta + i_{\theta}
\end{pmatrix} =
\begin{pmatrix}
\eta_\theta'\xi_\theta \eta_\theta & \eta_\theta'\xi_\theta \eta_\theta \\
\eta_\theta'\xi_\theta \eta_\theta & \eta_\theta'\xi_\theta \eta_\theta
\end{pmatrix}
$$

which is just the matrix $M_\theta$ defined at (2.2). Therefore

$$
\nu_{20}(\theta) = i_{\theta} = E_\theta \dot{\lambda}_\theta^2, \quad \nu_{11}(\theta) = E_\theta \dot{\lambda}_\theta \ddot{\lambda}_\theta = \text{cov}_\theta(\dot{\lambda}_\theta, \ddot{\lambda}_\theta),
$$

$$
\nu_{02}(\theta) = E_\theta \dot{\lambda}_\theta^2 - i_{\theta}^2 = \text{Var}_\theta \dot{\lambda}_\theta
$$

These definitions make no explicit reference to the geometrical structure of the curved exponential family. We will use them in Section 5 to provide the curvature definition for an arbitrary one-parameter family.
The two definitions of $M_{\theta}$, the geometrical one following (5.6) and the statistical one (5.21) give two useful invariance properties of the curvature $\gamma_{\theta}$.

i) Statistical curvature is an intrinsic property of the family $\mathcal{M}$ and does not depend on the particular parameterization used to index $\mathcal{M}$. If we let $\tilde{\theta} = g(\theta)$, where $g$ is any strictly monotone twice differentiable function, and $\tilde{r}_{\tilde{\theta}}(x) = g^{-1}(\tilde{\theta})(x)$, then $\tilde{\gamma}_{\tilde{\theta}} = \gamma_{\tilde{\theta}} = g^{-1}(\tilde{\theta})$ for every $\tilde{\theta} \in \tilde{\mathcal{M}} = g(\theta)$. This follows from the same property of the geometrical curvature (2.3). [Note: this is not true for the Fisher information: $\tilde{\gamma}_{\tilde{\theta}} = \frac{1}{g^{-1}(\tilde{\theta})} \frac{d^2}{d\tilde{\theta}^2}$.]

ii) If $t = T(x)$, is sufficient for $\theta$ then $\frac{\partial M_{\theta}^T}{\partial \theta} = \log f_{\theta}(t) = \tilde{k}_{\tilde{\theta}}(x)$, where $f_{\theta}^T$ indicates the density of $T$, implying by (3.18) that $M_{\theta}^T = M_{\theta}$ and $\gamma_{\theta}^T = \gamma_{\theta}$. The statistical curvature is invariant under any mapping to a sufficient statistic, including of course all one-to-one mappings of the sample space. This property would not hold if we had chosen an inner product other than $\mathbf{\gamma}_{\theta}$ in the definition of statistical curvature.

We can use property (ii) to transform an arbitrary curved exponential family into a form particularly convenient for theoretical calculations. Let $\theta_0$ be some value of $\theta$ at which we wish to investigate the local behavior of $\mathcal{M}$. Write $\mathbf{\gamma}_{\theta_0} = \mathbf{\Lambda}' \mathbf{\Lambda}$, $\mathbf{\Lambda}$ an $r \times r$ diagonal matrix with positive diagonal elements and $\mathbf{\Delta}$ an $r \times k$ matrix with orthonormal rows, $\mathbf{\mathbb{A}'} = \mathbf{I}_r$ (rank $\mathbf{\mathbb{A}} = r$, $\mathbf{I}_r$ the $r \times r$ identity matrix). Let $\tilde{x} = \mathbf{\Gamma} \mathbf{\mathbb{D}}^{-1/2} \mathbf{\mathbb{A}}(x-x_0)$ where $\mathbf{\Gamma}$ is an as yet unspecified
r × r orthogonal matrix. \( \tilde{x} \) is an r-dimensional sufficient statistic for the family (3.1). For \( \theta \in \Theta \) it has a curved exponential family of densities where we can take \( \tilde{\eta}_0 = \frac{1}{2} \Lambda (\eta_0 - \eta_0^0) \). (These statements are trivially true in the full rank case \( r = k \) and not difficult for \( r < k \).)

Notice that \( \tilde{\eta}_0 = \mathbb{Q} \), \( \tilde{\lambda}_0 = \mathbb{Q} \), and \( \tilde{\xi}_0 = \mathbb{I}_r \). Proper choice of the rotation matrix \( \Gamma \) makes \( \tilde{\eta}_0 \) proportional to \( \xi_1 = (1,0,\ldots,0)' \) and \( \tilde{\eta}_0 \) a linear combination of \( \xi_1 \) and \( \xi_2 = (0,1,0,\ldots,0)' \). By (3.6), \( \tilde{\lambda}_0 \) is then also proportional to \( \xi_1 \).

**Definition.** The family \( \mathcal{H} \) is in standard form at \( \theta = \theta_0 \) if \( k = r \), the dimension of \( \mathcal{H} \),

\[
\eta_0 = \lambda_0 = 0, \quad \xi_0 = \mathbb{I}_r \tag{4.1}
\]

and

\[
\hat{\eta}_0 = \hat{\lambda}_0 = \sqrt{\overline{\eta}_0} \xi_1, \quad \hat{\eta}_0 = \frac{\nu_{11}(\theta_0)}{\sqrt{\overline{\eta}_0}} \xi_1 + \overline{\eta_0} \gamma_0 \xi_2 \tag{4.2}
\]

(The constants in (4.2) are necessary to satisfy (2.2).) We will use standard form to simplify proofs in Sections 9 and 10.

If \( \mathcal{H} \) is not in standard form at \( \theta_0 \), the above transformation makes it so, and by property (ii) \( M_\theta \) and hence all information and curvature properties remain unchanged. We could use property (i) to further standardize the situation so that \( i_{\theta_0} = 1, \nu_{11}(\theta_0) = 0 \), but that does not simplify any of the theoretical calculations which follow.

Property (i) is useful for calculating curvatures, as will be shown in Section 7.
5. General Definition of Statistical Curvature.

Leaving exponential families, let

\[ \mathcal{J} \equiv \{ \theta(x), \theta \in \Theta \} \]  \hspace{1cm} (5.1)

be an arbitrary family of density functions indexed by the single parameter \( \theta \in \Theta \), a possibly infinite interval of the real line. The sample space \( \mathcal{X} \) and carrier measure for the densities can be anything at all so we have not excluded the possibility that \( \mathcal{J} \) consists of discrete distributions. Let

\[ l_\theta(x) = \log f_\theta(x), \quad l_\theta(x) = \frac{\partial}{\partial \theta} l_\theta(x), \quad l_\theta(x) = \frac{\partial^2}{\partial \theta^2} l_\theta(x) \]  \hspace{1cm} (5.2)

as in (3.12), (3.13). We assume the derivatives exist continuously and can be uniformly dominated by integrable functions in a neighborhood of the given \( \theta \), so that \( E_\theta \hat{l}_\theta = 0, E_\theta \hat{l}_\theta^2 = -E_\theta \hat{l}_\theta = 1 \) as in (3.14). Finally, as in (3.17)-(3.18) we let \( M_\theta \) be the covariance matrix of \( (\hat{l}_\theta, \ddot{l}_\theta) \),

\[ M_\theta = \begin{pmatrix} v_{20}(\theta) & v_{11}(\theta) \\ v_{11}(\theta) & v_{02}(\theta) \end{pmatrix} = \begin{pmatrix} E_\theta \hat{l}_\theta^2 & E_\theta \hat{l}_\theta \ddot{l}_\theta \\ E_\theta \hat{l}_\theta \ddot{l}_\theta & E_\theta \ddot{l}_\theta^2 \end{pmatrix} \]  \hspace{1cm} (5.3)

and define the statistical curvature of \( \mathcal{J} \) at \( \theta \) to be
\[ \gamma_\theta = \sqrt{M_0} \Big/ i_\theta^3 = \sqrt{\frac{v_{02}(\theta)}{i_\theta^2} - \frac{v_{11}(\theta)}{i_\theta^2}} \]  
\hspace{0.5cm} (5.4)

In making this definition we assume \( 0 < i_\theta < \infty \) and \( v_{02}(\theta) < \infty \). Properties (i) and (ii) of Section 4 are verified to hold for \( \gamma_\theta \) as defined in (5.4).

What does \( \gamma_\theta \) measure in this general situation? It is a measure of how quickly Fisher's score statistic is changing (more precisely, "turning") as \( \theta \) changes. An argument along those lines is given next, further support coming in the calculations of Section 8.

Comparing (5.3) with (2.2), we can connect the two definitions by thinking of \( \mathcal{L} = \{ \lambda_\theta, \theta \in \Theta \} \) as a curve through the space of random variables on \( \mathcal{L} \). The inner product \( \langle u, v \rangle_\theta = u' \mathbf{X}_\theta v \) of (2.2) is taken to be the covariance inner product in (5.3). (Section 3 makes the analogy precise in the exponential family case.) All of the quantities in Figure 1 can now be given a statistical interpretation.

The element of arc length along \( \mathcal{L} \), by analogy with (2.5), is \( ds_\theta/\theta = \sqrt{E_{\theta} \mathbf{X}_\theta^2} = \sqrt{i_\theta} \). Define

\[ u_\theta(x) = \frac{\lambda_\theta(x) + \theta}{i_\theta} \]  
\hspace{0.5cm} (5.5)

\( u_\theta \) is the version of Fisher's score statistic \( \lambda_\theta \) that is the best locally unbiased estimator for \( \theta \) near \( \theta_0 \); \( \text{var}_{\theta_0} u_\theta = 1/i_\theta \).
the Cramer-Rao lower bound, and \( E_{\theta_0} U_{\theta_0} = \theta_0, \frac{dE_{\theta_0} U_{\theta_0}}{d\theta} \bigg|_{\theta = \theta_0} = 1 \). Therefore

\[
\frac{d}{d\theta} \frac{E_{\theta} U_{\theta_0}}{\sqrt{\text{Var}_{\theta} U_{\theta_0}}} \bigg|_{\theta = \theta_0} = \frac{ds_{\theta}}{d\theta} \bigg|_{\theta = \theta_0} \tag{5.6}
\]

(The quantity on the left of (5.6) is called the "efficacy" of the statistic \( U_{\theta_0} \).) We see that

\[
\frac{ds_{\theta}}{d\theta} \bigg|_{\theta = \theta_0} \cdot (\theta - \theta_0) = \frac{E_{\theta} U_{\theta_0} - E_{\theta_0} U_{\theta_0}}{\sqrt{\text{Var}_{\theta_0} U_{\theta_0}}} + o(\theta - \theta_0)^2 \tag{5.7}
\]

Therefore \( s_{\theta} \) of Figure 1 can be interpreted locally as the number of \( (\theta_0) \) standard deviations from \( E_{\theta_0} U_{\theta_0} \) to \( E_{\theta} U_{\theta} \).

