The Annals of Statistics
1994, Vol. 22, No. 2, 668–695

SOME DIAGNOSTICS OF MAXIMUM LIKELIHOOD AND POSTERIOR NONNORMALITY[1]

BY ROBERT E. KASS AND ELIZABETH H. SLATE

Carnegie Mellon University

Standard large-sample maximum likelihood and Bayesian inference, based on limiting multivariate normal distributions, may be dubious when applied with small or moderate sample sizes. We define and discuss several measures of nonnormality of MLE and posterior distributions that may be used as diagnostics and can indicate whether reparameterization will be effective in improving inferences. We begin by showing how the nonlinearity measures introduced by Beale and Bates and Watts for nonlinear regression may be generalized to exponential family nonlinear models. We replace the exponential family regression surface with another surface defined in terms of the parameterization in which the third derivatives of the loglikelihood function vanish at the MLE, and then we compute "curvatures" of the latter surface. This generalization effectively replaces the normal-theory Euclidean geometry with an α-connection geometry of Amari identified by Kass, yet it may be understood and implemented without reference to that foundational argument.

We also discuss alternative diagnostics based on the observed third derivatives of the loglikelihood function, or the third derivatives of a log posterior density. These may be viewed as multiparameter generalizations of a nonnormality measure proposed by Sprott. We show how one of these diagnostics may be quickly and easily computed using approximations of Tierney, Kass and Kadane.

1. Introduction. In both maximum likelihood and Bayesian inference it is desirable to have sufficiently large samples that inferences may be based on limiting normal distributions. For any given data set, however, a data analyst needs some guidance as to whether or not the normal approximation is adequate. In this paper we define diagnostics that can be used to assess joint normality of an MLE or posterior distribution. We begin with the notion that in well-behaved problems reparameterization can bring the distribution to an approximately normal form, and we base the diagnostics on measures of departure from "optimal" parameterizations, which are defined in several ways.

Let $\ell(\theta)$ and $\tilde{\ell}(\theta)$ denote the loglikelihood and log posterior density functions; let $\hat{\theta}$ and $\tilde{\theta}$ denote the MLE and posterior mode; and let $\hat{\Sigma} = [-D^2\ell(\hat{\theta})]^{-1}$ and $\tilde{\Sigma} = [-D^2\tilde{\ell}(\tilde{\theta})]^{-1}$ be the inverse negative Hessian matrices of ℓ and $\tilde{\ell}$ at $\hat{\theta}$ and

Received September 1990; revised June 1993.

[1]This material was presented by Robert E. Kass as an IMS Special Invited Paper in March 1992. The research was supported by NSF Grants DMS-87-05646, DMS-88-0576 and DMS-90-05858, and NIH Grant RO1-CA54852-01.

AMS 1991 *subject classifications.* Primary 62F12; secondary 62F15, 62-07.

Key words and phrases. Alpha-connections, curvature measures, curved exponential family, differential geometry, generalized linear models, information metric, nonlinear regression, reparameterization.

$\tilde{\theta}$. (Here D^2 denotes the second-derivative operator with respect to the vector θ.) Assume Θ is m-dimensional. Then the usual large-sample maximum likelihood inferences are based on the asymptotic standard m-variate normality of $\hat{\Sigma}^{-1/2}(\hat{\theta} - \theta)$ or on its expected-information counterpart (in which the expected information evaluated at the MLE replaces the observed information $\hat{\Sigma}^{-1}$). Similarly, large-sample Bayesian inferences are based on the asymptotic standard m-variate normality of $\tilde{\Sigma}^{-1/2}(\theta - \tilde{\theta})$. In the case of maximum likelihood we will emphasize the "quadratic loglikelihood" parameterization, so called because it produces an approximately quadratic loglikelihood in the sense that the expectations of the third derivatives of the loglikelihood vanish. In the Bayesian case, we will say that a parameterization produces an approximately quadratic log posterior if the "observed" third derivatives of the log posterior, evaluated at the mode, vanish. Since these derivatives become the observed third derivatives of the loglikelihood evaluated at the MLE when the prior is uniform, the latter may be of use in non-Bayesian inference as well.

Our approach builds on methods that have been used to diagnose nonnormality of least-squares estimators in nonlinear regression. In that context, approximate inference is based on linear approximation to the regression surface at the least-squares fitted value, and the standard procedures perform poorly when the surface is highly nonlinear. To measure nonlinearity, Beale (1960) and Bates and Watts (1980) proposed summaries of second derivatives of the regression surface, which have the form of curvature measures. This general approach has received a fair amount of attention and has proved quite useful in practice [e.g. Ratkowsky (1983) and Seber and Wild (1989)]. Despite this success, however, there has been little attempt to generalize beyond nonlinear regression, one noteworthy exception being the paper by Cook and Tsai (1990). We describe here, in Section 2, what we consider to be a natural and direct generalization of the Bates and Watts (1980) methodology. Technically, it relies on the mathematical foundation of the α-connection geometries introduced by Amari (1982), exploiting an observation made by Kass (1984). From a practical point of view, however, it is easy to understand and compute the nonnormality diagnostics we describe without knowledge of the full-fledged foundation.

The basic idea is to define an appropriate surface, analogous to a nonlinear regression surface, the curvature of which will yield information relevant to the adequacy of inferences based on the normal approximation to the distribution of the MLE. The approach is easiest to understand in the context of a simple example, which we will analyze in detail in Section 4.

