CHAPTER 4

# Curvature and Information Loss

**4.1. Curvature.** This chapter contains a very informal discussion of curvature and connections along the lines of Efron (1975), who introduced curvature, and his discussant, Dawid (1975), who related Efron's curvature to the notion of "connection" in differential geometry. A more rigorous treatment is available in the articles in Amari, Barndorff-Nielsen, Kass, Lauritzen and Rao (1987); see also Kass (1989).

Consider a planar curve

$$y = y(x).$$

Then the curvature at $x$, $\gamma(x)$, is the rate at which the tangent changes direction in a neighborhood of $x$ and is defined as

$$\gamma(x) = \frac{da}{ds} = \lim_{h \to 0} \frac{a(x+h) - a(x)}{s(x+h) - s(x)},$$

where $a(x)$ is the angle that the tangent makes with a fixed line and $s(x)$ is the length along the curve up to $x$ from a fixed point.

Consider now a curved exponential, $\Theta$ an open interval and $\beta(\theta)$ a curve in $R^k$. Given two $K$-dimensional vectors, define the inner product by

$$\langle \beta_1, \beta_2 \rangle = \beta_1 \Sigma \beta_2^T,$$

where $\Sigma = \Sigma_{\theta_0}$ is a positive definite matrix to be specified later. Then,

$$\text{(4.1)} \qquad \cos(\text{angle between } \beta_1, \beta_2) = \langle \beta_1, \beta_2 \rangle / \| \beta_1 \| \cdot \| \beta_2 \|,$$

where $\| \beta \| = (\beta \Sigma \beta^T)^{1/2}$.

If $s_\theta$ is the length of the curve $\beta = \beta(\theta)$ from $\theta_0$ to $\theta$, then

$$\left. \frac{ds(\theta)}{d\theta} \right|_{\theta_0} = \left( \dot{\beta}(\theta_0) \Sigma_{\theta_0} \dot{\beta}(\theta_0) \right)^{1/2},$$

where $\dot{\beta}(\theta)$ is, as before, the derivative of $\beta(\theta)$ with respect to $\theta$. Curvature is defined, as before, by

$$\gamma(\theta)|_{\theta_0} = \left.\frac{da(\theta)}{ds(\theta)}\right|_{\theta_0}, \tag{4.2}$$

where $\cos(a(\theta))$ is given by (4.1) with $\beta_1 = \dot{\beta}(\theta_0)$ and $\beta_2 = \dot{\beta}(\theta_0 + d\theta)$.

Note that

$$\sin da(\theta) \sim da(\theta)$$

and hence, from (4.2),

$$\begin{aligned}\gamma(\theta_0) &= \left.\frac{da(\theta)}{ds(\theta)}\right|_{\theta_0} = \lim_{h\to 0}\frac{\sin a(\theta_0 + h)}{h}\,\frac{h}{s(\theta_0 + h) - s(\theta_0)}\\ &= \frac{\left[M_{11}M_{22} - M_{12}^2\right]^{1/2}}{M_{11}^{3/2}},\end{aligned} \tag{4.3}$$

where $M_{ij} = \beta_i \Sigma_{\theta_0} \beta_j^T$, $i, j = 1, 2$, and $\beta_1 = \dot{\beta}(\theta_0)$, $\beta_2 = \ddot{\beta}(\theta_0)$, $\ddot{\beta}(\theta_0)$ being the second derivative with respect to $\theta$.

So

$$\gamma(\theta_0) = \left\{|M|/M_{11}^3\right\}^{1/2}. \tag{4.4}$$

The choice of $\Sigma_{\theta_0}$ is now specified. Take it as the $(k \times k)$ information matrix of the original exponential (see Section 2.4), that is, of

$$p(x|\beta) = d(\beta)\exp\{\Sigma\beta_i(\theta)f_i(x)\}A(x),$$

and evaluate at $\beta = \beta(\theta_0)$. So

$$\Sigma_{\theta_0} = \left[E_{\theta_0}\frac{\partial \log p(x|\beta)}{\partial\beta_i}\frac{\partial \log p(x|\beta)}{\partial\beta_i}\bigg|_{\beta(\theta_0)}\right]. \tag{4.5}$$