By analogy with (2.6)

\[
\sin a_{\theta} = \sqrt{1 - \frac{[\text{cov}_{\theta_0} (k_{\theta_0}, k_{\theta})]^2}{\text{Var}_{\theta_0} k_{\theta_0} \cdot \text{Var}_{\theta_0} k_{\theta}}} = \sqrt{1 - \text{corr}^2_{\theta_0} (k_{\theta_0}, k_{\theta})}, \tag{5.8}
\]

so \( \sin^2 a_{\theta} \) is interpreted as the unexplained/variance in \( U_{\theta_0}(x) \) after linear regression on \( U_{\theta_0}(x) \), under density \( f_{\theta_0} \).

From (2.4) we get the following interpretation of the statistical curvature: \( \gamma_{\theta_0} \) is the derivative at \( \theta = \theta_0 \) of the unexplained/standard deviation of \( U_{\theta} \) given \( U_{\theta_0} \), the derivative being taken with respect to the efficacy distance \( (E_{\theta} U_{\theta_0} - E_{\theta_0} U_{\theta_0})/\sqrt{\text{Var}_{\theta_0} U_{\theta_0}} \) along \( \Delta \).
If this quantity is large then the locally best estimator (also the locally best test statistic) is changing quickly as $\theta$ changes and $\mathcal{H}$ is highly curved in a statistical sense. At the opposite extreme are one-parameter exponential families for which $a_{\theta} = 0$, so $U_{\theta}$ is statistically equivalent to $U_{\theta_0}$ for all $\theta$ and $\theta_0$. We pursue this interpretation of $r_\theta$ in Section 8 to decide what constitutes a seriously large value of the curvature.

In a certain sense any smooth one-parameter family $\mathcal{H}$ can be embedded in a suitably large exponential family. Suppose at some point $\theta_0$ in $\Theta$, $l_\theta$ is $k$ times differentiable. Consider the $k$-parameter exponential family

\[
g_\eta(x) = \exp[l_\theta_0(x) + \eta_1 \lambda_{\theta_0}(x) + \eta_2 \lambda_{\theta_0}(x) + \cdots + \eta_k \lambda_{\theta_0}^{(k)}(x) - \psi(\eta)], \tag{5.9}
\]

\[
l_{\theta_0}^{(k)}(x) = \frac{\partial^k}{\partial \theta^k} l_\theta(x) \bigg|_{\theta = \theta_0}, \quad \psi(\eta) \text{ being chosen to make (5.9) integrate}
\]
to one over $\mathcal{X}$ with respect to the carrying measure for $\mathcal{H}$. Choosing

\[
\eta_\theta = \left( \frac{(\theta - \theta_0)^2}{2}, \frac{(\theta - \theta_0)^3}{3}, \ldots, \frac{(\theta - \theta_0)^k}{k!} \right)
\]
gives a one-parameter family of densities $\tilde{f}_\theta = g_\eta$ approximating $f_\theta$ near $\theta = \theta_0$. If the Taylor expansion for $l_\theta$ converges at $\theta_0$ this approximation becomes increasingly accurate as $k \to \infty$. For any value of $k \geq 2$ definitions (5.3) and (3.21) show that $\tilde{M}_\theta = M_{\theta_0}$, so $\tilde{I}_\theta = I_{\theta_0}$.
and \( \hat{\gamma}_\theta = \gamma_\theta \). It is reasonable to expect results proved in the context of curved exponential families to hold for sufficiently smooth non-exponential families, though no justifying theorem has been proved to this effect. This is in the same spirit as approximating an arbitrary family by a multinomial with a large number of categories, as in Fisher \cite{7} and Barnett \cite{2}, but seems to make the approximation in a smoother way.

6. Repeated Sampling.

Suppose we sample \( x_1, x_2, \ldots, x_n \) independently and identically distributed with density \( f_\theta \). We will use boldface letters to indicate quantities connected with the repeated sample, \( \bar{x} = (x_1, x_2, \ldots, x_n) \),

\[ \hat{\xi}_\theta(\bar{x}) = \sum_{i=1}^{n} \hat{\xi}_\theta(x_i), \quad \hat{\eta}_\theta(\bar{x}) = \frac{\hat{\xi}_\theta(\bar{x})}{\hat{\gamma}_\theta} + \theta, \text{ etc.} \]

In particular

\[ M_\theta = nM_\theta \quad (6.1) \]

since, \( M_\theta \) is the covariance matrix of \( (\hat{\xi}_\theta(\bar{x}), \hat{\eta}_\theta(\bar{x})) = \sum_{i=1}^{n} (\hat{\xi}_\theta(x_i), \hat{\eta}_\theta(x_i)) \).

Besides the familiar relationship \( \hat{\gamma}_\theta = n\gamma_\theta \), this gives

\[ \hat{\gamma}_\theta = \frac{\gamma_\theta}{\sqrt{n}} \quad (6.2) \]

The curvature goes to zero at rate \( 1/\sqrt{n} \) under repeated sampling.

This makes sense since we know that linear methods work better in large samples.
In curved exponential families, (3.18)-(3.19) combine with
\[ \mathcal{L}_0(\bar{x}) = \sum_{r=1}^{n} \mathcal{L}_0(x_r) \] to give
\[ \tilde{\mathcal{L}}_0(\bar{x}) = n \eta_0'(\bar{x} - \lambda_0), \quad \tilde{\mathcal{I}}_0(\bar{x}) = n(\eta_0'(\bar{x} - \lambda_0) - n \eta_0), \tag{6.3} \]
\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \] being the sufficient statistic for the complete family (3.1).

If the \( x_i \) are independent but not necessarily identically distributed we still have \( \mathcal{L}_0(\bar{x}) = \sum_{i=1}^{n} \mathcal{L}_0^{(i)}(x_i) \), the superscript indicating the distribution for \( x_i \), and so \( M_0 = \sum_{i=1}^{n} M_0^{(i)} \). This explains the simple form of \( M_0 \) in example 2 of Section 3.

7. Some Examples.

Before discussing the statistical properties of \( \gamma_0 \) we will expand our catalog of examples to include several non-exponential families. Those results illustrate some simple principles that make \( \gamma_0 \) easy to calculate in familiar statistical situations. In the first three examples we assume that the densities given are with respect to Lebesgue measure on the real line, i.e., that we have just one observation of a continuous variable. For an i.i.d. sample of size \( n \) the curvature is obtained from formula (6.2). This last remark applies also to the last example, and to the examples of Section 3.
Example 4. Translation families. Let $g(x)$ be a probability density function and $f_0(x) = g(x-\theta)$. Also let $h(x) = \log g(x)$. Then

$$
\kappa_0(x) = -h^{(1)}(x-\theta), \quad \lambda_0(x) = h^{(2)}(x-\theta),
$$

where $h^{(1)}(x) = \frac{d}{dx} h(x)$,

so

$$
P_{\theta} x_{\theta} = \int_{-\infty}^{\infty} [\cdot h^{(1)}(x)] \cdot [h^{(2)}(x)]^j \cdot g(x) \, dx.
$$

Obviously $M_\theta$ and $r_\theta$ do not depend on $\theta$ in a translation family.

For the $t$ translation family, $f$ degrees of freedom,

$$
g(x) = \frac{\Gamma(f+1)}{\sqrt{f} \cdot \Gamma(\frac{1}{2}) \cdot \Gamma(\frac{f}{2})} \left(1 + \frac{x}{t}\right)^{-(f+1)/2} \tag{7.1}
$$

we calculate

$$

\nu_{20}(\theta) = 1(\theta) = \frac{f+1}{f+3}, \tag{7.2}
$$

$$
\nu_{02}(\theta) = \frac{f+1}{f+3} \left[ \frac{(f+2)(t^2+8f+19)}{f(f+5)(f+7)} - \frac{f+1}{f+3} \right]
$$

and $\nu_{11}(\theta) = 0$ (by symmetry), giving

$$
\gamma_\theta^2 = \frac{6\left(\frac{1}{2}t^2 + 18f + 19\right)}{f(f+1)(f+5)(f+7)} \tag{7.3}
$$

a monotone decreasing function of $f$. Some values are as follows:

<table>
<thead>
<tr>
<th>$f$</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>$\to \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_\theta^2$</td>
<td>2.5</td>
<td>1.063</td>
<td>.506</td>
<td>.107</td>
<td>.0334</td>
<td>$\sim 18/f^2$</td>
</tr>
</tbody>
</table>

21
The case $f = 1$ is the Cauchy translation family, and the value $5/2$ for $\gamma_\theta^2$ agrees with a closely related calculation in Fisher [7].

For the Gamma translation family

$$f_\theta(x) = \frac{(x-\theta)^{a-1} e^{-(x-\theta)}}{\Gamma(a)}, \quad x \geq \theta$$

(7.5)

For $a > 4$ a fixed constant, we calculate

$$\gamma_\theta^2 = \frac{2}{(a-3)^2} \frac{a-1}{a-4}$$

(7.6)

(For $a \leq 4$, $\gamma_\theta^2$ is infinite.)

**Example 5. Scale families.** $x \sim \theta \cdot z$ where $z$ has a known density, $\theta \in \Theta = (0, \infty)$. If $z$ is a positive random variable then $\log x = \log \theta + \log z$ is a translation family. By Section 4 the curvature will be the same for this family as for the original one, and by Example 4 it will not depend on $\theta$: For scale families $\gamma_\theta$ does not depend on $\theta$. (The argument above applied separately to the positive and negative axes gives the result in general. It can also be derived directly from (5.3).)

A particular example is the normal with known coefficient of variation, $x \sim \mathcal{N}(\theta, c^2)$, $c$ known. Here $x \sim \theta z$, $z \sim \mathcal{N}(1, c)$. We calculate $\gamma_\theta = 2(c + \frac{1}{2})/(c^2)$ and

$$\gamma_\theta^2 = \frac{c^2}{4(c + \frac{1}{2})^3}$$

(7.7)
(Notice that $x \sim \mathcal{N}(\theta, \sigma^2)$ is a curved exponential family, $k = 2$.)