EXAMPLE 1. We consider a model (and in Section 4 a data set), taken from Feigl and Zelen (1965) and discussed by many authors including, as an accessible reference, Cook and Weisberg (1982). In this example, the outcome (Y) is survival time in weeks among leukemia patients, and the predictor is white blood cell count (WBC), with $x = \log(\text{WBC}) - \text{mean}(\log(\text{WBC}))$. The model is

$$Y_i \sim \text{Exponential}(1/\mu_i),$$
$$\mu_i = \theta_1 \exp(-\theta_2 x_i),$$

independently, for $i = 1, \ldots, n$, where $E(Y_i) = \mu_i$. We write $\mu = (\mu_1, \ldots, \mu_n)$, noting that $\mu = \mu(\theta)$ and, since $m = 2$ in this case, as θ ranges over the parameter space Θ, $\mu(\theta)$ traces a two-dimensional surface in R^n. This is an example of an exponential family nonlinear model, a class we focus on in Section 2. When we return to this example in Section 4 we will, like other authors, confine our attention to the 17 AG-positive patients. Doing so, we have a sufficiently small data set that the loglikelihood function on θ may not be approximately quadratic, and large-sample inferences may not be reliable. Indeed, the likelihood contours for (θ_1, θ_2) shown in our Figure 3 are somewhat nonelliptical, whereas likelihood contours for an alternative parameterization (λ_1, λ_2) (defined in Section 4) are improved. The measures we study will help determine whether the loglikelihood for a particular parameterization is roughly quadratic.

To assess nonlinearity in this example, one might, at first glance, be tempted to imitate the computation of curvatures of the response surface in nonlinear regression by computing curvatures of the surface $\mu(\theta)$. The obvious difficulty, however, is that even for the linear model $\mu_i = \theta_1 + \theta_2 x_i$ the MLE of θ will not be approximately normal except when the sample size n is quite large. Thus, this particular kind of nonlinearity is of limited relevance. Instead, we measure nonlinearity of a different surface.

We begin by considering the individual exponential observations and introduce a transformation that makes this problem more like the normal nonlinear regression problem. For a single exponential distribution with mean parameter μ, the loglikelihood function on the transformed parameter $\tau = \mu^{-1/3}$ satisfies $E(\ell'''(\tau)) = 0$ and $\ell'''(\hat{\tau}) = 0$. The likelihood function based on an i.i.d. sample of exponential observations might therefore be expected to be more nearly normal as a function of τ than as a function of μ, at least in the sense that its logarithm should on average be more nearly quadratic in τ than in μ. [Beyond its intuitive appeal, this "quadratic-loglikelihood" parameterization has the property that the skewness of the usual pivot is reduced to order $O(n^{-3/2})$; see DiCiccio (1984).] As Slate (1992) shows in the context of natural exponential families having quadratic variance functions (NEF-QVF's), this transformation is remarkably effective in bringing the likelihood function to a more nearly normal form. In the case of the exponential distribution, the normal approximation may be judged adequate even for sample sizes as small as 1 or 2: as shown in Figure 1, for a sample size of $n = 2$, the likelihood function on μ is skewed and the normal approximation deviates substantially from it, while the exact and approximate likelihoods on $\tau = \mu^{-1/3}$ are, for practical purposes, identical. (In this example, the transformation is a special case of that due to Wilson and Hilferty 1931).] Thus, in Example 1, if we were to make the transformation $\zeta_i = \mu_i^{-1/3}$, we would obtain n observations $\hat{\zeta}_i = y_i^{-1/3}$, and the problem would now more closely resemble nonlinear regression. If the surface $\zeta = \zeta(\theta)$ turned out to be approximately linear, we would expect the likelihood function on θ to be approximately normal.

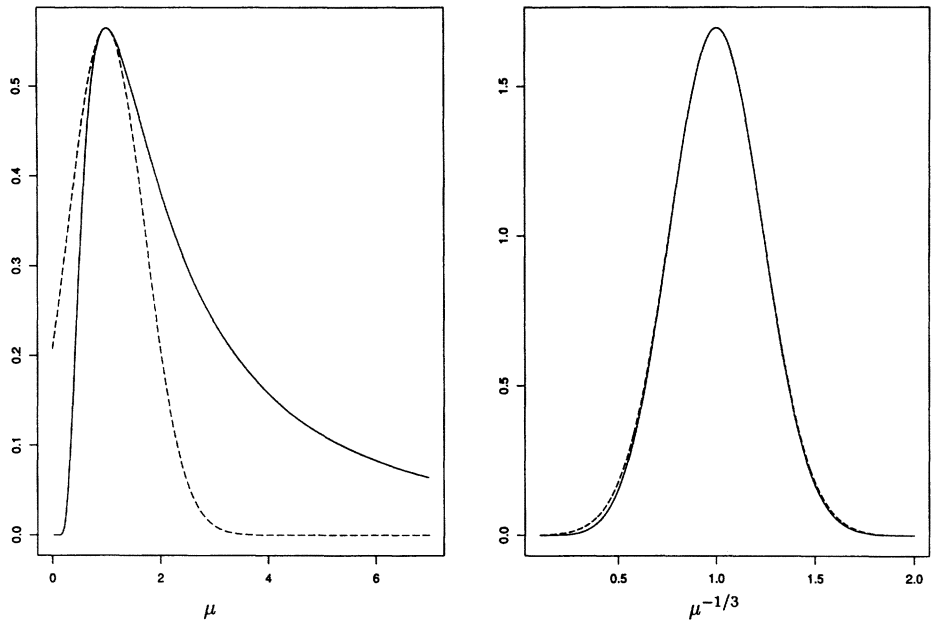

FIG. 1. *Likelihood (solid lines) and normal approximations (dashed lines) for* μ *and* $\mu^{-1/3}$ *with* $n = 2$ *and* $\bar{x} = 1$ *for the exponential model. The likelihoods have been scaled so that they match the value of the normal approximation at the MLE.*

In essence then, our method is to compute curvatures of the surface $\zeta = \zeta(\theta)$ following Bates and Watts (1980), except that we insert the Fisher information matrix into the formulas, to account for the inhomogeneity of the variances of the quantities $\widehat{\zeta}_i$. In Section 2.1 we provide necessary background on nonlinear regression; in Section 2.2 we describe the generalization to exponential family nonlinear regression models; in Section 2.3 we give the mathematical foundation based on α-connections (which provides extension of the methodology to any regular parametric family); and in Section 2.4 we give explicit formulae for the curvature measures.

