The Annals of Statistics
1991, Vol. 19, No. 3, 1570–1581

A GEOMETRIC APPROACH TO DETECTING INFLUENTIAL CASES[1]

BY PAUL W. VOS

University of Oregon

Amari's dual geometries are used to study measures of influence in exponential family regression. The dual geometries are presented as a natural extension of the Euclidean geometry used for the normal regression model. These geometries are then used to extend Cook's distance to generalized linear models and exponential family regression. Some of these extensions lead to measures already considered while other extensions lead to new measures of influence. The advantages of one of these new measures are discussed.

1. Introduction. In recent years, differential geometry has been playing an increasingly important role in statistics [see, e.g., Amari, Barndorff-Nielsen, Kass, Lauritzen and Rao (1987), McCullagh (1987), Barndorff-Nielsen (1986), Barndorff-Nielsen, Cox and Reid (1986), or Amari (1985)]. Although Amari (1985) has noticed that exponential families enjoy certain global geometric properties, most applications have been in asymptotic theory where local geometric structures, such as curvatures, are used. In this paper the global properties of exponential families are used to study measures of influential cases in exponential family regression and, in particular, in generalized linear models. The local geometric structure of exponential families, although important to some aspects of generalized linear models [Vos (1987)], will be for the purposes of this paper less important than the global properties. The geometric structure that we use is the same as the α-geometry (for $\alpha = \pm 1$) of Amari (1985), but the interpretation of these geometries as an extension of the more familiar Euclidean geometry will be emphasized. The advantage of this interpretation is that it allows us to extend influence measures for normal linear regression to exponential family regression.

2. Geometry of exponential family regression. We shall assume the vector $y = (y^1, \dots, y^n)'$ is a realization of a random vector $Y = (Y^1, \dots, Y^n)'$ having density p from some regular minimal n-dimensional exponential family. For each random variable Y^i we also have a column of covariates $x^i = (x^i_1, \dots, x^i_k)' \in \mathscr{X} \subset \mathbb{R}^k$ and a known function $f\colon \mathscr{X} \times \mathscr{B} \mapsto \mathbb{R}$ such that $E(Y^i) = f(x^i, \beta)$ for an unknown parameter $\beta = (\beta^1, \dots, \beta^m)' \in \mathscr{B} \subset \mathbb{R}^m$. For

Received June 1988; revised July 1990.

[1] This work originated while the author was at the University of Chicago under support from the McCormick Fellowship and NSF Grants MCS-83-01459, DMS-84-12233 and DMS-86-01732. Work at the University of Oregon supported by NSF Grant DMS-88-03584.

AMS 1980 *subject classifications.* Primary 62J02; secondary 62F99.

Key words and phrases. Dual geometries, exponential families, generalized linear models, influence measures, Kullback divergence, parameter invariance.

some σ-finite measure, p may be written as

$$p(y;\theta) = \exp(y'\theta - \psi(\theta)), \tag{2.1}$$

where $\theta \in \Theta \subset \mathbb{R}^n$. In most applications the components of Y are independent and each component belongs to the same exponential family so that $\psi(\theta) = \Sigma_{i=1}^{n}\psi_1(\theta^i)$, where $\psi_1(t + \theta^1) - \psi_1(\theta^1)$ is the cumulant generating function for Y^1. An important example of exponential family regression is the generalized linear model with errors from an exponential family. In this case, $f(x^i, \beta) = L^{-1}(\Sigma_{a=1}^{m}x_a^i\beta^a)$, where L is called a link function. The most common link is the canonical link for which $\theta^i = \Sigma_{a=1}^{m}x_a^i\beta^a$.

Amari (1985) has shown that the exponential family

$$S = \{p\colon p(y;\theta) = \exp(y'\theta - \psi(\theta)) \text{ for some } \theta \in \Theta\}$$

can be given the structure of a smooth manifold. We will also consider a particular subset of S that is defined as follows. Let $\mu(p)$ be the mean parameterization for $p \in S$ and let $\underline{f}(x, \beta) = (f(x^1, \beta), \dots, f(x^n, \beta))'$. We shall assume that $\underline{f}(x, \cdot\,)$ is an imbedding with domain $\mathscr{B}$ so that

$$M = \{p \in S\colon \mu(p) = \underline{f}(x, \beta) \text{ for some } \beta \in \mathscr{B}\}$$

becomes a smooth submanifold of S called a curved exponential family [Amari (1985), page 108] or an (n, m) exponential family [Barndorff-Nielsen (1980).] We let

$$M_{\mathbb{R}} = \{\mu = \mu(p)\colon p \in M\}, \qquad S_{\mathbb{R}} = \{\mu = \mu(p)\colon p \in S\}.$$

Amari (1985) also defines a pair of dual connections that make S flat. It is this dual geometric structure that we shall use to study measures of influence in exponential family regression.