Finally we define curvature for a general family of probability densities satisfying regularity conditions. Consider such a family $p(x|\theta)$, $\theta \in \Theta$, $\Theta$ an open interval. Let

$$\dot{l}(\theta) = \frac{d \log p(x|\theta)}{d\theta}. \tag{4.6}$$

We identify $\dot{l}(\theta)$ as a "tangent" to the "curve" $\theta \to p(\cdot|\theta)$. One defines an (abstract) tangent space at $\theta$ as the linear space of all random variables $Y$ with $E_\theta(Y) = 0$, and chooses $\dot{l}_\theta$ as a distinguished element of this space. The "inner product" of two such tangents $\dot{l}_{\theta_0}$ and $\dot{l}_\theta$ is defined by

$$\langle \dot{l}_{\theta_0}, \dot{l}_\theta\rangle = \mathrm{Cov}\left(\dot{l}_{\theta_0}, \dot{l}_\theta|\theta_0\right). \tag{4.7}$$

Now repeating the calculations made for a curved exponential with $\dot{l}(\theta)$ replacing $\dot{\beta}(\theta)$ and using the new inner product, we get "statistical curvature" at $\theta_0$ is

$$\begin{aligned}\gamma(\theta_0) &\underset{\text{def}}{=} \text{limiting rate at which } \dot{l}(\theta) \text{ changes its angle at } \theta_0 \\ &= \left[|M|/M_{11}^3\right]^{1/2},\end{aligned} \tag{4.8}$$

where $M = [M_{ij}]$ is the dispersion matrix (under $\theta_0$) of $\dot{l}(\theta_0)$ and $\ddot{l}(\theta_0) = (d^2 \log p(x|\theta))/d\theta^2|_{\theta_0}$.

It is easy to check that this is a proper generalization of the notion introduced earlier in the sense that for a curved exponential, statistical curvature equals (geometrical) curvature of the curve of $\theta \to \beta(\theta)$.

Note that if we have to calculate the curvature at a different point $\theta_1$, we must use the corresponding inner product. Inner products change with $\theta$. More about this when we discuss "connections."

**4.2. Geometry of information loss.** By passing to the space of the (minimal) sufficient statistic, we assume, without loss of generality, that the curved exponential is of the form

$$p(x|\beta(\theta)) = p(x|\phi) = c(\theta)\exp\left\{\sum_1^k \beta_i(\theta)x_i\right\} \tag{4.9}$$

(up to a factor involving only $x$).

Fix $\theta_0$. By making one-to-one linear transformations on the $x$'s, $\beta$'s and $\theta$'s, we may assume, without loss of generality,

$$\begin{aligned} &\theta_0 = 0, \qquad \text{dispersion matrix of } X = I \text{ (identity matrix)}, \\ &\mu(\theta_0) = \pi(\theta_0) = E(X|\theta_0) = 0, \\ &\beta(\theta_0) = 0, \\ &\dot{\beta}_i(\theta_0) = 0, \qquad i = 2, 3, \ldots, k. \end{aligned} \tag{4.10}$$

Efron (1975) calls this the standard form.

Let $l_1$ be the unit vector along $\dot{\beta}(\theta_0)$, that is, $(1, 0, 0, \ldots, 0)$ and let $l_2$ be the unit vector which is orthogonal to $l_1$ and lies in the linear space spanned by $\dot{\beta}(\theta_0)$, $\ddot{\beta}(\theta_0)$. Locally everything happens in this space, but to make our life easy, we will take $k = 2$.

Let $L_{\hat{\theta}}$ be the level curve of $\hat{\theta}$ in the space of the (minimal) sufficient statistic based on $n$ observations, that is,

$$\begin{aligned} L_{\hat{\theta}} &= \left\{(\bar{X}_1, \bar{X}_2);\ \hat{\theta}(\bar{X}_1, \bar{X}_x) = \hat{\theta}\right\} \qquad (\hat{\theta} \text{ being the observed value}) \\ &= \left\{(\bar{X}_1, \bar{X}_2);\ \dot{\beta}(\hat{\theta})\left(\bar{X}^T - \mu(\hat{\theta})^T\right) = 0\right\} \end{aligned} \tag{4.11}$$

since the observed value of $\hat{\theta}$ will be close to $\theta_0$, $\mu(\hat{\theta})$ will be close to $\mu(\theta_0)$, which is zero by assumption. Hence $L_{\hat{\theta}}$ may be approximated by a line

$$L_{\hat{\theta}} \simeq \left\{(\bar{X}_1, \bar{X}_2);\ \dot{\beta}(\hat{\theta})\bar{X}^T = 0\right\}. \tag{4.12}$$