What we have outlined so far concerns the use of expected third derivatives of the loglikelihood function: although for all regular exponential families the parameterization we begin with satisfies $\ell'''(\widehat{\tau}) = 0$, for other families this no longer holds. Thus, one might wish to consider diagnostics based directly on the observed third derivatives of the loglikelihood function or, in the Bayesian case, the log posterior. This we do in Section 3. The idea there is to generalize the one-dimensional standardized third derivative suggested by Sprott (1973) to higher dimensions. We do so by direct analogy with the diagnostics in Section 2, noting that, first, the curvature measures in nonlinear regression may be considered algebraically as summaries of the three-way array of second derivatives of the regression surface and, second, the third derivatives of the loglikelihood function again form a three-way array. Thus, we apply analogous three-way array summaries to obtain diagnostics based on the third derivatives.

One worry is that, in practice, at least in certain problems, it may be difficult or time-consuming to compute the third derivatives correctly. We show that one of our diagnostic third-derivative summaries may be computed approximately, with error of order $O(n^{-2})$, using "derivative-free" posterior expectation approximations described by Tierney, Kass and Kadane (1989) and available in LISP-STAT [Tierney (1990)].

In Section 4 we illustrate the approach, first returning to Example 1 and then analyzing a nonlinear binary response model as a second example. In Section 5 we discuss the methodology and our results.

2. General curvature measures. In this section we show how curvature measures of nonlinearity applied in normal nonlinear regression may be generalized to nonlinear models having other error structures. We begin with a review of pertinent results from normal nonlinear regression theory in Section 2.1, then present generalizations to exponential family nonlinear models in Section 2.2. In Section 2.3 we provide the mathematical foundation for the approach (which may be skipped by readers not wishing to wade through it), and in Section 2.4 we give the resulting curvature measures.

NOTATION. In writing components of parameters we will distinguish m-dimensional parameters, such as θ, from n-dimensional parameters, such as μ, by using as subscripts Latin letters at the beginning of the alphabet $a, b, c, \ldots$ for the former, and letters occurring later in the alphabet $i, j, k, \ldots$ for the latter. For arrays we will similarly distinguish components ranging from 1 to m from those ranging from 1 to n, but will also use Greek letters for those components that might range from either 1 to m, $m+1$ to n or 1 to n depending on the context.

2.1. *Nonlinear regression.* The usual normal nonlinear regression model begins with a model function $f(\theta, x)$ for $\theta \in \Theta \subseteq R^m$ and then

$$Y_i = \eta_i(\theta) + \varepsilon_i,$$

with $\eta_i(\theta) = f(\theta, x_i)$ and $\varepsilon_i \sim N(0, \sigma^2)$, independently, for $i = 1, \ldots, n$. Normality of the errors is not always assumed, but for the development of the generalization below we must begin with it here. The vector $Y = (Y_1, \ldots, Y_n)$ then has a distribution belonging to the n-dimensional multivariate normal family $N(\eta, \sigma^2 \cdot I_n)$, where $\eta = (\eta_1, \ldots, \eta_n)$ is restricted to lie on an m-dimensional surface specified by $\eta = \eta(\theta)$. Thus, for each fixed σ, this family is the subfamily of the n-dimensional normal location family that consists of distributions with location parameter restricted by $\eta = \eta(\theta)$, that is, with location parameter restricted by the nonlinear mapping $\theta \to \eta(\theta)$. Approximate inference is carried out through linear approximation of this nonlinear mapping, and diagnostics of poor performance of asymptotic inference procedures have been based on measures of the mapping's nonlinearity.

The starting point in assessing the accuracy of the linear approximation has been to examine the second derivatives of the model function $\eta_i(\theta) = f(\theta, x_i)$.

Beale (1960) made the geometrically elementary yet statistically interesting observation that the parameterization-invariant nonlinearity is associated with the normal components of the second derivatives. This implies that the three-way array of normal second-derivative components may be reduced to a parameterization-invariant scalar. As discussed by Kass [(1989), Section 3.5.5], there are two fairly natural such single-number summaries that have been studied in differential geometry and, as noted below, Beale used a linear combination of these as his measure of parameterization-invariant nonlinearity.

This leaves the parameterization-dependent nonlinearity to be assessed using the tangential components of the second-derivative array. Working by analogy with the normal components, this three-way array of tangential components may also be reduced to two different scalars. These are, of course, not fully parameterization-invariant, but they are invariant to affine transformations of the parameter space (i.e., transformations of the form $\theta \rightarrow A\theta + B$, where A is a full-rank $m \times m$ matrix and B is an $m \times 1$ vector). Affine invariance is desirable geometrically because affine transformations preserve linearity (or nonlinearity), and it is desirable statistically because they preserve normality (or nonnormality); that is, a diagnostic of nonnormality of the MLE (the least-squares estimator in the case of nonlinear regression) should not be affected by affine transformations, since these do not affect the degree to which the distribution of the MLE may be approximated by a normal distribution. The tangential components of the second derivatives are of primary interest here, and there is some evidence that they are the main concern in applications [Bates and Watts (1980) and Ratkowsky (1990), page 24].

We will now give several definitions, based mainly on the work of Beale (1960) and Bates and Watts (1980), that are relevant to the development in Sections 2.2 and 2.3. We first review the analysis used to assess parameterization-invariant nonlinearity, beginning with root-mean-squared curvature, and its calculation using a standardized array of normal components of second derivatives; summary of this second-derivative array is central to our motivation. We then present the analogous methods based on tangential, rather than normal, components of the second-derivative array, and we mention the interpretation of the array itself. We will generalize this array in Section 2.3 and interpret it in an example in Section 4.