We shall not discuss the details of these dual geometries but we introduce some notation that will be used in the following section. The tangent space of S and M at p will be denoted T_pS and T_pM, respectively. The corresponding tangent bundles are TS and TM. Each parameterization induces a natural basis on the tangent spaces. The natural basis for μ on T_pS is the set of score vectors for μ:

$$U_i(p) = \frac{\partial l(\mu;Y)}{\partial \mu^i}, \qquad i = 1, \dots, n,$$

and the natural basis for θ on T_pS is the set of score vectors for θ:

$$U_i^*(p) = \frac{\partial l(\theta;Y)}{\partial \theta^i}, \qquad i = 1, \dots, n,$$

where $l(\mu;Y)$ and $l(\theta;Y)$ are the log likelihood functions for μ and θ, respectively. The natural basis for β on T_pM is the set of score vectors for β:

$$U_a(p) = \frac{\partial l\big(\underline{f}(x,\beta);Y\big)}{\partial \beta^a}, \qquad a = 1, \dots, m.$$

Throughout we use the convention that the indices i, j, k correspond to quantities defined on S while the indices a, b, c correspond to quantities defined on M. Notice that $U_a(p) \neq U_i(p)$ even if $a = i$. The basis vectors can be extended to vector fields that we denote by U_i, U_i^*, and U_a. The dual connections on S can be defined,easily using the natural basis vector fields for μ and θ. The mean connection on TS is a derivation ∇ that satisfies $\nabla_{U_i} U_j = 0$ for all basis vectors fields U_i, U_j, while its dual, the exponential connection on TS, is a derivation ∇^* satisfying $\nabla^*_{U_i^*} U_j^* = 0$ for all U_i^*, U_j^*.

The following definitions do not appear in Amari (1985) but will be useful to us. Corresponding to each connection and at each point $p \in S$, there is a diffeomorphism between a neighborhood of the origin in T_pS and a neighborhood of p. This diffeomorphism is called the exponential map [Spivak (1979), page 452]. Using the dual bases U_i and U_j^*, we can define the exponential maps for ∇ and ∇^* by

$$\mu\big(\exp_p(\varepsilon U_i)\big) = \mu(p) + \varepsilon \mu_i,$$

$$\theta\big(\exp_p^*(\varepsilon^* U_j^*)\big) = \theta(p) + \varepsilon^* \theta_j,$$

where $\theta(p)$ is the natural parameter for p, μ_i has ith component μ^i and all others 0, θ_j has jth component θ^j and all others 0 and $\varepsilon, \varepsilon^* > 0$ are chosen so that εU_i and $\varepsilon^* U_j^*$ are in the domain of $\exp_p$ and $\exp_p^*$, respectively. Because S is flat in these two connections, the inverses of exp and exp* are defined on all S for each p. That is, for any two points $p, q \in S$, we can define vectors $\mathbf{v}$ and $\mathbf{v}^*$ in T_pS by

$$\mathbf{v} = \mathbf{v}(p,q) = \sum_{i=1}^{n} (\mu^i(q) - \mu^i(p)) U_i(p),$$

$$\mathbf{v}^* = \mathbf{v}^*(p,q) = \sum_{i=1}^{n} (\theta^i(q) - \theta^i(p)) U_i^*(p).$$

To understand how the dual geometries can be applied to exponential family regression, we consider the role of Euclidean geometry in normal regression. For regression with normal errors, the maximum likelihood estimate for β is also the least squares estimate. The expectation parameter space $S_{\mathbb{R}}$ and Θ are both equal to $\mathbb{R}^n$ and $M_{\mathbb{R}}$ is a submanifold of $\mathbb{R}^n$. For linear regression, $M_{\mathbb{R}}$ is a linear submanifold. The data y is represented by the point $\mu = y$ in $S_{\mathbb{R}}$. One important aspect of the Euclidean geometry is that the least squares estimate can be described geometrically: If $\hat{\beta}$ is the least squares estimate for β, then the residual vector $y - \mu(\hat{\beta})$ is orthogonal to the tangent space $T_{\hat{\beta}} M_{\mathbb{R}}$. The dual geometries allow us to extend this characterization to exponential family regression. Assume for the moment that y is the expectation vector for some density $p_y \in S$. Instead of minimizing a squared distance to obtain maximum likelihood estimates, we minimize the Kullback divergence

$$D(p_y, p) = 2E_{p_y}\{l(y;Y) - l(\mu(p);Y)\}, \tag{2.2}$$

where $\mu(p)$ is the expectation parameter for $p \in M$. [Strictly speaking, $D(y, \mu(\hat{\beta}))$ is twice the Kullback information number defined in Kullback

(1968).] Just as for normal regression, the geometry can be used to describe maximum likelihood estimates: If $\hat{\beta}$ is a maximum likelihood estimate for β and $\hat{p} = p(y; \mu(\hat{\beta}))$ is the maximum likelihood density, then the residual vector $\mathbf{v}(\hat{p}, p_y)$ is orthogonal to $T_{\hat{p}}M$. When $y \notin S_{\mathbb{R}}$, the Kullback divergence is not defined but we can still find maximum likelihood estimates and we can still characterize them geometrically. Although there is no point in S that corresponds directly to y, we can relate the data to the tangent bundle TS using the vectors $\mathbf{v}(\cdot, p_y)$ defined at each $p \in S$ by

$$\mathbf{v}(p, p_y) = \sum_{i=1}^{n} (\mu^i(p) - y^i) U_i(p).$$

Even though the Kullback divergence is not defined, the geometric description is now the same as when $y \in S_{\mathbb{R}}$, that is, if $\hat{\beta}$ is a maximum likelihood estimate for β, then the vector $\mathbf{v}(\hat{p}, p_y)$ is orthogonal to $T_{\hat{p}}M$.