The curvature is near 0 for all values of $c$, taking its maximum at $c = 1$:

$$
\begin{array}{|c|c|c|c|c|c|}
\hline
 c & 1/4 & 1/2 & 1 & 2 & 4 & \rightarrow \infty \\
\hline
Y_\theta^2 & .0370 & .0625 & .0740 & .0640 & .0439 & \sim \frac{1}{4c} \\
\hline
\end{array}
$$

(7.8)

**Example 6. Weibull Shape Parameter.** $f_\theta(x) = \theta x^{\theta-1} e^{-x^\theta}$ for $x \geq 0$, $\theta \in \Theta = (0, \infty)$. That is $x \sim z^{1/\theta}$, where $P(z < z_0) = 1 - e^{-z_0}$ for $z_0 \geq 0$. The transformation $\log x = \frac{1}{\theta} \log z$ makes this a scale family in $1/\theta$, so once again $Y_{\theta}^2$ does not depend on $\Theta$. Taking $\theta = 1$ for convenience gives $\mathbf{i}_1(x) = (1-x) \log x + 1$, $\mathbf{i}_2(x) = - (x \log^2 x + 1)$, $E_\theta \int_0^\infty [\mathbf{i}_1(x)]^4 [\mathbf{i}_1(x)]^j e^{-x} \, dx$. Numerical integration gives

$$
Y_{\theta}^2 = .704
$$

(7.9)

**Example 7. Mixture Problems.** $f_\theta(x) = (1-\theta) g(x) + \theta h(x)$, $g$ and $h$ known densities on an arbitrary space $\mathcal{X}$. The parameter space $\Theta$ contains $[0, 1]$. We see that

$$
\mathbf{i}_\theta = \frac{h - g}{g + \theta(h - g)}, \quad \mathbf{j}_\theta = - \mathbf{j}_\theta^2
$$

(7.10)

and for $\theta = 0$.
\[ k_0 = r - 1, \quad k_0 = -(r-1)^2 \] (7.11)

where \( r(x) = h(x)/g(x) \). Defining \( \alpha_j = E_j(r - 1)^j \) gives

\[
M_0 = \begin{pmatrix}
\alpha_2 & -\alpha_3 \\
-\alpha_3 & \alpha_4 - \alpha_2^2 \\
\end{pmatrix}, \quad r_0^2 = \frac{\alpha_4 - \alpha_2^2}{\alpha_2^2} = \frac{\alpha_2^2}{\alpha_2^2} \] (7.12)

If \( g \) and \( h \) are normal densities, say \( g(x) = \varphi(x) = e^{-x^2/2}/\sqrt{2\pi} \), \( h(x) = \varphi(x-\mu) \), we have \( r(x) = \exp(\mu x - \mu^2/2) \) and

\[
M_0 = \begin{pmatrix}
\xi - 1 & -[\xi^3 - 3\xi + 2] \\
-\xi^3 + 3\xi - 2 & \xi^6 - 4\xi^3 + 8\xi - 4 \\
\end{pmatrix} \] (7.13)

when \( \xi = e^{\mu^2} \). Therefore \( r_0 = \xi - 1 \) and

\[
r_0^2 = \xi^3(\xi + 1) \] (7.14)

The curvature approaches 2 for \( \mu \) near 0 but becomes enormous as \( \mu \) increases,

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>.5</th>
<th>.832</th>
<th>1</th>
<th>1.048</th>
<th>1.180</th>
<th>( \to \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_0^2 )</td>
<td>4.84</td>
<td>24</td>
<td>74.68</td>
<td>108</td>
<td>320</td>
<td>( \sim e^{4\mu^2} )</td>
</tr>
<tr>
<td>( r_0 )</td>
<td>.284</td>
<td>1</td>
<td>1.718</td>
<td>2</td>
<td>3</td>
<td>( \sim e^{2\mu} )</td>
</tr>
</tbody>
</table>

(7.15)
8. **Hypothesis Testing.**

So far we have not tried to say what constitutes a "large" curvature—a value of $r_\theta$ (or, in repeated sampling situations of $Y_\theta$, the curvature based on all the data) of sufficient magnitude to undermine techniques based on linear approximations to the log likelihood function. We can obtain a rough idea of this value by considering the problem of testing $H_0: \theta = \theta_0$ versus $A_0: \theta > \theta_0$.

Define

$$\theta_1 = \theta_0 + \frac{2}{\sqrt{1_\theta}}$$  \hspace{1cm} (8.1)

so that $\sqrt{1_\theta}(\theta_1 - \theta_0) = 2$. From the discussion (5.5)-(5.7) this means that, approximately,

$$\frac{E_{1,0} U_0 - E_{0,0} U_0}{\sqrt{\text{Var}_{\theta_0} U_{\theta_0}}} = 2$$  \hspace{1cm} (8.2)

(where in (5.7) we have used $\left.\frac{d s_\theta}{d \theta}\right|_{\theta = \theta_0} = \sqrt{1_\theta}$). The locally most powerful level $\alpha$ test of $H_0$ versus $A_0$, LMP$_\alpha$ for short, rejects for large values of $U_0$. From (8.2) we would expect LMP$_\alpha$ to have reasonable power at $\theta_1$ for the customary values of $\alpha$. That is $\theta_1$ should be a "statistically reasonable" alternative to $\theta_0$.

The discussion following (5.8) shows that the unexplained variance of $U_{\theta_1}$ after linear regression on $U_{\theta_0}$, calculated under
\( f_{\theta_0} \), is approximately \( 4\gamma_{\theta_0}^2 \). If this quantity is large, say \( 4\gamma_{\theta_0}^2 \geq \frac{1}{2} \), then \( U_{\theta_1} \) differs considerably from \( U_{\theta_0} \), and the test of \( H_0 \) based on \( U_{\theta_1} \) will substantially differ from that based on \( U_{\theta_0} \). Under these circumstances it is reasonable to question the use of LMP\(_\alpha\). Based on those very rough calculations a value of \( \gamma_{\theta_0}^2 \geq 1/8 \) is "large".

In the repeated sampling situation of Section 6 a sample of size \( n > n_0 \),

\[
n_0 = 8\gamma_{\theta_0}^2, \tag{8.3}
\]

makes \( \chi_{\theta_0}^2 = \gamma_{\theta_0}^2 / n < 1/8 \), and therefore reduces the curvature below the worrisome point. For the Cauchy translation family, Example 4, \( n_0 = 20 \). For the Weibull shape parameter, Example 6, \( n_0 = 5.6 \). For the normal with known coefficient of variation, Example 5, \( n_0 < 1 \) for all \( c \).

At the opposite extreme we have the normal mixture problem, Example 7, with \( \mu = 1 \), for which \( n_0 = 597.4 \). We expect linear methods to work well in Example 5 for any sample size, and poorly in the last example, even for large samples.

Moving from the vague to the specific, consider Example 1, Section 3, a bivariate normal vector \( x = (x_1, x_2)' \) with mean \( (\theta, \gamma_0 \theta^2/2) \) and covariance matrix \( I \). Assume we wish to test \( H_0: \theta = 0 \) versus \( A_0: \theta > 0 \) on the basis of observing \( x \). The LMP\(_\alpha\), which rejects for large values of \( x_1 \), has power function (probability of rejection)

\[
1 - \beta_0(\theta) = \Phi(\theta - z_\alpha),
\]
where \( z_\alpha \) and \( \phi \) are the upper \( \alpha \) point and cdf of a standard normal variate.

Figure 2. Bivariate Normal, Example 1, testing \( \theta = 0 \) versus \( \theta > 0 \).

The rejection region for the locally most powerful level \( \alpha \) test, \( \text{LMP}_{\alpha'} \), is compared with that for the most powerful level \( \alpha \) test of \( \theta \) versus \( \theta_1 \), \( \text{MP}_{\alpha}(\theta_1) \).

The Neyman-Pearson lemma says that the most powerful level \( \alpha \) test of \( \theta = 0 \) versus some specific positive alternative \( \theta = \theta_1 \), \( \text{MP}_{\alpha}(\theta_1) \) for short, rejects for large values of \( \eta_0'x \). It has power function
\[ 1 - \beta_{\theta_1}(\theta) = \Phi \left( \frac{\theta \sqrt{1 + y_0^2 \theta^2 / 4} \cos(A_{\theta_1} - A_\theta) - z_\alpha}{\theta} \right), \]  

\text{for} \ A_\theta \being the angle from the x_1 axis to \eta_\theta, \text{illustrated in Figure 2.}

As \ \theta_1 \to 0, \ \beta_{\theta_1}(\theta) \to \beta_0(\theta) \text{ for all } \theta, \text{justifying the notation } 1 - \beta_0(\theta) \text{ for the power function of LMP}_{\alpha}.

For a given value of \ \theta > 0, the power is maximized by taking \ \theta_1 = 0, \text{giving "power envelope"}

\[ 1 - \beta^*(\theta) = \Phi \left( \frac{\theta \sqrt{1 + y_0^2 \theta^2 / 4} - z_\alpha}{\theta} \right). \]  

Figure 3 compares the power envelope function, for four values of \ y_0', \text{with the power function of LMP}_{\alpha}, \ \alpha = .05 \text{ (which does not depend on } y_0').

As predicted the difference between \ 1 - \beta^*(\theta) \text{ and } 1 - \beta_0(\theta) \text{ increases}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Power of LMP}_{\alpha}, \ \alpha = .05, \text{ compared with power envelope function, Example 1.}
\end{figure}
with the curvature $\gamma_0$. In this case we can actually see that $\gamma_0$ measures how fast the locally optimum test statistic $U_0(x)$ becomes non-optimal as the alternative $\theta$ increases from 0. Also according to prediction the LMP$_\alpha$ has reasonable power properties for $\gamma_0^2 = 1/16$ and poor properties for $\gamma_0^2 \geq 1/4$.

Of course no level $\alpha$ test can achieve the power envelope for more than one value of $\theta > 0$. MP$_\alpha(\theta_1)$ achieves it for $\theta = \theta_1$ while LMP$_\alpha$ optimizes for $\theta$ near 0 in the sense that $\left. d\beta_0(\theta)/d\theta \right|_{\theta = \theta_0} = d\beta^*(\theta)/d\theta \bigg|_{\theta = \theta_0}$. By following prescription (8.1) in choosing $\theta_1$ we get a test which matches the power envelope at what should be a statistically interesting value of $\theta$, one where the power is reasonably but not unreasonably high. In our example this means choosing $\theta_1 = 2$, since $i_0 = 1$. Table 1 shows that $1 - \beta_2(\theta)$ stays remarkably close to $1 - \beta^*(\theta)$, and that MP$_{0.05}(2)$ has better power characteristics than LMP$_{0.05}$, especially for large values of $\gamma_0$.