Information loss for the mle, as defined by Fisher, is [see (3.2) and (3.3)]

$$nI(\theta_0) - I_{\hat\theta}(\theta_0) = E_{\theta_0}\big\{\mathrm{Var}\big[\dot l(\theta_0)|\hat\theta\big]\big\}, \tag{4.13}$$

where $\dot l$ is now based on $n$ observations,

$$\begin{aligned}\dot l(\theta_0) &= \frac{d\log p(X_1,\dots,X_n|\theta)}{d\theta}\bigg|_{\theta_0} = n\big\{\dot\beta(\theta_0)\big(\overline{X}^T - \mu^T(\theta_0)\big)\big\}\\ &= n\{I(\theta_0)\}^{1/2}\overline{X},\end{aligned} \tag{4.13a}$$

since $\dot\beta_2(\theta_0) = 0$, $\dot\beta_1^2(\theta_0) = I(\theta_0)$ by (4.10). Since $\overline{X} \in L_{\hat\theta}$, $\hat\theta$ is the observed value and $L_{\hat\theta}$ is approximated by the line (4.12),

$$-\overline{X}_1 = \{\tan a(\hat\theta)\}\overline{X}_2 \quad \text{(approximately)}, \tag{4.14}$$

where $a(\hat\theta)$ is the angle between $\dot\beta(\hat\theta)$ and $\dot\beta(\theta_0)$. Since $\hat\theta$ is close to $\theta_0 = 0$, we may think of $\hat\theta$ of $d\theta$. Hence, using the definition of curvature,

$$\tan a(\hat\theta) = \gamma(\theta_0)\hat\theta \quad \text{(approximately)}. \tag{4.15}$$

Hence making use of (4.13a), (4.14) and (4.15),

$$\mathrm{Var}\big(\dot l(\theta_0)|\hat\theta\big) = n^2I^2(\theta_0)\gamma^2(\theta_0)\hat\theta^2\,\mathrm{Var}_{\theta_0}\big(\overline{X}_2|\hat\theta\big). \tag{4.16}$$

Since [see, e.g., (3.9a)]

$$\begin{aligned}\big(\hat\theta - \theta_0\big) &\simeq \dot l(\theta_0)/nI(\theta_0)\\ &= \text{a linear function of } \overline{X}_1\end{aligned} \tag{4.17}$$

and $\overline{X}_1$ and $\overline{X}_2$ are independent under $\theta_0$,

$$\mathrm{Var}_{\theta_0}\big(\overline{X}_2|\hat\theta\big) \simeq \mathrm{Var}_{\theta_0}\big(\overline{X}_2|\overline{X}_1\big) = \mathrm{Var}_{\theta_0}\big(\overline{X}_2\big) = 1/n. \tag{4.18}$$

So information loss is, by (4.13),

$$\begin{aligned}E_{\theta_0}\big(\mathrm{Var}\big(\dot l(\theta_0)|\hat\theta\big)\big) &\simeq E_{\theta_0}\big(nI^2(\theta_0)\gamma^2(\theta_0)\hat\theta^2\big)\\ &\simeq nI^2\gamma^2\frac{1}{nI} = \gamma^2(\theta_0)I(\theta_0).\end{aligned} \tag{4.19}$$

We have used various approximations to deduce the formula

$$\text{(limiting) information loss in using } \hat\theta = \gamma^2(\theta_0)I(\theta_0). \tag{4.20}$$

The relation (4.20) follows immediately from Theorem 3.1 and the definition of $\gamma^2(\theta_0)$, if we adopt the Fisher–Rao modified formulation of the notion of information loss and replace the left-hand side of (4.20) by the measure $E_2$ (with $T_n$ replaced by $\hat\theta$).

Ghosh notes in his discussion in Efron (1975) that the quantity $\gamma^2$ (or rather some multiple of it) appears naturally in inference questions, even if one does not seek a geometric interpretation. Pfanzagl was led to it through a study of the local power of tests. Ghosh and Subramanyam (1974) were led to it through their evaluation of the minimum value of $E_2$ and a study of the condition when the minimum $E_2$ can be zero. They noted that the minimum $E_2$, which equals $E_2$ for $\hat{\theta}$, can be zero for all $\theta$ if and only if the curved exponential is linear exponential.