3. Measures of influence. In this section we explore how the geometry developed in the previous section can be used in the study of influential cases. By the ith case we mean the ith observation y^i together with its covariate values $x_1^i, \dots, x_k^i$. We call the ith case influential if the inferences or summaries made with all the data are significantly different from those made with all the data except the ith case. For a more complete description of influential cases see Cook and Weisberg (1982). How we measure influence will depend on what aspects of the model are of greatest interest. If there is a parameterization that is of special interest, then we may want to measure the influence of the ith case by calculating the difference between the parameter estimates using all the data and the parameter estimates using all the data except the ith case. In this paper, however, we assume the parameterization is not a part of the model formulation and we shall prefer influence measures that are parameter invariant. Instead of comparing estimated parameters, we compare the probability distributions named by these parameters. We only consider single-case deletions since extensions to multicase deletions are obvious, although the computations grow rapidly.

In order to study measures of influence for exponential family regression, we begin with a special case, normal linear regression, where measures of influence are better understood. An important and widely used measure of influence in normal linear regression is Cook's (1977) distance. If we let $\hat{\mu}, \hat{\beta}, \hat{p}$ represent maximum likelihood estimates using all the data while $\hat{\mu}_{(i)}, \hat{\beta}_{(i)}, \hat{p}_{(i)}$ represent maximum likelihood estimates using all the data except the ith case, then Cook's distance is

$$C_i = \frac{1}{m\sigma^2} \left(\hat{\mu} - \hat{\mu}_{(i)}\right)' \left(\hat{\mu} - \hat{\mu}_{(i)}\right). \tag{3.1}$$

Our goal is to extend C_i to all exponential class regressions in such a manner that the important properties of C_i are maintained. We proceed by giving a geometric interpretation to C_i and then extending the geometry to other error structures. For the normal distribution with σ^2 known, it is easily seen that

$$C_i = D\big(\hat{p}, \hat{p}_{(i)}\big)/m. \tag{3.2}$$

Cook and Weisberg (1982), page 183, define a measure called the likelihood distance that is simply twice the difference of two log likelihoods

$$LD(p_1, p_2) = 2\{l(\mu_1; y) - l(\mu_2; y)\}. \tag{3.3}$$

Since $LD(p_1, p_2)$ is not a distance, Cook (1986a) uses the term likelihood displacement instead; we shall do the same here. For normal linear regression it is easily shown that $LD(\hat{p}, \hat{p}_{(i)}) = D(\hat{p}, \hat{p}_{(i)})$ so that

$$C_i = LD\big(\hat{p}, \hat{p}_{(i)}\big)/m. \tag{3.4}$$

Cook's distance can also be interpreted as the squared length of the vector connecting $\hat{p}$ and $\hat{p}_{(i)}$:

$$C_i = \big\|\mathbf{v}\big(\hat{p}, \hat{p}_{(i)}\big)\big\|^2/m \tag{3.5}$$

since $\mathbf{v}(\hat{p}, \hat{p}_{(i)}) = \sum_{j=1}^{n}(\hat{\mu}^j_{(i)} - \hat{\mu}^j)U_j(\hat{p})$ and the inner product used to define the norm is the Fisher information $\sigma^{-2}I_{n\times n}$.

Equations (3.2), (3.4) and (3.5) show that there are at least three different ways to extend Cook's distance beyond normal linear regression. Finding the estimate $\hat{p}_{(i)}$ will generally require several iterations of an estimation algorithm, so that it is common to replace $\hat{p}_{(i)}$ with the single iteration estimate $\hat{p}_{(i)1}$ [Cook and Weisberg (1982), Pregibon (1981) and Moolgavkar, Lustbader and Venson (1984)]. Two of these measures then become

$$\begin{aligned} D_i &= D\big(\hat{p}, \hat{p}_{(i)1}\big)/m, \\ LD_i &= LD\big(\hat{p}, \hat{p}_{(i)1}\big)/m. \end{aligned} \tag{3.6}$$