Davies performs similar evaluations for the Neyman-Davies model of Example 2. The curvatures for the upper and lower cases graphed on page 532 of [4] are $\gamma_0^2 = .488$ and $\gamma_0^2 = .244$ respectively, while the two on page 533 are $\gamma_0^2 = .00629$ and $\gamma_0^2 = .0364$. Ignoring the "Wald's test" curve, one sees that the magnitude of $\gamma_0^2$ is indeed a good predictor of the relative performance of LMP$_\alpha$ compared to MP$_\alpha(\theta_1)$. His results are quite similar to those for our Example 1. (Davies choses $\theta_1$ so that $1 - \beta^*(\theta_1) = .8$. This is a more precise way of accomplishing what (8.1) is intended to do, but is computationally difficult in most situations.)
<table>
<thead>
<tr>
<th>$\theta$</th>
<th>0</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image1" alt="" /></td>
<td>.05</td>
<td>.126</td>
<td>.262</td>
<td>.453</td>
<td>.662</td>
<td>.835</td>
<td>.941</td>
<td>.985</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.125</td>
<td>.260</td>
<td>.452</td>
<td>.662</td>
<td>.834</td>
<td>.938</td>
<td>.983</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.126</td>
<td>.260</td>
<td>.442</td>
<td>.639</td>
<td>.804</td>
<td>.912</td>
<td>.968</td>
</tr>
<tr>
<td>.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.127</td>
<td>.269</td>
<td>.483</td>
<td>.723</td>
<td>.904</td>
<td>.982</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.121</td>
<td>.261</td>
<td>.479</td>
<td>.723</td>
<td>.901</td>
<td>.980</td>
<td>.998</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.126</td>
<td>.260</td>
<td>.442</td>
<td>.639</td>
<td>.804</td>
<td>.912</td>
<td>.968</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.129</td>
<td>.300</td>
<td>.591</td>
<td>.882</td>
<td>.991</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.115</td>
<td>.280</td>
<td>.583</td>
<td>.882</td>
<td>.990</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.126</td>
<td>.260</td>
<td>.442</td>
<td>.639</td>
<td>.804</td>
<td>.912</td>
<td>.968</td>
</tr>
<tr>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.139</td>
<td>.409</td>
<td>.855</td>
<td>.998</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.115</td>
<td>.381</td>
<td>.850</td>
<td>.998</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>.126</td>
<td>.260</td>
<td>.442</td>
<td>.639</td>
<td>.804</td>
<td>.912</td>
<td>.968</td>
</tr>
</tbody>
</table>

Table 1. Power comparison, Example 1.

a) Power envelope.

b) Power MP .05(2)

c) Power IMP .05.
Section 10 shows that the Cramer-Rao lower bound for the variance of an unbiased estimator erra roughly by a factor of $1 + \frac{2}{L_0}$, rejustifying the definition of $\frac{2}{L_0} \geq \frac{1}{8}$ as a "large" curvature.


We again assume an i.i.d. sample $x_1, x_2, \ldots, x_n, \ldots$ as in Section 6. Result (1.1), originally stated by Fisher in his fundamental paper on estimation theory [7], can be restated as

$$\lim_{n \to \infty} \left( I_\theta - I_\hat{\theta} \right) = I_\theta \gamma_\theta^2$$  \hspace{1cm} (9.1)

since $\gamma_\theta^2$ equals the bracketed term in (1.1). (9.1) is derived from (1.1) and (5.4) by means of the relationships $\nu_{20}(\theta) = \mu_{20} = 1_\theta$, $\nu_{11}(\theta) = \mu_{11} - \mu_{30}$, and $\nu_{02}(\theta) = \mu_{02} - 2\mu_{21} + \mu_{40} - \mu_{20}$, these following from (1.2), (5.3) and

$$\tilde{I}_\theta = \tilde{r}_\theta / r_\theta$$, \hspace{1cm} $\tilde{I}_\theta = \tilde{r}_\theta / r_\theta - (\tilde{r}_\theta / r_\theta)^2$  \hspace{1cm} (9.2)

To use Fisher's evocative language, asymptotically the MLE $\hat{\theta}(x_1, x_2, \ldots, x_n)$ extracts all but $I_\theta \gamma_\theta^2$ of the information in the sample $X = (x_1, \ldots, x_n)'$. Since a single observation contains an amount $I_\theta$ of information this is equivalent to a reduction in effective sample size from $n$ to $n - \gamma_\theta^2$, for example from $n$ to $n - 5/2$
in the Cauchy translation parameter problem. This is the price one pays for a one-dimensional summary of the data and, also according to Fisher, any summary statistic other than the MLE would pay a greater price. (Rao's substantial contributions to this argument are discussed toward the end of the section.)

The geometrical argument which follows shows clearly why the curvature $\gamma_0$ plays the role that it does in (9.1). It also leads quickly to a counterexample to (9.1) and shows that by working within multinomial families, Fisher and Rao chose perhaps the least tractable curved exponential families. We will work with a general curved exponential family in the standard form (4.1)-(4.2). For notational convenience we let $\theta_0$, a particular value of $\theta$ where we wish to evaluate $\lim_{n \to \infty} (i_{\theta} - i_{\hat{\theta}})$, equal 0. Then we have $\eta_0 = \lambda_0 = 0$, $\lambda_0 = 1$, $\eta_0 = \lambda_0 = \sqrt{10} \sigma_1$, and $\sigma_0 = (\nu_1(0)/\sqrt{10}) \sigma_1 + i_0 r \sigma_2$.

Fisher's argument depends on two useful results which we borrow:

1) If $T(\chi)$ is any statistic, with density $f_{\theta}^T(\chi)$ and score function (log derivative) $\dot{z}_{\theta}(t) = \frac{\partial}{\partial \theta} \log f_{\theta}^T(t)$, then $\dot{z}_{\theta}(t) = E_{\theta}(\dot{z}_{\theta}(\chi) | T = t)$ (where we recall from Section 6 that $\dot{z}_{\theta}(\chi)$ is the score based on all the data). This implies that the loss of information in going from $\chi$ to $T(\chi)$ is

$$i_{\theta} - i_{\hat{\theta}} = E_{\theta} \text{Var}_{\theta}(\dot{z}_{\theta}(\chi) | T)$$

(9.3)

since $i_{\theta} - i_{\hat{\theta}} = \text{Var}_{\theta} \dot{z}_{\theta} - \text{Var}_{\theta} \dot{z}_{\theta}^T$.
2) Let \( L_\theta \) be the set of values of \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i / n \) for which

\[ \ell_\theta(\bar{x}) = 0; \]

\( L_\theta \) consists of those values of the sufficient statistic \( \bar{x} \) for which \( \theta \) is a solution to the likelihood equations

\[ \ell_\theta(\bar{x}) = 0. \]

Then, since \( \ell_\theta = n\eta_\theta'(\bar{x} - \lambda_\theta) \),

\[ L_\theta = \{ \bar{x}: \eta_\theta'(\bar{x} - \lambda_\theta) = 0 \} \tag{9.4} \]

the \( r \)-dimensional hyperplane through \( \lambda_\theta \), orthogonal to \( \eta_\theta \).

---

Figure 4. A curved exponential family of dimension \( r = 2 \). \( \hat{\theta} \)
is the set of \( \bar{x} \) for which \( \hat{\theta} \) is a solution to the maximum likelihood equations. \( M_{\hat{\theta}} \) is the level curve for another consistent efficient estimator.
Figure 4 illustrates the situation for the case $r = 2$. (Notice that the sample space, the space of possible $\bar{x}$ values, has been superimposed on $\Lambda$, the space of possible mean vectors $\lambda$.) Actually this two-dimensional picture is appropriate for any dimension since curvature is locally a two-dimensional property, as pointed out at the end of Section 2. A heuristic proof of (9.1) based on this picture now follows in five easy steps:

(i) $\hat{\lambda}_0(\bar{x}) = n \sqrt{\hat{I}_0} \bar{x}_1$ (by (6.3)).

(ii) $\sqrt{n} \bar{x} \rightarrow N_r(\mu, I)$ if $\theta = 0$.

(since $\lambda_0 = \mu$, $\mu_0 = I$, and central limit conditions are satisfied inside an exponential family).

(iii) Let $\hat{\theta}$ be the MLE and $\hat{\lambda}_0$ the angle between $\hat{\lambda}_0$ and $\hat{\eta}_0 = \sqrt{\hat{I}_0} \bar{z}_1$. Then $\hat{a} = \sqrt{\hat{I}_0} \hat{\theta}^2 + o(\hat{\theta}^2)$. (Since $\eta_{\theta}^2 = \hat{\eta}_0 + o(\hat{\theta}^2) = \sqrt{\hat{I}_0} \hat{\theta}^2 + o(\hat{\theta}^2)$, the element of arc-length in Figure 1 is $s = \| \eta_{\theta}^2 \| = \sqrt{\hat{I}_0} \hat{\theta} + o(\hat{\theta}^2)$. By (2.4) we have $a_{\theta} = \hat{a} = \sqrt{\hat{I}_0} \hat{\theta} + o(\hat{\theta}^2)$.)

(iv) $\text{Var}_0(\hat{\lambda}_0(x) | \hat{\theta}) = n^2 \hat{I}_0 \tan^2 \hat{a} \cdot \text{Var}_0(\bar{x}_2 | \hat{\theta})$. (In the case $r = 2$ this follows immediately from (i) and the geometry of the situation. For $r > 2$ $\bar{x}_2$ is replaced by $\nabla^2 ||x||$ where $\nabla = \hat{\eta}_0 - \| \hat{\eta}_0^2 \| \cos \hat{a} \cdot \hat{\eta}_0$, the part of $\hat{\eta}_0$ orthogonal to $\hat{\eta}_0$.)

(v) $\text{Var}_0(\bar{x}_2 | \hat{\theta}) = \frac{1}{n} + o\left(\frac{1}{n}\right)$. (This is plausible because of (ii) and the fact that near $\theta = \hat{\theta}$ the partition of the sample space generated by the "lines" $L_{\theta}$ looks like the partition generated by lines parallel to $L_{\theta}$.)
Steps (iii) and (iv) together give \( \text{Var}_0(\hat{\beta}_0 | \hat{\theta}) = n^2 \text{Var}_0(\hat{x}_2 | \hat{\theta}) \), which, combined with (v), gives

\[
\text{Var}_0(\hat{\beta}_0(x) | \hat{\theta}) = n^2 \text{Var}_0(\hat{x}_2 | \hat{\theta}) (1 + o(\hat{\theta})) \{1 + o_n(1)\} \tag{9.5}
\]

\( o_n(1) \rightarrow 0 \) as \( n \rightarrow \infty \), \( o(\hat{\theta}) \rightarrow 0 \) as \( \hat{\theta} \rightarrow 0 \). The heuristic proof of (9.1) is completed by (9.3), giving

\[
\lim_{n \rightarrow \infty} \frac{\hat{\beta}_0 - \beta_0}{\text{Var}_0(\hat{\beta}_0 | \theta)} = \lim_{n \rightarrow \infty} \frac{\text{Var}_0(\hat{\beta}_0 | \theta)}{\text{Var}_0(\beta_0 | \theta)} = \text{Var}_0(\beta_0 | \theta) \tag{9.6}
\]

Here we have used

\[
\lim_{n \rightarrow \infty} nE_0|\hat{\theta}|^3 = 0, \quad \lim_{n \rightarrow \infty} nE_0\hat{\theta}^2 = i_0^{-1}, \tag{9.7}
\]

which one might hope for in view of \( \sqrt{n} \hat{\theta} \rightarrow \mathcal{N}(0, i_0^{-1}) \).