The curvature $\gamma$ has an interesting invariance property. If one transforms to $w$, where $w$ is a one-to-one twice differentiable function of $\theta$, then with the new parametrization,

$$|M \text{ for } w| = \left(\frac{d\theta}{dw}\right)^6 |M \text{ for } \theta|,$$

$$I \text{ for } w = \left|\frac{d\theta}{dw}\right|^2 I \text{ for } \theta.$$

Hence $\gamma$ is invariant under the reparametrization. In fact if one wants a "normalized" $E_2$ of the form $E_2/\{I(\theta)\}^k$, which is invariant under reparametrization, then one must have $k = 3$. This explains the otherwise mysterious appearance of the power 3 in the denominator of $\gamma$.

Efron's original idea of using $\gamma^2$ as a diagnostic tool for when not to use a locally most powerful test has not been followed up. For use of $\gamma^2$ in Bayesian inference, see Kass (1988) and Kass, Tierney and Kadane (1989).

**4.3. Curvature and connection.** "Connections" were first used in statistics by Chentsov (1972). Revival of interest arose from Dawid's comments on a paper by Efron [Dawid (1975)]. Much of the recent work is due to Amari (1985).

To motivate connections, note that in Efron's work on curvature, as given in Section 4.1, the tangent $T_\theta$ at $\theta$ has to be "displaced" to be brought into same (inner product) space as the tangent $T_{\theta_0}$, so that the angle between them can be measured. More generally, one can have a "connection," as in differential geometry, which tells you how to correspond to the tangent $T_{\theta_0+d\theta}$ an element, say $T'_{\theta_0}$ in the tangent space at $\theta_0$. It is then possible to compute the angle between $T'_{\theta_0}$ and $T_{\theta_0}$ using the inner product of the tangent space at $\theta_0$. Note that both in the context of differential geometry and (statistical) curvature, the tangent spaces at different $\theta$-points have a different geometry (leading, in most cases, to different inner products) and elements from two such spaces do not have a well-defined angle between them without a "connection" which lifts elements from one space to the other in a suitable way.

Efron's algorithm for doing this tacitly defines a connection which has come to be known as the exponential connection. In this framework, curvature vanishes if and only if the family $P(\cdot|\theta)$, $\theta \in \Theta$, is a linear exponential family.

Another connection, of importance in statistics, is the "mixing connection" due to Dawid, under which the curvature is zero if and only if the family $P(\cdot|\theta)$ is a mixture $\theta P_0 + (1 - \theta)P_1$, $0 \le \theta \le 1$.

These two connections happen to be the two most useful members of a family of connections, first introduced by Chentsov; see his book [Chentsov (1972)] or his recent article [Chentsov (1990)] on Kolmogorov in *The Annals of Statistics*. Each kind of curvature measures how far a given family of densities is from a family with a particular structure that is called linear. A good overview is available in the introductory article by Kass (1987).

The best application in statistics still seems to be Efron's. However, Amari has shown the mixture connection also has an interesting role in third order efficiency. For example, the loss of information due to use of a FC, FOE $T_n$ equals $\gamma^2 I + \beta_T^2/2$, where $\beta_T^2$ is the curvature arising from the mixture connection and the suffix $T$ indicates we are looking at the family of densities of $T_n$. See Kass (1987) for more details.

A natural analytical tool arising from all this is the notion of a covariant derivative $D$, which is defined as follows. Let the lifting under a given connection be indicated by the map $M(\cdot)$. Then

$$\lim_{h \to 0} \frac{\left(M(T_{\theta_0 + h}) - T_{\theta_0}\right)}{h} = DT_{\theta_0},$$

where in the above the tangent, elements $M(T_{\theta_0} + h)$ and $T_{\theta_0}$ are best thought of as operators on some suitable function space. For example, we can think of $T_{\theta_0}$ as $d/d\theta_0$ and so $DT_{\theta_0}$ is a second order differential operator. Often connections are defined by specifying $D$. An example of use of this notion, due to Amari and Kumon (1984), is in the treatment of Neyman–Scott problems.