The invariance properties of these measures will depend on what algorithm is used to produce $\hat{p}_{(i)1}$. An obvious extension of the interpretation given in (3.5) is to replace $\hat{p}_{(i)}$ with $\hat{p}_{(i)1}$. This extension is rarely considered, probably because the resulting measure is not parameter invariant. An extension that is invariant is defined as follows. Consider the $(n - 1)$-dimensional manifold S_{-1} obtained from S by deleting the ith case. We assume that the submanifold M_{-i} obtained by deleting the ith case is still m-dimensional. Let $\hat{p}_{-i} \in S_{-i}$ be obtained by deleting the ith case from $\hat{p}$ and define

$$\mathbf{v}_{-i} = \sum_{j\neq i}(y^j - \hat{\mu}^j)U_j(\hat{p}) \in T_{\hat{p}_{-i}}S_{-i} \subset T_{\hat{p}}S,$$

$$\mathbf{w}_{-i} = P(\mathbf{v}_{-i}),$$

where $P(\cdot)$ is the orthogonal projection defined in $T_{\hat{p}_{-i}}S_{-i}$ onto $T_{\hat{p}_{-i}}M_{-i}$. An easy calculation shows that

$$\mathbf{w}_{-i} = \sum_{a=1}^{m}\big(\hat{\beta}^a_{(i)1} - \hat{\beta}^a\big)U_a(\hat{p}) = \sum_{a=1}^{m}\sum_{j\neq i}\big(\hat{\beta}^a_{(i)1} - \hat{\beta}^a\big)\frac{\partial f(x^j, \beta)}{\partial \beta^a}U_j(\hat{p}),$$

where $\hat{\beta}_{(i)1}$ is the single iteration estimate obtained from the Fisher scoring algorithm using $\hat{\beta}$ as the starting value. Corresponding to $\mathbf{w}_{-i} \in T_{p_{-i}}M_{-i}$, there is a vector $\mathbf{w}_{(i)} \in T_{\hat{p}}M$ defined by

$$\mathbf{w}_{(i)} = \sum_{a=1}^{m} \sum_{j=1}^{n} \left(\hat{\beta}^{a}_{(i)1} - \hat{\beta}^{a}\right) \frac{\partial f(x^{j}, \beta)}{\partial \beta^{a}} U_{j}(\hat{p}). \tag{3.7}$$

For normal linear regression, Fisher's scoring algorithm converges in one iteration so that $\mathbf{w}_{(i)} = \mathbf{v}(\hat{p}, \hat{p}_{(i)1})$ and the extension to (3.5) becomes

$$W_i = \|\mathbf{w}_{(i)}\|^2/m. \tag{3.8}$$

Notice that each of these measures could be defined using the Kullback divergence or Fisher information for the $(n-1)$-dimensional densities $\hat{p}_{-i}$ and $\hat{p}_{-i1}$ and the vector $\mathbf{w}_{-i}$. Although these measures based on $(n-1)$-dimensional quantities may be reasonable measures of influence, they are not generalizations of Cook's distance and we do not consider them here. The geometry for the $(n-1)$-dimensional measures is similar to that of the n-dimensional measures given in (3.6) and (3.8), although it is not identical. Choosing between measures based on the full likelihood and the likelihood based on $(n-1)$ observations is an extension of the problem in normal linear regression of choosing between Cook's distance and DFFITS [Chatterjee and Hadi (1986)]. The $(n-1)$-dimensional measures of influence and their geometric structure are considered in Vos (1987).

Before comparing the properties of the measures listed in (3.6) and (3.8), we discuss the relationship between the Kullback divergence and the likelihood displacement. Suppose there exists $p_y \in S$ such that $\mu(p_y) = y$. From the definitions given in (2.2) and (3.3) and (2.1), we have

$$LD(p_1, p_2) = D(p_y, p_2) - D(p_y, p_1). \tag{3.9}$$

From (3.9) we see that $LD(p_1, p_2)$ measures how much better p_2 fits the saturated model compared to p_1, while $D(p_1, p_2)$ measures directly how close p_1 and p_2 are. Although the Kullback divergence and the likelihood displacement will often indicate the same cases as influential, these measures can be quite different. If p_y, $\hat{p}$ and $\hat{p}_{(i)}$ are vertices of an "isosocles triangle" with $D(p_y, \hat{p}) \approx D(p_y, \hat{p}_{(i)})$, then LD_i will be near 0 while D_i can be large and LD_i would fail to indicate this as an influential case. The definitions in (2.2) and (3.3) also show that

$$D(p_1, p_2) = E_{p_1}\{LD(p_1, p_2)\}. \tag{3.10}$$

In light of (3.10), we can interpret the Kullback divergence as the expected likelihood displacement. We have already noted that LD_i and D_i are equal for normal linear regression; using the following identity

$$D(p_y, \hat{p}_{(i)1}) = D(p_y, \hat{p}) + D(\hat{p}, \hat{p}_{(i)1}) + 2\langle \mathbf{v}(\hat{p}, p_y), \mathbf{v}^*(\hat{p}, \hat{p}_{(i)1})\rangle, \tag{3.11}$$

it can be shown that the equality of these two measures extends to all generalized linear models that use the canonical link. Equation (3.11) is a generalization of the well-documented Pythagorean relationship that holds in exponential families [Amari (1985), page 92, and Hastie (1987)] and is verified by making substitutions from (2.2) and the definitions of $\mathbf{v}(\cdot,\cdot)$ and $\mathbf{v}^*(\cdot,\cdot)$. For the canonical link the third term of (3.11) is 0 because $\mathbf{v}^*(\hat{p}, \hat{p}_{(i)1}) \in T_{\hat{p}}M$ and $\mathbf{v}(\hat{p}, p_y)$ is orthogonal to $T_{\hat{p}}M$. Hence (3.9) and (3.11) show that $LD_i = D_i$ whenever the canonical link is used.