All of the weak links in this chain of reasoning can be made solid except for (v). Its fatal flaw is shown by a counterexample to (9.1) based on the trinomial distribution

\[
P(\text{Observed object is in category } j) = \lambda_j, \quad j = 1, 2, 3 \tag{9.8}
\]

\( \lambda_j \geq 0, \lambda_1 + \lambda_2 + \lambda_3 = 1 \).

The trinomial can be considered as an exponential family of form (3.1) with \( k=r=2; \lambda = (\lambda_1, \lambda_2)' \), \( \eta = (\eta_1, \eta_2)' \), \( \eta_j = \log[\lambda_j/(1-\lambda_1-\lambda_2)] \), \( j = 1, 2 \), and \( \psi(\eta) = \log(1 + e^{\eta_1} + e^{\eta_2}) \). The \( x \) vector takes on
three possible values: (1,0), (0,1), (0,0), corresponding to the observed object being in the first, second, or third categories respectively. The carrier measure \( m(\cdot) \) puts mass one at each of these three \( x \) values.

Figure 5. Counterexample to (9.1) based on trinomial. Each line \( L_\theta \) contains at most one possible sample point \( \bar{x} \).
The counterexample is a one parameter family $\mathcal{V}$ with $I_\theta$ passing through the fixed point $c = (\sqrt{2}, -1)$ as illustrated in Figure 5, the parameter $\theta$ being the angle between $L_0$ and $L_\theta$. Such a family does exist, as the following construction shows: let $\lambda_0 \equiv (\frac{1}{3}, \frac{1}{3})$ and

$$\lambda_\theta = \lambda_0 + \int_0^\theta \mu_\theta (\xi_\theta \phi_\theta) d\theta$$

where $\mu_\theta = ||\lambda_0 - c||/(||\xi_\theta \phi_\theta|| \sin B_\theta)$, $\xi_\theta$ is the covariance matrix of $x$ under $\Gamma_\theta$, the vector $\phi_\theta$ and the angle $B_\theta$ being defined as in Figure 5. Definition (9.9) gives $\lambda_\theta \in L_\theta$ and also that, by (3.6), $\hat{n}_0 = \rho_\theta$, the normal vector to $L_\theta$, as necessitated by (9.4).

$\mathcal{V}$ is a curved exponential family having the following property: if $\bar{x}(1)$ and $\bar{x}(2)$ are two values of $\bar{x} = \sum_{i=1}^n x_i/n$ giving the same MLE $\hat{\theta}$, then both $\bar{x}(1)$ and $\bar{x}(2)$ lie on $L_\theta$. But $\bar{x}(1) = (n_{1(1)}/n, n_{2(1)}/n)$, $i = 1, 2$, the $n_{j(1)}$ being nonnegative integers. This implies either $\bar{x}(1) = \bar{x}(2)$ or

$$\frac{n_{2(2)} - n_{2(1)}}{n_{1(2)} - n_{1(1)}} = \frac{n_{2(1)} + 1}{n_{1(1)} + \sqrt{2}}.$$ (9.10)

(9.10) would make $\sqrt{2}$ a rational number, hence $\bar{x}(1)$ must equal $\bar{x}(2)$. In short there is at most one possible $\bar{x}$ value corresponding to any $\hat{\theta}$, and so the MLE is a sufficient statistic in $\mathcal{V}$, implying $\bar{y} - \hat{\theta} = 0$ for all $n$. But $\gamma_\theta^2$ must be positive for all $\theta$ values since $\hat{\eta}_\theta$ is always changing direction. This completes the counterexample.
Remark 1. Let \( \varphi(t) = E_0 e^{it'x} \) be the characteristic function of \( f_0 \).
If \( \varphi(t) \) is integrable for some \( p \geq 1 \) then \( \sqrt{n} \bar{x} \) has a density function converging uniformly to \( (2\pi)^{-k/2} \exp(-\|x\|^2/2) \). See [6] and [8]. Under those conditions (9.1) can be verified. The technical details, which depend on an exponential bound to the density of \( \bar{x} \), are indicated in the Appendix.

Remark 2. Instead of working with the MLE \( \hat{\theta} \) itself we can consider the coarser statistic which only records which interval \( \hat{\theta} \) lies in, among intervals of the form \( (i\epsilon_n, (i+1)\epsilon_n) \), \( i = 0, \pm 1, \pm 2, \ldots \). The line \( L_\theta \) in Figure 4 is now replaced by a pair of lines \( L_{i\epsilon_n}, L_{(i+1)\epsilon_n} \), and step (v) can be weakened to say only that the conditional distribution of \( \bar{x}_2 \), given that \( \bar{x} \) is between the two lines, has variance \( 1/n + o(1/n) \). However in order for statement (iv) to still have meaning we need to take \( \epsilon_n = o(1/n) \) (so that the conditional variance of \( \bar{x}_0 \) will still be due mainly to the slope of the lines \( L_{i\epsilon_n}, L_{(i+1)\epsilon_n} \), and not to the distance between them). It turns out, [6], to be possible to choose \( \epsilon_n \) in this way and to get the proper convergence of the conditional variance if \( f_0 \) is non-lattice, \( |\varphi(t)| < 1 \) for all \( t \neq 0 \). (This excludes the multinomial.) In this case it is possible to show that \( \lim_{n \to \infty} (\bar{z}_\theta - \bar{z}_\hat{\theta})^2 \leq \sup_{\theta} \gamma^2 \).
Remark 3. If \( \tilde{\theta}(x) \) is any other consistent efficient estimator of \( \theta \), and \( M_{\theta} \) is the set of \( \tilde{x} \) values having \( \tilde{\theta}(\tilde{x}) = \theta \), then as in Figure 4, \( M_{\theta} \) passes through \( \lambda_{\theta} \) and is tangent to \( L_{\theta} \) at that point. See Section 10. The increment of \( \lim_{n \to \infty} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \theta}{\partial \theta} - \hat{\theta}_{\theta} \right) \) above zero is due to the quadratic term in the expansion of \( M_{\theta} \) near \( \lambda_{\theta} \). The details are almost identical to those of Section 10 and will not be given here. (See (10.25).)

Remark 4. It is possible for two of the surfaces (9.4), say \( L_{\theta} \) and \( \hat{L}_{\theta} \), to intersect. If \( \tilde{x} \in L_{\theta} \cap \hat{L}_{\theta} \) then both 0 and \( \hat{\theta} \) are solutions to the likelihood equation. As \( \theta \) decreases to zero in Figure 4, \( L_{\theta} \cap \hat{L}_{\theta} \) converges to a point (in general an \( r-2 \) dimensional flat) on \( L_{\theta} = \{c e_{\theta} \} \) a distance \( \rho_{\theta} = 1/\gamma_{\theta} \) above \( Q \). It turns out that values of \( \tilde{x} \) on \( L_{\theta} \) which lie above this point are local maxima of the likelihood function, while those lying below are local minima.

Remark 5. Rao [10,11,12] uses a different definition of the information which avoids the difficulty illustrated by the counterexample. (9.3) can be written as \( \frac{1}{\gamma_{\theta}} = \inf_{\theta} \left( \frac{1}{\theta} \sum_{i=1}^{n} \frac{\partial \theta}{\partial \theta} (x) - h(T(x)) \right)^2 \), the infimum being over all choices of the function \( h(\cdot) \). Rao redefines \( \frac{1}{\theta} \) by restricting the function \( h \) to be quadratic. Rao states that he believes the two definitions to be equivalent, but the counterexample can be used to show that they are not.
Remark 6. Is (9.1) a useful fact, assuming it is true? Fisher seemed to think of Fisher information as a perfect measure of the amount of information available to the statistician. For ordinary "first order efficiency" calculations in large samples this is true enough, in the following sense: let \( T(x) \) be a statistic having Fisher information \( \frac{T}{\theta} \). Then in a neighborhood of any given value \( \theta_0 \) of \( \theta \) we can construct, under suitable regularity conditions, a function \( \hat{T}(T) \), that is approximately \( \mathcal{N}(\theta, 1/\mathcal{I}_0^T) \), as compared with \( \mathcal{N}(\theta, 1/\mathcal{I}_0^\hat{\theta}) \) for the MLE. If \( \mathcal{I}_0^T/\mathcal{I}_0^\hat{\theta} = .8 \) for example, then any statistic \( h(\hat{T}) \) will have almost the same distribution as \( h(\hat{\theta}) \) with \( \hat{\theta} \) based on a sample \( \mathcal{O}_0 \) as large.

This argument breaks down for information discrepancies as small as those contemplated in (9.1), since the central limit theorem is in general not capable of supporting such fine distinctions. To give substance to Fisher and Rao's theorem we must demonstrate that in specific statistical problems the Fisher information determines relative performance at the level of accuracy suggested by (9.1). Rao [12] showed that this indeed was the case for the problem of estimating \( \theta \) with squared error loss. We review his results from the point of view of this paper in Section 10.

Suppose we wish to estimate the parameter $\theta$ in a curved exponential family on the basis of an i.i.d. sample $x_1, x_2, \ldots, x_n$, using a squared error loss function to evaluate possible estimators. We will only consider estimators that are smooth functions of the sufficient statistic $\bar{x}$ and are consistent and efficient in the usual sense (see (10.5)-(10.7) below). The following result will be discussed:

let $\hat{\theta}(\bar{x})$ be such an estimator, the form of $\hat{\theta}$ not depending on $n$, and let $\varphi(\theta) = \mathbb{E}_{\theta} U_{\theta}(\bar{x})$ where as before $U_{\theta}(\bar{x}) = \frac{\partial^2}{\partial \theta^2} + \theta$ is the best locally unbiased estimator of $\theta$ near $\theta_0$. Also let $b_\theta = \mathbb{E}_\theta \hat{\theta}(\bar{x}) - \theta$ be the bias of $\hat{\theta}$, a quantity which will turn out to be of order $O(1/n)$ in the theory below. Then

$$\text{Var}_{\theta_0} \hat{\theta} = \frac{1}{n i_{\theta_0}} + \frac{1}{n^2 i_{\theta_0}^2} \left( \gamma_{\theta_0}^2 + 4 \frac{\Gamma_{\theta_0}^2}{\eta_{\theta_0}} + \Delta_{\theta_0}^2 \right) + 2 \frac{\mathbf{b}_{\theta_0}}{ni_{\theta_0}} + o\left(\frac{1}{n}\right) \quad (10.1)$$

where $\Delta_{\theta_0}^2 \geq 0$ and for the MLE $\hat{\theta}$, $\Delta_{\theta_0} = 0$. The quantity $\Gamma_{\theta_0}^2$ is the ordinary curvature at $\theta = \theta_0$ of the two-dimensional curve $(\theta, \varphi(\theta))$ as defined at (2.1).