Let $\widehat{\theta}$ be the least-squares estimator of θ. The measure of parameterization-invariant nonlinearity proposed by Beale (1960) is an average curvature among certain curves that slice through the regression surface at the least-squares fitted value $\eta(\widehat{\theta})$. The curves have the form $c_v(t) = \eta(\widehat{\theta} + tv)$ with $c_v(0) = \eta(\widehat{\theta})$, where v is a nonzero vector in R^m, and have been called *lifted lines* by Bates and Watts (1980). Each such curve has a second-derivative vector which decomposes into normal and tangential components $c_v''(0) = c_v''(0)_N + c_v''(0)_T$, and the curvature of the curve is

$$\kappa_N(v) = \|c_v'(0)\|^{-2}\|c_v''(0)_N\|. \tag{2.1}$$

This curvature may be considered a measure of curvature of the surface at $\eta(\widehat{\theta})$

in the direction of the tangent vector $c_v'(0)$. Beale's measure is then

$$\gamma_{\mathrm{RMS}}^2 = \frac{1}{A_m}\int_S \left(\kappa_N(v)\right)^2 dS, \tag{2.2}$$

where the integral is over the sphere $\{v: \|c_v'(0)\| = 1\}$, and $A_m = \pi^{m/2}/\Gamma(m/2)$ is the surface area of the unit sphere in R^m. [As Bates and Watts (1980) showed, Beale's N_ϕ is actually equal to $\frac{1}{4}\gamma_{\mathrm{RMS}}^2$.]

Computationally it is easiest to perform this integration by taking advantage of a standardization, introduced by Bates and Watts (1980), which is convenient for many calculations. There are three steps involved. Let TM be the tangent plane to the surface $\eta(\theta)$ at $\hat{\theta}$, and let $V = D\eta(\hat{\theta})$ so that the columns of V span TM. Geometrically, the first step is to rotate the n-dimensional Euclidean coordinate system so that the first m coordinates span TM, and the last $n-m$ are orthogonal to it. Algebraically, this is accomplished via the QR-decomposition, with $QR = V$. The second step is to introduce a linear transformation of the parameter θ such that the vector derivatives of η with respect to the components of the new parameter ϕ coincide with the orthonormal basis for TM in the rotated coordinate system. This transformation is

$$\phi = R_1(\theta - \hat{\theta}), \tag{2.3}$$

where R_1 is the $m \times m$ upper portion of the $n \times m$ matrix R. (The rotation here is not unique; one possibility is to choose R_1 to have positive diagonal elements, but the choice is inconsequential for what follows.) Finally, the surface in these new coordinates and parameterization becomes

$$\xi(\phi) - Q^T\left(\eta(\hat{\theta} + R_1^{-1}\phi) - \eta(\hat{\theta})\right),$$

and the second-derivative array A is defined by its components

$$a_{\lambda ab} = \frac{\partial \xi_\lambda}{\partial \phi_a \partial \phi_b}(\hat{\phi}), \tag{2.4}$$

with the index λ ranging over the m tangential components followed by the $n-m$ normal components as indicated in the notation $A = A^T \mid A^N$. Note that this definition does not include the normalizing factor $s \cdot m^{1/2}$, where s is the residual root-mean-square, used by Bates and Watts. The computation of the array may be carried out most conveniently using

$$A = \left[Q^T\right]\left[L^T D^2\eta(\hat{\theta})L\right], \tag{2.5}$$

where $L = R_1^{-1}$ and the brackets indicate multiplication of three-way-arrays, as in Bates and Watts (1980). Specifically, $A = [B][C]$ when the components satisfy $A_{\lambda ab} = \sum_{\mu=1}^n B_{\lambda\mu}C_{\mu ab}$.

The elements of the three-way array A^N could be examined for diagnostic purposes, but it is most convenient to consider invariant one-number summaries.

There are two such scalars that play an important role in the geometry of surfaces. They are analogous to two forms of squared trace for a matrix h, $\mathrm{tr}(hh^T)$ and $\mathrm{tr}(h)^2$. Including a factor of σ^2 in their definition, they are

$$\gamma^2 = \sigma^2 \sum_{\lambda=m+1}^{n} \sum_{b,c} (a_{\lambda bc})^2 \tag{2.6}$$

and

$$\bar{\gamma}^2 = m^{-2}\sigma^2 \sum_{\lambda=m+1}^{n} \left(\sum_{b} a_{\lambda bb} \right)^2, \tag{2.7}$$

where λ is summed over the $n-m$ normal components. Here we should perhaps remark that the use of the factor m in the definition of $\bar{\gamma}$ but not γ occurs as a notational quirk because the latter is σ times what is usually called the *mean curvature* of the surface, while γ is the *statistical curvature* [see Kass (1989), Sections 3.5.5 and 3.5.6). The root-mean-squared curvature may now be rewritten in terms of these summaries of the A array according to the relation

$$m(m+2)\gamma_{\mathrm{RMS}}^2 = m^2\bar{\gamma}^2 + 2\gamma^2. \tag{2.8}$$

This corrects the corresponding equation in Kass (1989), which omitted the factor $m(m+2)$.

The tangential-component analogues begin with

$$\kappa_T(v) = \|c_v'(0)\|^{-2} \left\| \left(c_v''(0)\right)_T \right\|, \tag{2.9}$$

which may be averaged over the unit sphere, as in (2.2),

$$\omega_{\mathrm{RMS}}^2 = \frac{1}{A_m} \int_S \left(\kappa_T(v)\right)^2 dS. \tag{2.10}$$

As in (2.6) and (2.7), we define

$$\omega^2 = \sigma^2 \sum_{\lambda=1}^{m} \sum_{b,c} (a_{\lambda bc})^2 \tag{2.11}$$

and

$$\bar{\omega}^2 = m^{-2}\sigma^2 \sum_{\lambda=1}^{m} \left(\sum_{b} a_{\lambda bb} \right)^2, \tag{2.12}$$

where λ is now summed over the m tangential components. We then have the tangential-component analogue of (2.8),

$$m(m+2)\omega_{\mathrm{RMS}}^2 = m^2\bar{\omega}^2 + 2\omega^2. \tag{2.13}$$

The quantities $\omega, \overline{\omega}$ and ω_{RMS}, which are computed using (2.5), are invariant to affine transformations of the parameter space. They are not standard geometrical objects, and $\kappa_T(v)$ is not a curvature in the usual sense. Instead of indicating nonlinearity inherent to the surface $\eta(\theta)$ sitting in R^n, they indicate nonlinearity due to the parameterization of the surface. For instance, in ordinary linear regression the regression surface is an m-dimensional plane in R^n. If the regression coefficients were replaced by new parameters arising from a nonlinear transformation (as would occur if the original regression coefficients were rewritten in terms of certain of their ratios and products), the surface would remain flat within R^n but would become nonlinear in these new parameters. Thus, the quantities $\lambda, \overline{\lambda}$ and λ_{RMS} would remain zero but $\omega, \overline{\omega}$ and ω_{RMS} typically would become nonzero.