In normal nonlinear regression one of the most common extensions of Cook's distance is the measure W_i. This measure is often defined as

$$W_i = \left(\hat{\beta} - \hat{\beta}_{(i)1}\right)' F'F\left(\hat{\beta} - \hat{\beta}_{(i)1}\right)/m\sigma^2, \tag{3.12}$$

where F is the $n \times m$ matrix with elements $\partial f(x^i, \beta)/\partial\beta^a$, $i = 1,\dots,n$, $a = 1,\dots,m$. From (3.12) it is not clear that this measure of influence is parameter invariant. However, from (3.7) we can see that this definition is equivalent to the definition of W_i given in (3.8) and so must be parameter invariant.

For generalized linear models, Pregibon (1981) suggests a related measure $W_i^N = \|\mathbf{w}^N\|_N^2$, where $\mathbf{w}^N$ is the vector obtained from the first iteration of the Newton–Raphson algorithm and $\|\cdot\|_N^2$ is the squared norm defined by the matrix with components $\partial^2 l/\partial\beta^i\beta^j$. For the canonical link, the Newton–Raphson algorithm and Fisher's scoring algorithm are identical so that $W_i = W_i^N$. For other links, W_i should be preferred because it is invariant under reparameterization while W_i^N is not. McCullagh and Nelder (1983) generalize Cook's distance as

$$\left(\hat{\beta} - \hat{\beta}_{(i)}\right)' F'V^{-1}F\left(\hat{\beta} - \hat{\beta}_{(i)}\right)/m\sigma^2, \tag{3.13}$$

where $V^{-1} = I(\hat{\mu})$ is the inverse variance matrix for Y at $\hat{\beta}$. Since $F'V^{-1}F$ is the matrix for the inner product relative to the basis $\{U_a\}$, we see that by replacing $\hat{\beta}_{(i)}$ with $\hat{\beta}_{(i)1}$, display (3.13) equals W_i. Moolgavkar, Lustbader and Venson (1984) also suggest the measure W_i for generalized linear models. By using the fully iterated value $\hat{\beta}_{(i)}$ in (3.13) we are left with a measure that is not parameter invariant. If one can do all the computations to obtain $\hat{\beta}_{(i)}$, then the fully iterated likelihood displacement $LD(\hat{p}, \hat{p}_{(i)})/m$ or divergence $D(\hat{p}, \hat{p}_{(i)})/m$ are preferred because they are parameter invariant.

In the case of nonnormal error structure, there is reason to prefer measures of influence based on the likelihood displacement to those based on squared norms, such as W_i and C_i defined in (3.5). The measure W_i is a function of only the first two moments of $\hat{p}$ and $\hat{p}_{(i)1}$ while LD_i is defined on the densities themselves. When the skewness and higher-order cumulants are large, W_i and LD_i can be quite different. For this reason, Cook and Weisberg (1982) suggest using the measure LD_i rather than W_i. Although there is no guarantee that either of these measures of influence will always behave correctly, a simple example will illustrate one of the problems with measures based on the squared norm that is avoided by using LD_i. Suppose $y = (2, 1/2)'$ are independent observations from a gamma family with dispersion parameter $\phi = 10$ so that the log likelihood is

$$l(\theta) = \tfrac{1}{10}(y^1\theta^1 + y^2\theta^2 + \log(-\theta^1) + \log(-\theta^2)) + h(y),$$

where $h(y)$ is not a function of θ. Suppose that $-\theta^1 = x^1\beta + c$ and $-\theta^2 = x^2\beta + c$, where $x = (1, 2)'$ and $c = 1$ is known as an offset. The maximum likelihood estimates for β and μ using both data values are $\hat{\beta} = 0$ and $\hat{\mu} = (1, 1)$. The estimate for $\hat{\beta}_{(1)}$ can be found without the need for iteration. Taking $\hat{\beta}_{(1)1} = \hat{\beta}_{(1)} = 1/2$ in (3.7) gives $W_1 = 0.125$. The corresponding fitted values are $\hat{\mu}_{(1)} = (2/3, 1/2)'$. The maximum likelihood estimate $\hat{\beta}_{(2)} = -1/2$ so that $W_2 = W_1 = 0.125$. The measure W_i shows that cases 1 and 2 have the same influence, even though $\hat{\mu}_{(2)} = (2, \infty)'$ and $\hat{\beta}_{(2)} \notin \mathscr{B}$. The one-step likelihood displacement clearly shows that the second case is influential; $LD_1 = 0.08$ while $LD_2 = \infty$. The difficulty illustrated by this example is not that $\hat{\beta}$ need not lie in $\mathscr{B}$. Rather, the influence of the ith case depends not only on how much the estimate for β changes, but also on the direction of this change. The measure W_i is insensitive to the direction and can thereby miss influential observations. Notice in most applications the dimension of β is greater than 1 and so the direction of the change in β involves more than just a change of sign.