Before verifying (10.1) several remarks are pertinent. 1) The term $1/n i_{\theta_0}$ is the Cramer-Rao lower bound for the variance of an unbiased estimator. The bracketed quantity in (10.1) expresses the coefficient of the $1/n^2 i_{\theta_0}^2$ term as the sum of three non-negative quantities: $\gamma_{\theta_0}^2$, the statistical curvature, which is invariant under
transformations of $\theta$, $4r^2_{\theta_0}/i_{\theta_0}$, the "naming curvature", which depends on how $\mathcal{H}$ is parametrized (however, notice that $4r^2_{\theta_0}/i_{\theta_0}$ is invariant under linear reparametrizations $\theta \rightarrow \alpha + \beta \theta$); and $\Delta_{\theta_0}$, which can be made zero by using the MLE. Taken literally (10.1) says that the MLE is superior to other efficient estimations with the same bias structure.

2) The estimators $\hat{\theta}$ will generally be biased by an amount of order $1/n$. This effect means square error to order $1/n^2$. A simple adjustment, noted below at Remark 11, produces estimators biased only to order $1/n^2$; (10.1), with the bias term $2b_{\theta_0}/ni_{\theta_0}$ removed, is valid for such estimators. Among such bias corrected estimators, (10.1) says that the MLE has asymptotically smallest variance.

3) The Fisher information is essentially invariant under reparametrizations of $\mathcal{H}$, in the sense that if $\mu = \mu(\theta)$ is a differentiable monotonic function then $i^T_{\mu} = i^T_{\theta}(d\theta/d\mu)^2$ for every statistic $T(x)$. The squared error estimation problem is not invariant under reparametrization and this accounts for the presence of the $4r^2_{\theta_0}$ term in (10.1). For a given $\theta_0$, the "best" parametrization is in terms of $\phi(\theta)$, the expectation of the best locally unbiased estimator of $\theta$. (Notice that $\phi$ will be the same, except for scale and translation constants, no matter what "$\theta" we begin with.) It will turn out that if the MLE $\hat{\theta}$ is unbiased for $\theta$ then $\phi = \theta$ for all choices of $\theta_0$, so we are automatically using the best parametrization.
4) (10.1) is not a special case of the Bhattacharya lower bounds. The second Bhattacharya bound, applying to estimators biased by amount \( O(1/n^2) \) or less, is of the form

\[
\operatorname{Var}_{\theta_0} \hat{\theta} \geq \frac{1}{ni_{\theta_0}} + \frac{1}{n^2i_{\theta_0}} \left( 4r_{\theta_0}^2 \right) + O\left( \frac{1}{n^3} \right),
\]

and the higher Bhattacharya bounds are identical until order \( O(1/n^3) \), so these bounds relate only to the naming part of the estimation problem. It is possible for an estimator to achieve equality in (10.2), but then it cannot be efficient in a neighborhood of \( \theta_0 \), so (10.1) is not contradicted.

5) Even if \( \mathcal{Y} \) is not a curved exponential family we can use (10.1) to get an improved approximation to \( \operatorname{Var}_{\theta_0} \hat{\theta} \), compared with the Cramer-Rao lower bound \( 1/n_i_{\theta_0} \). The Cauchy translation family discussed at (7.4) has \( i_{\theta_0} = 1/2, r_{\theta_0}^2 = 5/2 \). The MLE \( \hat{\theta} \) is unbiased in this case, so \( r_{\theta_0}^2 = 0 \) and (10.1) is of the form

\[
\operatorname{Var}_{\theta_0} \hat{\theta} = \frac{1}{ni_{\theta_0}} + \frac{r_{\theta_0}^2}{n^2i_{\theta_0}} + O\left( \frac{1}{n^3} \right).
\]

Numerical comparisons of this formula with the Monte Carlo studies of Barnett [2] and also of Andrews et al. [1] are shown in Figure 6. The theoretical values are obviously too small for \( n \leq 11 \), but seem to be more accurate than the Monte Carlo results for \( n \geq 13 \). For \( n = 40 \) Andrews et al. estimate \( \operatorname{Var}_{\theta_0} \hat{\theta} - 1/n_i_{\theta_0} = .0025 \pm .0017 \) while (10.1) gives .0031.
Figure 6. Variance of MLE minus Cramer-Rao lower bound, for estimating the Cauchy translation parameter. Theoretical value from (10.1) compared with Monte Carlo results.

6) For estimating a translation parameter Pitman's estimator is known to have smaller variance than the MLE. However (10.1) suggests that this effect must be of magnitude at most \( O(1/n^3) \).

7) Nothing in (10.1), except the application to general curved exponential families, is new. Rao [12] states the result for curved multinomial families, and notes that for the MLE it was previously derived by Haldane and Smith [9]. The identification of the bracketed terms with curvatures is new, as well as the line of proof which leads to a rigorous verification.

8) The similarity of (9.1) and (10.1) can be viewed as a vindication of the belief that Fisher information is an accurate measure of the information contained in a given statistic. This
conclusion is premature; the squared error estimation problem is very closely related to the information calculation, a fact which would be more obvious if we had presented a geometric argument below, as in Section 9, instead of using analytic methods. It is more reasonable to say that the curvature $\gamma_0$ is the leading term defining the non-linearity of a family $\mathcal{H}$, and must play a central role in all calculations like (9.1) and (10.1). On the other hand in the absence of evidence to the contrary it seems difficult to dispute Fisher and Rao's assertion that the MLE provides the most informative one-dimensional summary statistic even when there is no one-dimensional sufficient statistic.

Our derivation of (10.1) will be done with the curved exponential family $\mathcal{H}$ in standard form, and assuming $\theta_0 = 0$. Neither of these assumptions affects the generality of the result. (The transformation to standard form maps any estimator into an estimator having the same variance, and leaves the quantities $i_{\theta_0}$, $\gamma_{\theta_0}$, and $\Gamma_{\theta_0}$ unchanged.) We assume that the estimator $\widetilde{\theta}(\bar{x})$ has continuous third partial derivatives with respect to the components of $\bar{x}$, so that around $\bar{x} = 0$ it has the Taylor's series expansion

$$\widetilde{\theta}(\bar{x}) = a_0 + a'\bar{x} + (\bar{x}'A\bar{x})/2 + o(\bar{x}^2),$$

(10.3)

where $a_0$ is a scalar, $a$ is a $r \times 1$ vector, and $A$ an $r \times r$ matrix, $r$ being the dimension of the full exponential family containing $\mathcal{H}$.
Here $O(x^3)$ indicates a term that near the origin is bounded in absolute value by some polynomial in the components of $\tilde{x}$ containing only terms of order 3.

Differentiating (10.3) with respect to the components of $\tilde{x}$ gives the gradient vector

$$\nabla \tilde{\theta}(\tilde{x}) = a + A \tilde{x} + O(\tilde{x}^2).$$ (10.4)

In order for $\tilde{\theta}$ to be consistent and efficient, (10.3) must have the special form shown in the lemma:

**Lemma.**

A consistent, efficient estimator $\tilde{\theta}(\tilde{x})$, having continuous third partial derivatives near $\tilde{x} = 0$, has the Taylor series expansion

$$\tilde{\theta}(\tilde{x}) = \frac{\tilde{x}_1}{\sqrt{\lambda_0}} - \frac{\lambda_{11}}{2} \frac{\tilde{x}_1^2}{\sqrt{\lambda_0}} + \frac{\gamma_0}{\sqrt{\lambda_0}} \tilde{x}_1 \tilde{x}_2 + \frac{\tilde{x}_1^{(1)} A(1) \tilde{x}_1^{(1)}}{2} + O(\tilde{x}^3)$$ (10.5)

assuming $\mathcal{H}$ is in standard form at $\theta = 0$. Here $\tilde{x}_1$ indicates the $1^{st}$ component of $\tilde{x}$, $\tilde{x}_1^{(1)} = (\tilde{x}_2^{(1)}, \tilde{x}_3^{(1)}, \ldots, \tilde{x}_p^{(1)})$, and $A(1)$ is the matrix $A$ with its first row and column removed. For the MLE $\hat{\theta}(\hat{x})$, $A(1) = 0$.

The proof of the lemma is based on two simple facts: in order for a continuous estimator $\tilde{\theta}(\tilde{x})$ to be consistent it must have "Fisher consistency", 

46
\[ \tilde{e}(\lambda_\theta) = \theta, \quad (10.6) \]

since \( \bar{x} \xrightarrow{D} \lambda_\theta \) under repeated independent sampling from \( f_\theta \).

Moreover, letting

\[ \nabla_\theta = \left. \nabla \tilde{e}(\bar{x}) \right|_{\bar{x} = \lambda_\theta}, \]

\[
\lim_{n \to \infty} \frac{1}{n} \text{Var}_\theta \tilde{e} \xrightarrow{\text{in}} \frac{(\tilde{\eta}_\theta \nabla_\theta \lambda_\theta)^2}{(\tilde{\eta}_\theta \nabla_\theta \tilde{e}) (\nabla_\theta \tilde{e} \nabla_\theta \tilde{e})} \quad (10.7)
\]

so \( \tilde{\theta} \) will be first order efficient at \( \theta \), \( \lim_{n \to \infty} \frac{1}{n} \text{Var}_\theta \tilde{e} = 1 \), if and only if

\[ \nabla_\theta = \left. \nabla \tilde{e}(\bar{x}) \right|_{\bar{x} = \lambda_\theta} = c_\theta \hat{\eta}_\theta \quad (10.8) \]

for some scalar \( c_\theta \). Taken together (10.6) and (10.8) say that the level surface \( M_\theta \equiv \{ \bar{x} : \tilde{e}(\bar{x}) = 0 \} \) of an efficient consistent estimator \( \tilde{\theta} \) must cross \( \{ \lambda_\theta, \theta \in \Theta \} \) at \( \lambda_\theta \), and at that point must be parallel to the level surface (9.4) of the MLE, as shown in Figure 4. (10.7) merely says that the linear term in the expansion of \( \tilde{e}(\bar{x}) \) about \( \lambda_\theta \),

\[ \tilde{\theta} = \theta + \nabla_\theta' (\bar{x} - \lambda_\theta) + o((\bar{x} - \lambda_\theta)^2), \]

must be proportional to the score statistic \( \tilde{Z}_\theta = n \tilde{\eta}_\theta'(\bar{x} - \lambda_\theta) \) in order to get first order efficiency.