Bates and Watts (1981) emphasized the value of examining the A^T array and its summaries to assess the effects of reparameterization. Referring to the ideal parameterization as providing a "uniform coordinate system" for the surface, they interpreted elements of the A^T array as indicating departures from uniformity. In particular, they mentioned what they called *compansion, arcing* and *fanning* entries and illustrated their effects graphically using a two-dimensional example. In Section 4 we provide a similar interpretation for the exponential regression model described in the Introduction.

2.2. *Generalization to exponential family nonlinear regression.* Consider a set of densities

$$p^{(1)}(z \mid \nu, \sigma) = \exp\Big\{ \big[z\nu(\theta) - \psi(\nu)\big]/\sigma^2 + b(z, \sigma) \Big\}$$

that, for fixed σ, form a one-dimensional regular exponential family with natural parameter ν. Allowing σ to vary produces a family that may be called a regular exponential dispersion model [Jorgensen (1987)]. Taking n copies of such a family and writing $y = (y_1, \ldots, y_n)$ with y_i replacing z, for $i = 1, \ldots, n$, and $\eta = (\eta_1, \ldots, \eta_n)$ with η_i replacing ν, for $i = 1, \ldots, n$, the product family of densities $p(y|\eta, \sigma) = \Pi_{i=1}^n p^{(1)}(y_i|\eta_i, \sigma)$ is, for fixed σ, a regular exponential family of order n. We let the natural parameter space of this family be denoted by N. When $\eta_i = f(\theta; x_i)$ for some function f, parameter vector θ and values of an explanatory variable or vector x_i, $i = 1, \ldots, n$, we obtain an *exponential family nonlinear model*, which will have densities of the form

$$p(y \mid \eta, \sigma) = \prod_{i=1}^{n} \exp\Big\{ \big[y_i\eta_i(\theta) - \psi(\eta_i)\big]/\sigma^2 + b(y_i, \sigma) \Big\}. \tag{2.14}$$

Generalized linear models [McCullagh and Nelder (1989)] are exponential family nonlinear models, and when the dispersion parameter σ is known, exponential family nonlinear models become curved exponential families. We assume the parameter space Θ is an open subset of R^m and $\theta \to \eta(\theta)$ is an embedding (i.e., the mapping from Θ into N is one-to-one and infinitely differentiable, with full-rank Jacobian and an infinitely differentiable inverse mapping).

As in the case of nonlinear regression, geometrical analysis of these models may be based on the embedding $\theta \to \eta(\theta)$, ignoring the presence of the parameter σ and effectively treating it as if it were known. We therefore write $\Omega_0 = \{p(\cdot|\eta(\theta), \sigma): \theta \in \Theta\}$ and take $\Omega = \{p(\cdot|\eta, \sigma): \eta \in N\}$ to be the unrestricted exponential dispersion model, speaking of θ and η as parameterizations for these models, ignoring σ. (In applications, when σ is needed for some formula an estimator of it may be substituted.) By construction, Ω is an n-fold product of a one-dimensional regular exponential family with itself, and we let $\Omega^{(1)}$ denote this one-dimensional exponential family.

In the exponential regression model described in the Introduction, $\Omega^{(1)}$ is the Exponential(μ^{-1}) family. In general, as noted at the end of Section 2.3, there exists a parameterization τ of $\Omega^{(1)}$, which we will call the *quadratic loglikelihood parameterization*, for which the loglikelihood function satisfies $E(\ell'''(\tau)) = 0$ and $\ell'''(\hat{\tau}) = 0$ [see Hougaard (1982)]. We can use this parameterization for each of the n copies of $\Omega^{(1)}$ comprising Ω to define a parameterization $\zeta = (\zeta_1, \ldots, \zeta_n)$ for Ω (ζ_i replacing τ, for $i = 1, \ldots, n$). We then obtain an m-dimensional surface $\zeta(\theta)$ in R^n that represents Ω_0. In this setting we will now be more explicit about the idea, motivated in the Introduction, of computing curvatures of the surface $\zeta(\theta)$.

The curvature measures in nonlinear regression are computed from the A array of (2.5). We carry out the steps leading to (2.5) using ζ in place of η, with one modification. In the full exponential family model Ω, the variances of the components $\hat{\zeta}_1, \ldots, \hat{\zeta}_n$ may be inhomogeneous. To accommodate this, we weight the observations according to the Fisher information matrix $i(\zeta)$. Thus, letting $G = i(\zeta(\hat{\theta}))$, we begin with the decomposition

$$QR = G^{1/2} D\zeta(\hat{\theta}) \tag{2.15}$$

and take R_1 to be the upper $m \times m$ part of R. We then define ϕ using (2.3) with this new definition of R_1, and we write the surface in the rotated coordinate system as

$$\xi(\phi) = Q^T G^{1/2}\big(\zeta(\hat{\theta} + R_1^{-1}\phi) - \zeta(\hat{\theta})\big).$$

As in (2.4), if we define

$$a_{\lambda ab} = \frac{\partial \xi_\lambda}{\partial \phi_a \, \partial \phi_b}(\hat{\phi}), \tag{2.16}$$

we arrive at the computational formula

$$A_\zeta = \big[Q^T G^{1/2}\big]\big[L^T D^2 \zeta(\hat{\theta}) L\big], \tag{2.17}$$

where $L = R_1^T$, the brackets again indicating multiplication of the three-way arrays, and the subscript ζ being used to indicate explicit dependence on the parameterization ζ. This new A array may be decomposed according to the first m and last $n - m$ values of the index λ as $A_\zeta = A_\zeta^T \mid A_\zeta^N$, representing tangential and normal components with respect to the inner product based on the Fisher information matrix G.