Now we consider $LD_i = LD(\hat{p}, \hat{p}_{(i)1})/m$ and $D_i = D(\hat{p}, \hat{p}_{(i)1})/m$ more closely. Since the single iteration estimate $\hat{p}_{(i)1}$ is generally not parameter invariant, neither will these measures be invariant. Recall that simply replacing $\hat{p}_{(i)}$ with $\hat{p}_{(i)1}$ in (3.5) also resulted in a measure dependent on the parameterization. To obtain an invariant measure, we considered the vector $\mathbf{w}_{(i)}$ and defined W_i in terms of this vector. If $\hat{t}_{(i)} = \exp^*_{\hat{p}}(\mathbf{w}_{(i)}) \in S$, then we can define W_i be replacing $\hat{p}_{(i)}$ with $t_{(i)}$ in (3.5)

$$W_i = \left\| \mathbf{v}(\hat{p}, \hat{t}_{(i)}) \right\|^2.$$

Usually $\exp^*_{\hat{p}}(\mathbf{w}_{(i)}) \in S$, but when it is not we can still define $\hat{t}_{(i)} = \exp^*_{\hat{p}}(\varepsilon \mathbf{w}_{(i)}) \in S$ for an appropriately chosen $\varepsilon > 0$. One way to choose $\varepsilon < 1$, is to find the largest value for ε for which $\hat{t}_{(i)}$ lies in S. Since $\mathbf{w}_{(i)}$ and $\exp^*$ are parameter invariant, so is $\hat{t}_{(i)}$. Hence, to obtain parameter measures based on the likelihood, we can replace $\hat{p}_{(i)}$ in (3.2) and (3.4) with $\hat{t}_{(i)}$ to obtain

$$ILD_i = LD(\hat{p}, \hat{t}_{(i)}), \qquad ID_i = D(\hat{p}, \hat{t}_{(i)}).$$

From the definition of $\hat{t}_{(i)1}$ and $\exp^*$, we find that $\hat{t}_{(i)1} = \hat{p}_{(i)1}$ for a generalized linear model having the canonical link and parameterization β satisfying $\theta^i = \sum_{a=1}^m x_a^i \beta^a$. In this special case, $LD_i = ILD_i$ and the one-step likelihood displacement corresponds to a parameter-invariant measure.

From the definition of ILD_i and ID_i and (3.9) and (3.11), we see that

$$ILD_i = ID_i + 2\langle \mathbf{v}(\hat{p}, p_y), \mathbf{v}^*(\hat{p}, \hat{t}_{(i)}) \rangle. \tag{3.14}$$

In the dual geometries, the residual vector $\mathbf{v}(\hat{p}, p_y)$ is orthogonal to $T_{\hat{p}}M$. Since $\mathbf{v}^*(\hat{p}, \hat{t}_{(i)}) = \mathbf{w}_{(i)} \in T_{\hat{p}}M$, the inner product in (3.14) is 0 and $ILD_i = ID_i$. Clearly, ILD_i is parameter invariant and since it is defined using the one-step likelihood displacement, it incorporates the skewness and higher-order cumulants. For the example with $y = (2, 1/2)'$, ILD_i also shows the second case is influential. Since $LD(p_1, p_2) = \|\mu(p_1) - \mu(p_2)\|^2/\sigma^2$ for normal errors, we have $ILD_i = W_i$ for normal regression and $ILD_i = C_i$ for normal linear regression.

For normal regression it is usually necessary to find an estimate $\widetilde{\sigma^2}$ for the variance parameter σ^2. Measures of influence are now defined with $\widetilde{\sigma^2}$ replacing σ^2. For some generalized linear models there is a dispersion parameter ϕ that also needs to be estimated. When there is a dispersion parameter, the density p_1 can be written as

$$\exp\{(y'\theta - \psi(\theta))/a(\phi)\}, \tag{3.15}$$

where $a(\cdot)$: $\mathbb{R} \mapsto (0, \infty)$. Notice that when ϕ is known, (3.15) is a density from an exponential family. If $\tilde{\phi}$ is an estimate for ϕ, then the measures of influence such as W_i, LD_i and ILD_i are redefined using $\tilde{\phi}$ in place of ϕ. Although there are many ways to define $\tilde{\phi}$, most of these estimates are based on either all the data y or all the data less the ith case, $y_{(i)}$. One advantage of using an estimate $\tilde{\phi}$ based on all the data is that $\tilde{\phi}$ will be parameter invariant. Not all the single iteration estimates $\tilde{\phi}_{(i)}$ based on $y_{(i)}$ will have this property. If one is interested in the influence of the ith case on both β and ϕ, then Cook (1986b) suggests the likelihood displacement