**4.4. Asymptotic ancillaries and conditional loss of information.** Fisher had always advocated inference conditional on a suitable ancillary statistic, which captures some significant aspect of the observed data but whose distribution is free of $\theta$. For example, as first pointed out by Cox (1958), if the sample size $n$ is a random variable, taking, say, two values 2 and 100 with positive probability, it would be absurd not to condition on its observed value. A general ancillary statistic is somewhat like the sample size in Cox's example. A paradigm for statistics, in which inference is conditional on suitably chosen ancillary statistics, is half way between the frequentist paradigm and the Bayesian paradigm, where inference is conditional on the full data. For a discussion, see Cox and Hinkley (1974), Basu (1988) and Lehmann (1986); see also Barndorff-Nielsen (1988).

In the context of information loss, Fisher also advocated conditioning on

$$\ddot{l}(\hat{\theta}) = \left.\frac{d^2 \log p(X_1, \dots, X_n | \theta)}{d\theta^2}\right|_{\hat{\theta}},$$

the negative of which is often called Fisher's observed information. Like the sample size $n$ in Cox's beautiful example, it does seem to measure something like how informative the sample is.

Pierce (1975) was probably the first to suggest that $\ddot{l}(\hat{\theta})$ can be used to construct an asymptotically ancillary statistic. It is unclear if Fisher also had anything like this in mind when he advocated conditioning on $\ddot{l}(\hat{\theta})$. Efron and Hinkley (1978) proposed the use of the asymptotically ancillary statistic

$$\varphi = \frac{1 - \ddot{l}(\hat{\theta})/I(\hat{\theta})}{\gamma(\hat{\theta})}.$$

In this connection Barndorff-Nielsen (1978) has criticized the concept by pointing out that if one takes a linear exponential family, then $\ddot{l}(\hat{\theta})$ will involve the (minimal) sufficient statistic and hence conditioning on that would not make sense (except, of course, to a Bayesian). It is not clear to us whether there are compelling reasons for conditioning on $\varphi$, but Barndorff-Nielsen's comments do not seem to be justified for two reasons. For a linear exponential, $\gamma = 0$, and so $\varphi$ is not well defined. Second, with $\theta$ as the natural parameter,

$$\log p = n \log c(\theta) + \theta \sum_1^n x_i.$$

Then $n^{-1}\ddot{l}(\hat{\theta}) - I(\hat{\theta}) = 0$ identically, and so, trivially, is exactly ancillary.

Let us get back to Fisher's idea of conditioning on $\ddot{l}(\hat{\theta})$. The following statements are no more than a modern version of what Fisher knew. The mle $\hat{\theta}$ is asymptotically sufficient to a first order (in a sense that can be made precise) and this is closely related to its being FOE. However, though TOE, it is not asymptotically sufficient to a third order, this being clear from the limiting positive value of loss of information. If we condition on $\ddot{l}(\hat{\theta})$, it is not hard to see that conditionally the (limiting) loss of information due to use of $\hat{\theta}$ is zero. It is easiest to see this in the framework of Section 4.1. We approximate $\ddot{l}(\hat{\theta})$ by $\ddot{l}(\theta_0)$ and note that $\operatorname{Var}(\dot{l}(\theta_0)|\hat{\theta}, \ddot{l}(\theta_0)) = 0$ exactly. More generally, if we replace the measure of loss of information $E_2$ of (3.5) in Chapter 3 by

$$\begin{aligned} E_2 = \inf_{\lambda} \lim_{n\to\infty} \operatorname{Var} \theta \Bigg\{ & \frac{d \log p}{d\theta} - nI(\hat{\theta} - \theta) - n\lambda_{11}(\hat{\theta} - \theta)^2 \\ & - n\lambda_{22}\left( \frac{1}{n} \frac{d^2 \log p}{d\theta^2}\bigg|_{\hat{\theta}} + I(\theta) \right)^2 \\ & - n\lambda_{12}(\hat{\theta} - \theta)\left( \frac{1}{n} \frac{d^2 \log p}{d\theta^2} + I(\theta) \right) \Bigg\}, \end{aligned}$$

then it is clear from the expansion of $(\hat{\theta} - \theta)$ in (3.9a) that this new $E_2$ equals zero. The point of all this is that, in a sense, $\hat{\theta}$ and $\ddot{l}(\hat{\theta})$ together seem to carry all the information in the data up to third order.