We state the method in terms of an arbitrary parameterization $\zeta = (\zeta_1, \ldots, \zeta_n)$ constructed similarly from components that are each a particular parameterization of $\Omega^{(1)}$, but not necessarily the one for which the expected third derivative of the loglikelihood vanishes. For instance, we might take the asymptotic variance stabilizing parameterization of $\Omega^{(1)}$ to define the components of ζ. To be more precise, if τ is a parameterization of $\Omega^{(1)}$, we may consider the parameter mapping $t: \Omega^{(1)} \to R$ defined by the operation on densities (elements of $\Omega^{(1)}$) as $t(p^{(1)}(\cdot|\nu(\tau), \sigma)) = \tau$. Then, rewriting each element of Ω according to its product form $p(\cdot|\eta, \sigma) = \prod_{i=1}^{n} p_i^{(1)}(\cdot|\eta_i, \sigma)$, where the subscript on $p_i^{(1)}$ indicates, of course, the i-th instance of the density, we define a parameterization $\zeta = (\zeta_1, \ldots, \zeta_n)$ of Ω from $\zeta_i = t(p_i^{(1)}(\cdot|\eta_i, \sigma))$, for $i = 1, \ldots, n$. We will call ζ the *product parameterization* of Ω defined from τ.

METHOD FOR EXPONENTIAL FAMILY NONLINEAR MODELS. For the product parameterization ζ of Ω constructed as described above from a parameterization τ of $\Omega^{(1)}$, we consider the array A_ζ to be the generalization of the A array.

Once A_ζ is substituted for A, the calculation of curvatures proceeds exactly as in the nonlinear regression setting. Explicit formulae are given in Section 2.4. We think of τ as the quadratic loglikelihood parameterization, but we left it unspecified in the procedural statement above because other choices may be of interest as well. Expressions for the quadratic loglikelihood parameterization in the NEF-QVF families may be found in Slate (1994).

2.3. *Foundation using α-connections.* In Section 2.2 we described a generalization to exponential family nonlinear models of the normalized second-derivative array $A = A^T|A^N$ in nonlinear regression. In this section we provide a foundation for the calculations using the α-connections introduced by Amari (1982). For simplicity, and because the parameter-effects portion of the array are of primary interest in practice, we confine ourselves to a discussion of parameter effects. We begin by pointing out that the array A^T may be considered a special case of a connection coefficient array. We then review relevant definitions and results from Amari (1982) and indicate the way the α-connection coefficient arrays provide generalizations of the A^T array defined in Section 2.1.

To a reader not knowledgeable about differential geometry, we try to give a rough idea of the way connection coefficients enter into derivative calculations on a p-dimensional smooth manifold M. For us, p could be either the sample size n or the dimension of the parameter space m. If $E_1, \ldots, E_p$ are the β-coordinate basis tangent vectors in the tangent space of M at β and if $\nabla_\lambda = \nabla_{E_\lambda}$ is the covariant derivative (specified by the affine connection on M) in the λth coordinate direction, then

$$\nabla_\lambda E_\mu = \sum_\nu \Gamma^\nu_{\lambda\mu}(\beta) \cdot E_\nu,$$

that is, $\Gamma^\nu_{\lambda\mu}(\beta)$ is the νth component of the covariant derivative of the μth coordinate tangent vector in the direction of the λth coordinate. For rectangular coordinates in Euclidean space ∇_λ is the usual λth partial derivative operator,

the basis vectors are constant throughout R^p and the connection coefficients vanish. For curvilinear coordinates, however, the basis tangent vectors vary from point to point and the coefficients do not vanish. The reader may recall that for polar coordinates (r, θ) in R^2 the gradient of a real-valued function f is the vector $(\partial f/\partial r, (1/r)(\partial f/\partial \theta))$, the appearance of the coefficient $1/r$ being a consequence of the curvilinearity. Connection coefficients enter analogously for curvilinear coordinates when vector fields are differentiated, as they are implicitly in the calculation of curvatures. Thus, the connection coefficients of the Euclidean connection vanish in rectangular coordinates but not in polar coordinates. When M_0 is a submanifold of M, there is a "natural" manner in which a connection is inherited on M_0 from that on M; the most familiar case is when M_0 is an m-dimensional smooth surface in $M = R^n$. This latter case is that which occurs in nonlinear regression, with the regression surface being M_0.

Having given this description we may return to the A^T array for nonlinear regression (Section 2.1) and note that

$$a_{cab} = \Gamma^c_{ab}(\phi), \tag{2.18}$$

where ϕ is defined in (2.3) and $\Gamma^c_{ab}(\phi)$ are the connection coefficients in ϕ coordinates for the connection inherited on the regression surface from the Euclidean connection on R^n. This observation is of interest because it indicates the role that connection coefficients play in determining the effects of parameterization.

2.3.1. *General method.* We will now define the α-connection coefficients and use them to obtain another version of (2.16). Our purpose is to understand the method for exponential family models (Section 2.2) more deeply and thereby generalize it. In this section we use (2.18) to provide a sense in which certain α-connection coefficients may be considered generalizations of the A^T array, and in the next section we show that this does indeed produce a generalization of the method of Section 2.2.

In general, if β is a parameterization of a regular parametric family of densities $\mathcal{P}$ and $\ell(\beta)$ is the loglikelihood function, then the *information metric* has the Fisher information matrix $i(\beta)$ as its β-coordinate expression and we will write

$$g_{\lambda\mu} = i(\beta)_{\lambda\mu} = E\left(\frac{\partial \ell}{\partial \beta_\lambda}\frac{\partial \ell}{\partial \beta_\mu}\right). \tag{2.19}$$