$$2\left\{l(\hat{\beta}, \hat{\phi}; y) - l\left(\hat{\beta}_{(i)}, \hat{\phi}_{(i)}; y\right)\right\},$$

where $\hat{\phi}_{(i)}$ is estimated using one iteration or until convergence of the algorithm. If a single iteration estimate is used, then it is desirable to base this estimate on $\hat{t}_1 = \exp^*(\mathbf{w})$ [or the corresponding $(n-1)$-dimensional density] in order to make the measure parameter invariant. Further details on choosing an estimate for ϕ, measures of influence based on $y_{(i)}$ and how the geometry changes for this situation can be found in Vos (1987). A geometric approach to influence measures that does not use the dual geometries can be found in Ross (1987).

The computations involved in the measures of influence ILD_i, LD_i, D_i and W_i are all about the same. The measure W_i is somewhat easier to calculate since it is based on the m-dimensional quantity $\mathbf{w}_{(i)}$. This difference is small, however, compared to the extra calculations required to find fully iterated parameter estimates for the deletion of each case. When $\exp_{\hat{p}}(\mathbf{w}_{(i)}) \notin S$, little computation should be spent to find the largest ε for which $\exp_{\hat{p}}(\varepsilon \mathbf{w}_{(i)})$ lies in S. It seems reasonable to investigate more closely any case that causes the single iteration estimates to leave the parameter space; so if $\exp_{\hat{p}}(\mathbf{w}_{(i)}) \notin S$, the ith case is tagged as influential. For general families of distributions the measure D_i would require heavy computation since the Kullback divergence can be difficult to calculate; for exponential families, however, D_i is easily calculated.

We shall consider an example given in Cook and Weisberg (1982), pages 178 and 185, to compare the LD_i and the ILD_i in practice. The data come from 17

TABLE 1
Leukemia data

Case	WBC	y	ILD_i	LD_i	FLD_i
1	2,300	65	0.03	0.03	0.03
2	750	156	0.09	0.10	0.11
3	4,300	100	0.03	0.03	0.03
4	2,600	134	0.06	0.07	0.07
5	6,000	16	0.02	0.02	0.02
6	10,500	108	0.07	0.07	0.08
7	10,000	121	0.10	0.10	0.11
8	17,000	4	0.08	0.08	0.07
9	5,400	39	0.01	0.01	0.01
10	7,000	143	0.13	0.14	0.16
11	9,400	56	0.00	0.00	0.00
12	32,000	26	0.03	0.03	0.03
13	35,000	22	0.05	0.05	0.04
14	100,000	1	0.57	0.43	0.35
15	100,000	1	0.57	0.43	0.35
16	52,000	5	0.24	0.20	0.18
17	100,000	65	1.40	1.63	9.89

patients with leukemia, a cancer characterized by a high white blood cell (WBC) count. The initial white blood cell count is used as an explanatory variable for y^i, the survival time in weeks.

Cook and Weisberg (1982) suggest the following model

$$Y^i = \xi^1 \exp(\xi^2 x^i)\varepsilon^i, \tag{3.16}$$

where ε^i is standard exponential and $x^i = w^i - \bar{w}$ with w^i being the logarithm (base 10) of the white blood cell count. In this example the ith case is (y^i, w^i), so that deletion of the ith case not only changes the estimate for β but x^j, $j \neq i$, as well. By reparameterizing, (3.16) is seen to be equivalent to the generalized linear model

$$\begin{aligned} Y^i &\sim \text{Exponential}, \\ E(Y^i) &= \exp(\beta^1 + \beta^2 x^i), \end{aligned} \tag{3.17}$$

where $\beta^i = \log(\xi^1)$ and $\beta^2 = \xi^2$. For the exponential (gamma) distribution the negative reciprocal link is canonical, so that the generalized linear model specified by (3.17) does not use the canonical link. Table 1 lists the data and three measures of influence. The measure FLD_i is the likelihood displacement based on the fully iterated estimate, that is, $FLD_i = LD(\hat{p}, \hat{p}_{(i)})$. For this example, the likelihood displacement is

$$LD(p_1, p_2) = 2\sum_{j=1}^{n} \left\{ y^j \left(\frac{1}{\mu_2^j} - \frac{1}{\mu_1^j} \right) + \log\left(\frac{\mu_2^j}{\mu_1^j} \right) \right\}, \tag{3.18}$$

where $\mu_1 = \mu(p_1)$ and $\mu_2 = \mu(p_2)$. If $\hat{\beta}_{(i)1}$ is the single iteration estimate obtained for β using $\hat{\beta}$ as the starting value in Fisher's scoring algorithm, then

$$\mu^j(\hat{p}_{(i)1}) = \exp\left(\hat{\beta}^1_{(i)1} + \hat{\beta}^2_{(i)1} x^j\right),$$