A proof follows from a greatly simplified version of the argument below, but the result is well known and will not be derived here.
The proof of (10.5) is obtained by seeing what form of (10.3) is necessary in order that (10.6) and (10.8) hold for \( \lambda_\theta \) near 0. We will need the Taylor series expansions

\[
\eta_\theta = \sqrt{1_0} \varepsilon_1 + \left[ \frac{\gamma_{11}}{\sqrt{1_0}} \varepsilon_1 + 1 \gamma_0 \varepsilon_2 \right] \theta + o(\theta), \tag{10.9}
\]

\[
\lambda_\theta = \sqrt{1_0} \varepsilon_1 \theta + o(\theta^2)
\]

and a more accurate expansion for the first component of \( \lambda_\theta \),

\[
\varepsilon_1' \lambda_\theta = \sqrt{1_0} \varepsilon_1 \theta + \frac{\mu_{11}}{\sqrt{1_0}} \frac{\theta^2}{2} + o(\theta^2) \tag{10.10}
\]

(10.9) follows from the standard form relationships (4.1)-(4.2). To prove (10.10) notice that \( \varepsilon_1' \lambda_\theta = E_\theta x_1 = \frac{1}{\sqrt{1_0}} E_\theta \hat{k}_0(x) \) (see (3.18)). Formally

\[
E_\theta \hat{k}_0 = \int \frac{f_0'}{f_0}(x) \left[ f_0(x) + \theta f_0(x) + \frac{\theta^2}{2} \tilde{f}_0(x) + o(\theta^2) \right] dm(x)
\]

\[
= 1_0 \theta + \frac{\mu_{11} \theta^2}{2} + o(\theta^2), \tag{10.11}
\]

a result which is easy to verify rigorously in an exponential family.
(10.4), (10.9), and (10.8) combine to give

\[ c_0 = c_0 + \delta_0 \theta + o(\theta) \]

\[ a + A(\sqrt{t_0} \varepsilon_1) + o(\theta^2) \]

\[ = c_0 \sqrt{t_0} \varepsilon_1 + \left( \delta_0 \sqrt{t_0} \varepsilon_1 + \frac{c_0 \nu_{11}(0)}{\sqrt{t_0}} \varepsilon_1 + c_0 \varepsilon_0 \right) \theta + o(\theta) \]

(10.12)

implying

\[ a = c_0 \sqrt{t_0} \varepsilon_1 \]

(10.13)

and

\[ \sqrt{t_0} A\varepsilon_1 = (\delta_0 \sqrt{t_0} + \frac{c_0 \nu_{11}(0)}{\sqrt{t_0}}) \varepsilon_1 + c_0 \varepsilon_0 \varepsilon_2 \]

(10.14)

Notice that (10.14) shows that

\[ A_{31} = A_{41} = \cdots = A_{r1} = 0 \]

(10.15)

(10.9), (10.10), (10.13), (10.6), and (10.3) combine to give

\[ \theta = a_0 + c_0 \sqrt{t_0} \left( \sqrt{t_0} \theta + \frac{\mu_{11} \theta^2}{2} \right) + \frac{1}{2} \frac{A_{11}}{t_0} \theta^2 + o(\theta^2) \]

(10.16)

implying

\[ a_0 = 0 \]

(10.17)

\[ c_0 = 1/t_0, \text{ and } c_0 \mu_{11} + \frac{1}{2} A_{11} = 0. \text{ Therefore} \]

49
\[ a = \frac{1}{\sqrt{1_0}} \xi_1, \quad A_{11} = -\frac{\mu_{11}}{\sqrt{1_0}}, \quad A_{21} = \frac{\gamma_0}{\sqrt{1_0}} , \quad (10.18) \]

de the first of these following from (10.13), the last from (10.14).

Taken together, (10.15), (10.17) and (10.18) are equivalent to (10.5). Finally, for the MLE, \( \hat{\theta}(0, \bar{x}_{(1)})' = 0 \), implying \( \hat{A}_{(1)} = 0 \).

This completes the proof of (10.5).

Two more simple results give (10.1) from (10.5). First of all, the Cramer-Rao lower bound for the variance of a possibly biased estimator \( T(\bar{x}) \) can be rewritten as an equality in the following useful form:

\[ E_0T^2 = \frac{1}{n_0} + E_0 \left( T - \frac{\bar{x}_1}{\sqrt{1_0}} \right)^2 + 2 \frac{b_0}{n_0} . \quad (10.19) \]

((10.19) follows from \( \text{cov}_0(T, \bar{z}_0) = 1 + b_0 \).

Notice that \( \bar{z}_0/1_0 = \bar{x}_1/\sqrt{1_0} \) by (6.3) so this statistic is just the best locally unbiased estimator of \( \theta \), \( U_0 \), introduced at (5.5). For an unbiased estimator, (10.19) says that \( \text{Var}_0 T \) exceeds the Cramer-Rao lower bound by the expected squared error of \( T \) in predicting \( U_0 \).

In a curved exponential family the regularity conditions necessary for (10.19) are satisfied if \( E_0T^2 < \infty \) for \( \theta \) in a neighborhood of 0. The second fact needed is that if \( z \) is standard multivariate normal, \( z \sim \mathcal{N}_r(0, I) \), and \( A \) is an \( r \times r \) symmetric matrix, then \( E(z'Az)/2 = \text{tr} A^2 \) and

\[ \text{Var} \left( \frac{z'Az}{2} \right) = \frac{1}{2} \text{tr} A^2 \quad (10.20) \]
As \( n \to \infty \), \( z_n = \sqrt{n} \bar{x} \to \mathcal{N}(0, \Sigma) \), and because \( f_0 \) is inside an exponential family the moments of \( z_n \) converge to the moments of \( z \sim \mathcal{N}(0, \Sigma) \). Ignoring the \( O(x^3) \) term, an omission justified (under an additional restriction on \( \delta \)) in Remark 12 below, (10.3) and (10.5) give

\[
E_0 \delta = E_0 \frac{x'A\bar{x}}{2} = \frac{1}{n} \frac{\text{tr} A}{2} = \frac{1}{n} \left( -\frac{\mu_{11}}{2i_0} + \frac{\text{tr} A^{(1)}}{2} \right) \tag{10.21}
\]

Moreover (10.5) combines with (10.19) and (10.20) to give

\[
E_0 \delta^2 = \frac{1}{n_1^0} + \frac{1}{n} \left( \frac{\nu_1^2}{i_0} + \frac{\mu_{11}^2}{2i_0^4} + \frac{\text{tr} A^{(1)}_2}{2} + \frac{\text{tr} A_4}{4} \right) + \frac{2b_0}{n_1^0} + o\left( \frac{1}{n^2} \right) \tag{10.22}
\]

Therefore

\[
\text{Var}_0 \delta = \frac{1}{n_1^0} + \frac{1}{n} \left( \frac{\nu_1^2}{i_0} + \frac{\mu_{11}^2}{2i_0^4} + \frac{\text{tr} A^{(1)}_2}{2} \right) + \frac{2b_0}{n_1^0} + o\left( \frac{1}{n^2} \right) \tag{10.23}
\]

Finally, (10.11) gives \( \phi(\theta) = \theta + (\mu_{11}/2i_0) \theta^2 + o(\theta^2) \), where \( \phi(\theta) = E_0 \hat{\theta}_0/\hat{\theta}_0 = E_0 \hat{\theta}_0(x)/i_0 \), and then (2.1) gives the curvature squared of \( (\theta, \phi(\theta)) \) equal to \( \mu_{11}^2/8i_0^2 \) at \( \theta = 0 \). This completes the proof of (10.1). We see that the term \( \Delta_0 \) is

\[
\Delta_0 = i_0 \frac{\text{tr} A^{(1)}_2}{2} \tag{10.24}
\]

and so equals 0 for the MLE.
Several more remarks can now be made about (10.1).

9) The bias of the MLE up to \( O(1/n) \) is, by (10.21), equal to 
\[ \frac{-\mu_{11}}{2(1_n^2 n)}. \]
If \( \hat{\theta} \) is unbiased to \( O(1/n) \), as it is for example in any translation parameter estimation problem involving a symmetric density, then we must have \( \mu_{11} = 0 \). By (10.23) we then have 
\[ \text{Var}_0 \hat{\theta} = \frac{1}{n} \hat{\sigma}_x^2 + \gamma_0^2 / n^2 I_0 + o(1/n^2). \]
The naming curvature term disappears from (10.1) in this case, so \( \theta \) must be equivalent to the best name, \( \phi \), at every point in \( \mathcal{F} \).

10) The expression (10.24) for the excess variance of \( \tilde{\sigma} \) over the MLE also occurs in the theory of Section 9,
\[ \lim_{n \to \infty} I_0 - \tilde{\sigma}_x \sim I_0 \gamma_0^2 + \Delta_0, \quad (10.25) \]
see Rao [12].

11) Let \( \mathbb{A}(\theta_0) \) be the matrix \( \mathbb{A} \) in the Taylor expansion (10.3) when we have put \( \mathcal{M} \) into standard form at \( \theta = \theta_0 \), and define 
\[ B_0^{\tilde{\sigma}} = \text{tr} \mathbb{A}(\theta)/2. \]
Then up to \( O(1/n) \), \( B_0^{\tilde{\sigma}} / n \) is the bias of \( \tilde{\sigma} \) when \( \theta = \theta_0 \). It is easy to show, by calculations similar to those in Remark 12 below, that \( \tilde{\sigma}_n = \tilde{\sigma} - \frac{B_0^{\tilde{\sigma}}}{n} \tilde{\sigma}(\hat{\theta})/n \) has bias of order \( O(1/n^2) \) and variance as given in (10.23) but with the term \( 2b_0 / nI_0 \) removed.
See Rao [12]. For the MLE \( \hat{\theta} \), \( B_0^{\tilde{\sigma}} = -\mu_{11}(\theta)/2I_0^2 \). The estimator
\[ \hat{\theta} - B_0^{\tilde{\sigma}} / n + \frac{B_0^{\tilde{\sigma}}}{\hat{\theta}(\tilde{x})} \tilde{\sigma}(\tilde{x}) / n \]
has variance as given in (10.1) but with the term \( \Delta_0^2 \) removed. The point is that by modifying the MLE we can obtain an estimator with the same bias structure and smaller variance than any other consistent, efficient estimator \( \tilde{\theta} \).
12. We have ignored the $O(x^3)$ term in (10.3) in the derivation of (10.23) and (10.1). To justify this requires the following result:

let $\mathcal{G}_n$ be the cube $\{z: |z_i| \leq n^\alpha, i = 1, 2, \ldots, r\}$, $0 < \alpha < 1/6$, and $I_n(z)$ the indicator function of $\mathcal{G}_n$. Define $z_n = \sqrt{n} \bar{x}$ (so $z_n \to \mathcal{N}_r(Q, I)$) and let $p(z_n)$ be a polynomial of degree $i$ in the coordinates of $z_n$. Then

$$E_0 p(z_n)[1-I_n(z_n)] = O(n^{k\alpha} \exp(-\frac{1}{2n^{2\alpha}})) \quad (10.26)$$

as discussed in the Appendix.