The elements of the inverse of this matrix will be written

$$g^{\lambda\mu} = [i(\beta)^{-1}]_{\lambda\mu}.$$

We use $\langle \cdot, \cdot \rangle_{i(\beta)}$ and $\| \cdot \|_{i(\beta)}$ to denote the information inner product and its norm, so that, for a vector u in the β-space $\|u\|^2_{i(\beta)} = \sum_{\lambda,\mu} g_{\lambda\mu} u_\lambda u_\mu$. The α-connection coefficients are defined by

$$\overset{1}{\Gamma}_{\lambda\mu\nu}(\beta) = E\left(\frac{\partial^2 \ell}{\partial \beta_\lambda \partial \beta_\mu}\frac{\partial \ell}{\partial \beta_\nu}\right),$$

$$\overset{-1}{\Gamma}_{\lambda\mu\nu}(\beta) = \overset{1}{\Gamma}_{\lambda\mu\nu}(\beta) + E\left(\frac{\partial \ell}{\partial \beta_\lambda}\frac{\partial \ell}{\partial \beta_\mu}\frac{\partial \ell}{\partial \beta_\nu}\right)$$

and

$$\overset{\alpha}{\Gamma}_{\lambda\mu\nu}(\beta) = \frac{1-\alpha}{2}\overset{-1}{\Gamma}_{\lambda\mu\nu}(\beta) + \frac{1+\alpha}{2}\overset{1}{\Gamma}_{\lambda\mu\nu}(\beta).$$

Note that $\overset{1}{\Gamma}_{\lambda\mu\nu}$ and $\overset{-1}{\Gamma}_{\lambda\mu\nu}$ correspond to $\alpha = 1$ and $\alpha = -1$, respectively; these are the *exponential* and *mixture* connection coefficients. The $\overset{\alpha}{\Gamma}_{\lambda\mu\nu}$ coefficients are in covariant form. An alternative form includes one contravariant (or "raised") index and is defined by

$$\overset{\alpha}{\Gamma}{}^{\nu}_{\lambda\mu}(\beta) = \sum_{\kappa} g^{\nu\kappa}\overset{\alpha}{\Gamma}_{\lambda\mu\kappa}(\beta).$$

Now, by virtue of the very special structure of normal location models, in normal-family nonlinear regression it happens that the α-connection coefficients are equal to the Euclidean connection coefficients, for all α. Thus, in the usual nonlinear regression setting we have,

$$a_{cab} = \overset{\alpha}{\Gamma}{}^{c}_{ab}(\phi), \tag{2.20}$$

for all α [cf. (2.16)]. Geometrically, (2.20) provides a sense in which the α-connection coefficients in terms of ϕ may be considered generalizations of the A^T array.

To obtain statistical motivation for the use of α-connection coefficient arrays, we consider the equations

$$\overset{\alpha}{\Gamma}{}^{k}_{ij}(\beta) = 0,$$

$i = 1,\ldots,p; j = 1,\ldots,p; k = 1,\ldots,p$. Reinterpreting a formula discussed by Hougaard (1982), Kass (1984) observed that these equations characterize various parameterizations. If the equations hold for $\alpha = 1, 0, -\frac{1}{3}, -1$, then β is, respectively, the natural parameterization, the variance-stabilizing parameterization, the asymptotic skewness-reducing parameterization and the mean-value parameterization. These values of α occur in the work of Amari (1982, 1985). In addition, and more important for the methodology described here, if the equations hold for $\alpha = \frac{1}{3}$, then β is the parameterization in which the expected third derivatives of the loglikelihood function vanish. Thus, the latter parameterization provides a "uniform coordinate system" in the sense of Bates and Watts (1981) with respect to the $\alpha = \frac{1}{3}$ geometry. Departures from uniformity in this geometry therefore correspond to deviations away from the "quadratic loglikelihood parameterization."

With this motivation, we specify the general method. In doing so, although $\alpha = \frac{1}{3}$ deserves special attention, we allow the value of α to remain arbitrary. For any given parameterization θ we take

$$\phi = i(\hat{\theta})^{1/2}(\theta - \hat{\theta}), \tag{2.21}$$

where $i(\theta)$ is the Fisher information in θ, and $\hat{\theta}$ is the MLE. In this parameterization the information matrix at the MLE $\hat{\phi} = 0$ becomes the identity and

$$\overset{\alpha}{\Gamma}{}^{c}_{ab}(\hat{\phi}) = \overset{\alpha}{\Gamma}_{abc}(\hat{\phi}). \tag{2.22}$$

Notice, however, that since θ is a full-rank linear transformation of ϕ, $\overset{\alpha}{\Gamma}{}^{c}_{ab}(\hat{\phi}) = 0$ if and only if $\overset{\alpha}{\Gamma}{}^{c}_{ab}(\hat{\theta}) = 0$. As in the nonlinear regression setting, it is convenient to assess the parameterization θ by using its linearly standardized version ϕ.

GENERAL METHOD. We interpret and compute curvatures from $\overset{\alpha}{\Gamma}{}^{c}_{ab}(\hat{\phi})$ by analogy with the interpretation of, and curvatures based on, A^T.

2.3.2. *Method for exponential family nonlinear models.* We now show that the method of Section 2.2 is a special case of that given in Section 2.3.1.

In nonlinear regression the model of interest, which is determined by the restrictions $\eta_i(\theta) = f(\theta, x_i)$, is considered an m-dimensional submodel of the unrestricted n-dimensional multivariate Normal$(\eta, \sigma^2 I_n)$ model. As indicated in Section 2.3.1, its geometry is thereby inherited from the Euclidean geometry of the unrestricted normal family, for which the connection coefficients vanish. Correspondingly, in general, if we have a subfamily Ω_0 of a larger family Ω, the α-connection coefficients on Ω_0 may be determined by inheritance formulas from the α-connection coefficients on Ω. This assumes the regularity condition that Ω_0 is an imbedded submanifold [see, e.g., Kass (1989)].

PROPOSITION 1. *If ζ is a parameterization of a family Ω such that $\overset{\alpha}{\Gamma}_{ijk}(\zeta) = 0$, for $i, j, k = 1, \ldots, n$, and if θ is a parameterization of Ω_0, then, with ϕ satisfying* (2.21), *the α-connection coefficients on Ω_0 satisfy*

$$\overset{\alpha}{\Gamma}_{abc}(\hat{\phi}) = \langle \partial_{ab}\zeta, \partial_c \zeta \rangle_{i(\zeta(\hat{\theta}))} = \sum g_{ij}\, \partial_{ab}\zeta_i\, \partial_c \zeta_j, \tag{2.23}$$

where the derivatives are evaluated at $\hat{\theta}$.