$$\mu^j(\hat{t}_{(i)1}) = \left(-\hat{\theta}^j + \hat{\theta}^j\left(\hat{\beta}^1_{(i)1} - \hat{\beta}^1\right) + \hat{\theta}^j x^j\left(\hat{\beta}^2_{(i)1} - \hat{\beta}^2\right)\right)^{-1},$$

where $\hat{\theta}^j = -\exp(-\hat{\beta}^1 - \hat{\beta}^2 x^j)$. Replacing μ_2 with $\mu(\hat{p}_{(i)1})$ and $\mu(\hat{t}_{(i)1})$ and replacing μ_1 with $\mu(\hat{p})$ in (3.18) gives LD_i and ILD_i, respectively. Since the link is not canonical $LD_i \neq ILD_i$. The measures LD_{17} and ILD_{17} are quite different from FLD_{17}, but all three measures indicate that the seventeenth case is influential. Since LD_i is not parameter invariant we can make this measure arbitrarily large or small simply by choosing a different parameterization. The one-step likelihood displacement is also sensitive to what estimation algorithm is used to find $\hat{\beta}_{(i)1}$. Using the Newton–Raphson algorithm, we find $LD_{17} = 13.87$; but, if we use Fisher's scoring algorithm, $LD_{17} = 1.63$. Using the ξ parameterization, $LD_{17} = 18.78$ for the Newton–Raphson algorithm and $LD_{17} = 1.66$ for Fisher's scoring algorithm.

Cook and Weisberg (1982) calculate the maximum likelihood residual and Studentized residual for the seventeenth case. These are 3.47 and 4.18, respectively. Neither of these is large enough to clearly indicate that the seventeenth case is an outlier. From Table 1 we see that the seventeenth patient had a high white blood cell count and yet survived a long time. Cook and Weisberg (1982) point out that measurement error in the white blood cell count may contribute to the large influence of case 17. For whatever reasons this case is influential, the conclusions from these data should be viewed with caution.

4. Conclusion. We have shown that the dual geometries can be used to compare measures of influence suggested in the literature and to see how these measures are related to Cook's distance. In particular, there is reason to prefer the extension of Cook's distance, W_i, in normal nonlinear regression while the one-step likelihood displacement LD_i is preferred for generalized linear models when the canonical link is used. The geometry is also used to define a new measure, ILD_i, that extends the advantages of both W_i and LD_i to generalized linear models with other link functions and to exponential family regression in general. One important aspect of the dual geometries is that they allow us to study exponential family regression from a parameter-invariant perspective.

In a broader sense, the result on influence measures can be viewed as an example of the role of the dual geometries in exponential family regression. Euclidean geometry has been used to study many aspects of the normal regression model and, although results obtained geometrically can also be found by other methods, the geometric approach often brings more clarity and intuition to the solution. In exponential family regression, the dual geometries play a similar role. The measures W_i and LD_i can each be studied without the dual geometries, but the invariance of LD_i under the canonical link was not realized using these other approaches. Furthermore, many aspects of normal regression can be extended by the dual geometries to exponential family regression because the dual geometries contain the normal regression geometry as a special case. These extensions can lead to new results (ILD_i) as well as a better understanding of existing results (W_i and LD_i). Since the development of normal theory regression is extensive when compared to what is known for generalized linear models, the possibility of geometrically extending normal regression ideas is particularly important.

Acknowledgment. I wish to thank the referees for their helpful and constructive comments.

REFERENCES

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

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

BARNDORFF-NIELSEN, O. E. (1980). Conditionality resolutions. *Biometrika* **67** 293–310.

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

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

CHATTERJEE, S. and HADI, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression (with discussion). *Statist. Sci.* **1** 379–416.

COOK, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19** 15–18.

COOK, R. D. (1986a). Assessment of local influence (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 133–169.

COOK, R. D. (1986b). Comment on "Influential observations, high leverage points, and outliers in linear regression" by S. Chatterjee and A. S. Hadi. *Statist. Sci.* **1** 393–397.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

HASTIE, T. (1987). A closer look at the deviance. *Amer. Statist.* **41** 16–20.

KULLBACK, S. (1968). *Information Theory and Statistics*. Dover, New York.

McCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, New York.

McCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, New York.

MOOLGAVKAR, S., LUSTBADER, E. and VENSON, D. (1984). A geometric approach to nonlinear regression diagnostics with applications to matched case-control studies. *Ann. Statist.* **12** 816–826.

PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9** 705–724.

ROSS, W. H. (1987). The geometry of case deletion and the assessment of influence in nonlinear regression. *Canad. J. Statist.* **15** 91–103.

SPIVAK, M. (1979). *A Comprehensive Introduction to Differential Geometry* **1**. Publish or Perish, Wilmington, Del.

VOS, P. W. (1987). Dual geometries and their applications to generalized linear models. Ph.D. dissertation, Univ. Chicago.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OREGON
EUGENE, OREGON 97403