Now write (10.5) as $\bar{\theta} - \bar{x}_1/\sqrt{\lambda_0} = Q + R$ where $Q$ is the quadratic term $\bar{x}' \bar{A} \bar{x}$, $\bar{A}$ having the special form indicated in the Lemma, and $R$ is the remainder term $O(x^3)$. Also define $S(\bar{x}) = Q(\bar{x}) I_n(\sqrt{n} \bar{x})$, $T(\bar{x}) = Q(\bar{x}) [1 - I_n(\sqrt{n} \bar{x})]$, and $V = T + R$ (so $Q = S + T$, $\bar{\theta} - \bar{x}_1/\sqrt{\lambda_0} = S + V$). Notice that

$$|V| = |O(x^3)| < K n^{-3\left(\frac{1}{2} - \alpha\right)} \quad \text{for} \quad \sqrt{n} \bar{x} \in \mathcal{G}_n \quad (10.27)$$

for some positive constant $K$. (We use below the same symbol $K$ to represent any bounding constant.) To the assumptions of the Lemma we now add that $|\bar{\theta} - \bar{x}_1/\sqrt{\lambda_0}|$ is uniformly bounded, giving

$$|V| < K, \quad \sqrt{n} \bar{x} \notin \mathcal{G}_n. \quad (10.28)$$
(With only slightly greater effort below, the boundedness condition can be
relaxed to \(|\tilde{\theta}| \leq K(n^{3/2}||\tilde{x}||)^k\) for \(n^{3/2} \notin \mathcal{O}_n\) for some positive constants
K, k). By (10.26) and (10.27),

\[
E_0|V|^\ell = o(n^{-3(\ell/2) - \alpha}),
\]

for any \(\ell \geq 0\), while

\[
E_0|T|^\ell = o(n^{2\alpha \ell} e^{-n^{\alpha\ell/2}}).
\]

Formulas (10.21) and (10.23) were derived assuming
\(\tilde{\theta} - \tilde{x}_1/\sqrt{t_0} = Q\). But

\[
|E_0Q - E_0S| \leq E_0|T| = o(n^{\alpha} e^{-n^{2\alpha\ell/2}})
\]

and

\[
|E_0(\tilde{\theta} - \tilde{x}_1/\sqrt{t_0}) - E_0S| \leq E_0|V| = O(n^{-3(\ell/2) - \alpha}).
\]

Since \(\alpha < 1/6\) this shows that \(E_0 \tilde{\theta} = E_0Q + o(1/n)\), so (10.21) is
valid. Likewise

\[
|E_0(\tilde{\theta} - \tilde{x}_1/\sqrt{t_0})^2 - E_0\tilde{x}^2| = |E_0(S + V)^2 - E_0(S + T)^2|
\]

\[
= |E_0[2SV + V^2 - T^2]|.
\]

(since \(ST = 0\)), which is \(\leq 2E_0|SV| + E_0|V|^2 + E_0|T|^2\). The last two
terms are \(o(n^{-2})\) by (10.29) and (10.30). Notice that \(SV = O(x^5)\)
and \(SV = 0\) for \(\sqrt{n} \tilde{x} \not\in \mathcal{O}_n\), so

54
\[ |SV| < Kn^{\frac{1}{2} - \alpha} \]  \hspace{1cm} (10.31)

Taking \( \alpha < 1/10 \) makes \( E_0 |SV| = o(n^{-2}) \), completing the proof that (10.23) is valid. We remark that a more careful proof, assuming \( \mathcal{G} \) four times continuously differentiable, allows one to replace \( o(1/n^2) \) by \( O(1/n^3) \) in (10.1).
REFERENCES


Appendix. Complete proofs of the statements made in Sections 9 and 10 require large deviation results of the type discussed in Chernoff [3] and the references therein. Suppose \( x_1, x_2, \ldots, x_n \) are independent, identically distributed random variables such that \( E x_i = 0, \text{Var } x_i = 1 \), and \( \psi(s) = E e^{sx} \) exists for \( |s| < s_0 \), some positive constant. Then \( \psi(s) = 1 + s^2/2 + o(s^2) \) for \( s \) near 0, so

\[
\log \psi(s) = s^2/2 + o(s^2). \tag{A1}
\]

Define \( I_{[y, \infty)}(z) = 1 \) or 0 for \( z \geq y \) or \( z < y \) respectively. Because \( e^{s(x_n - y)/\sqrt{n}} \geq I_{[y, \infty)}(\bar{x}_n) \) for all values of \( \bar{x}_n = \sum_{i=1}^{n} x_i/n \) we have, for \( |s| < s_0 \),

\[
P(\bar{x}_n \geq y) \leq E e^{s(x_n - y)/\sqrt{n}} \leq [\psi(s) e^{s y}]^n. \tag{A2}
\]

Lemma. For \( c_n \) a sequence of numbers going to infinity, \( c_n = o(n^{1/6}) \), and \( \ell \) a non-negative integer,

\[
E(\sqrt{n} \bar{x}_n)^\ell I_{[c_n^{\infty}) (\sqrt{n} \bar{x}_n)) \leq c_n^{-\ell/2} e^{c_n^2/2+o(n(1))}. \tag{A3}
\]

Proof. Let \( \bar{F}_n(y) = P(\bar{x}_n \geq y) \), so \( \bar{F}_n(y) \leq [\psi(s) e^{s y}]^n \) for \( |s| < s_0 \) by (A2). We have

\[
E(\sqrt{n} \bar{x}_n)^\ell I_{[c_n^{\infty}) (\sqrt{n} \bar{x}_n)) = -\sqrt{n} \int_{c_n^{\ell}}^{\infty} x^\ell d\bar{F}_n(x)
\]

and integration by parts gives

Al
- \int_{c_n\sqrt{n}}^{\infty} x^\ell \, dF_n(x) = \left( \frac{c_n}{\sqrt{n}} \right)^\ell \frac{c_n}{\sqrt{n}} \left( \frac{c_n}{\sqrt{n}} \right) + \ell \int_{c_n\sqrt{n}}^{\infty} x^\ell-1 \frac{c_n}{\sqrt{n}} \, dF_n(x) \, dx
\left( \frac{c_n}{\sqrt{n}} \right)^\ell \left[ \psi \left( \frac{c_n}{\sqrt{n}} \right) e^{-s c_n/\sqrt{n}} \right] \, dx + \ell \int_{c_n\sqrt{n}}^{\infty} x^\ell-1 [\psi(s) e^{-s x}] \, dx.

Taking \( s = c_n/\sqrt{n} \) gives

\( E(\sqrt{n} \overline{x_n}^\ell \mathbf{1}_{[c_n, \infty)} (\sqrt{n} \overline{x_n})) \leq \psi \left( \frac{c_n}{\sqrt{n}} \right) e^{-c_n^2 \left[ c_n^\ell - \ell c_n^2 E(c_n + G/c_n)^{\ell-1} \right]} \), (A4)

where \( G \) has density \( e^{-g} \) for \( g \geq 0 \), 0 otherwise. Finally

\[ \psi \left( \frac{c_n}{\sqrt{n}} \right) = e^{\frac{n}{2} \log c_n/\sqrt{n}} \left( c_n^2 / 2 + o(c_n^{3/2} / n^{1/2}) \right) \]

(A5)

by (A1). Combining (A4) and (A5) gives (A3) with

\[ o_n(1) = O(\left[ c_n^{1/6} \right]^3) + \log(1 + \ell c_n^2 E(1 + G/c_n)^{\ell-1}) \), (A6)

where we now use \( c_n = o(n^{1/6}) \), \( c_n \to \infty \).

Now let \( x_1, x_2, \ldots, x_n, \ldots \) be independent identically distributed vectors, dimension \( k \), \( \mathbb{E} x_i = 0 \), \( \text{Cov} x_i = I_k \), such that \( \psi(t) = \mathbb{E} e^{t' x_i} \) exists for \( \|t\| < t_0 \), some positive constant. For any unit vector \( v \) define \( x_i^v = v' x_i \). Then (A3) holds with \( \overline{x_n} \) replaced by \( \overline{x_n}^v \). The term \( o_n(1) \) is defined as in (A6), with the big \( O \) term being the one in the expression \( \log \psi(t) = \|t\|^2/2 + o(t^3) \). (Notice that \( o_n(1) \) does not depend on \( v \).) (10.26) now follows easily.
Lemma. If \(|E_0 e^{it'x}|^p\) is integrable as a function of \(t\) for some \(p \geq 1\) then \(g_n(z)\), the density of \(z = \sqrt{n} \tilde{x}_n\), exists and satisfies
\[
g_n(z) \leq \frac{3^2}{2} \frac{\|z\|}{(2\pi)^{3/2}} e^{-\frac{\|z\|}{4} \min\{c_n, \|z\|\} + o_n(1)}, \tag{A7}
\]
\(c_n = o(n^{1/2}), c_n \to \infty\).

Proof. Consider the univariate case, with \(n\) even. Define
\[
h(z) = \int_{-\infty}^{\infty} g_{n/2}(w) g_{n/2}(z-w) dw = \int_{-\infty}^{z/2} g_{n/2}(w) g_{n/2}(z-w) dw + \int_{z/2}^{\infty} g_{n/2}(w) g_{n/2}(z-w) dw. \tag{A8}
\]
Here \(g_{n/2}(z)\), the density of \((n/2)^{1/2} \tilde{x}_{n/2}\), is known to exist and to converge uniformly to \((2\pi)^{-1/2} \exp(-z^2/2)\), see p. 224 of [8]. Then
\[
M_n = \sup_z |g_n(z)| = (2\pi)^{-1/2} + o_n(1), \quad \text{so for } 0 \leq z \leq c_n
\]
\[
h(z) \leq M_{n/2} \int_{-\infty}^{z/2} g_{n/2}(z-w) dw + \int_{z/2}^{\infty} g_{n/2}(w) dw \\
\leq 2 M_{n/2} e^{-z/2 \min\{2c_n, z\} + o_1(1)}
\]
where we have used the bound \(P[\sqrt{n} \tilde{x}_n \geq z] \leq \exp\left[-\frac{z}{2} \min\{c_n, z\} + o_n(1)\right]\) obtained by setting \(y = z/\sqrt{n}\) and \(s = \min\{z/\sqrt{n}, c_n/\sqrt{n}\}\) in A(2).

But \(g_n(z) = \sqrt{2} h(\sqrt{2} z)\), giving (A7). The same proof with trivial modifications works for \(n\) odd. For the multivariate case the integrals in (A8) are over the regions \(R_1 = \{w: z'w < \|z\|^2/2\}\) and \(R_2 = \{w: z'w > \|z\|^2/2\}\).
Remark 1 of Section 9 follows because (A7) makes step (v) of the heuristic proof valid. All the other approximations involved in the proof are handled by power series expansions and the bounding arguments of Remark 12, Section 10.
ACKNOWLEDGEMENT

Much of this work was done while I was visiting Imperial College, London, Department of Mathematics. I appreciate the assistance of Margaret Ansell in carrying out the more difficult numerical computations. The Associate Editor provided extensive help, especially with the Appendix.