PROOF. This is an immediate consequence of Theorem 3 and equation (4.15) of Amari (1982). □

Now suppose Ω_0 is an exponential family nonlinear model, constructed as a subfamily of $\Omega = \Omega^{(1)} \times \cdots \times \Omega^{(1)}$.

PROPOSITION 2. *For each α there exists a parameterization τ of $\Omega^{(1)}$, unique up to an affine transformation, such that the product parameterization ζ of Ω defined from τ satisfies $\overset{\alpha}{\Gamma}_{ijk}(\zeta) = 0$, for $i, j, k = 1, \ldots, n$.*

PROOF. As shown in Kass (1984), for any regular one-dimensional exponential family $\Omega^{(1)}$, τ satisfies $\overset{\alpha}{\Gamma}_{111}(\tau) = 0$ if and only if it is the solution of a

second-order differential equation discussed in Hougaard (1982). Such solutions exist and are unique up to an affine transformation. The result follows from the product-structure of Ω. □

EXAMPLE 1 (Continued). In the exponential regression example described in the Introduction, if $\Omega^{(1)}$ is the Exponential(μ^{-1}) family, then $\tau = \mu^{-1/3}$ satisfies $\overset{\alpha}{\Gamma}_{11}^{1}(\tau) = 0$ for $\alpha = \frac{1}{3}$, this being the only coefficient since the family is one-dimensional. We may parameterize $\Omega = \Omega^{(1)} \times \cdots \times \Omega^{(1)}$ by $\mu = (\mu_1, \ldots, \mu_n)$; but if we instead use $\zeta = (\mu_1^{-1/3}, \ldots, \mu_n^{-1/3})$, then $\overset{\alpha}{\Gamma}_{ijk}(\zeta) = 0$, for $\alpha = \frac{1}{3}$. (Note that it is a coincidence of no apparent importance that the transformation for $\alpha = \frac{1}{3}$ involves the power $-\frac{1}{3}$ in this example.)

We now consider the nonlinear surface in the ζ-space, defined by $\theta \to \zeta(\theta)$, rather than in the natural parameter space of (2.14) defined by $\theta \to \eta(\theta)$ [and rather than in the mean-value parameter space defined by $\theta \to \mu(\theta)$, with which we began Example 1]. Doing so, we recognize (2.23) as a computation of the tangential components of the second derivatives of ζ, with respect to ϕ. This is precisely the way the A^T array was defined in (2.4), except that the information inner product has been used in (2.23).

It is straightforward to verify that $i(\theta) = R_1^T R_1$ as defined from (2.15) so that (2.21) is satisfied when ϕ is defined from (2.3). Since the contravariant and covariant forms of the connections coincide at $\hat{\phi}$ as in (2.22), Proposition 1 and (2.16) give

$$\overset{\alpha}{\Gamma}{}_{ab}^{c}(\hat{\phi}) = a_{cab}.$$

Together with Proposition 2, this shows that when ζ is a product parameterization defined from τ with $\overset{\alpha}{\Gamma}_{111}(\tau) = 0$, the method of Section 2.2 may be applied (essentially uniquely) and becomes a special case of the general method of Section 2.3.1.

2.4. *Curvature measures.* As we have already indicated, the curvature measures we will use are analogous to those of Section 2.1. In this section we will write specific formulae for the case of an exponential family nonlinear model Ω_0 and, as in Section 2.2, will work with the product parameterization ζ determined from an arbitrary parameterization τ of $\Omega^{(1)}$. Three choices of τ will be highlighted. First, and most important, there is the quadratic log-likelihood parameterization. The curvatures based on this parameterization will be subscripted by Q. In addition, when the natural parameterization and the mean-value parameterization of $\Omega^{(1)}$ are used, the curvatures will carry the subscripts of E and M, respectively. These latter two come from the names "exponential" and "mixture," which are part of the general terminology used in the geometrical foundation discussed in Section 2.3. In terms of that foundation, we would be defining τ from α to obtain the characterization on Ω, $\overset{\alpha}{\Gamma}_{ijk}(\zeta) = 0$ for all i, j, k, and the subscripts of Q, E and M correspond to $\alpha = \frac{1}{3}, 1$ and -1.

Another possible choice of τ is the MLE skewness-reducing parameterization, which corresponds to $\alpha = -\frac{1}{3}$. However, in assessing skewness the usual asymptotic pivot $I(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta)$, where $I(\widehat{\theta})$ is either observed or expected information, may be considered more relevant (since inferences are based on the pivot); its skewness is reduced [to order $O(n^{-3/2})$] when the quadratic loglikelihood parameterization is chosen [Hougaard (1982) and DiCiccio (1984)].

Using the subscript X generically to indicate E, Q, or M (or any other of the infinitely many possible choices of τ or α used to determine ζ), we define the following:

$$\gamma_X^2 = \sum g^{ac} g^{bd} g_{ij} \cdot \left((\partial_{ab}\zeta)_N\right)_i \left((\partial_{cd}\zeta)_N\right)_j \tag{2.24}$$

$$m^2\overline{\gamma}_X^2 = \sum g^{ab} g^{cd} g_{ij} \cdot \left((\partial_{ab}\zeta)_N\right)_i \left((\partial_{cd}\zeta)_N\right)_j \tag{2.25}$$

$$\omega_X^2 = \sum g^{ac} g^{bd} g_{ij} \cdot \left((\partial_{ab}\zeta)_T\right)_i \left((\partial_{cd}\zeta)_T\right)_j \tag{2.26}$$

$$m^2\overline{\omega}_X^2 = \sum g^{ab} g^{cd} g_{ij} \cdot \left((\partial_{ab}\zeta)_T\right)_i \left((\partial_{cd}\zeta)_T\right)_j \tag{2.27}$$

where g_{ij} is the $n \times n$ information matrix in terms of ζ at $\zeta(\widehat{\theta})$; g^{ab} is the inverse of the $m \times m$ information matrix in terms of θ at $\widehat{\theta}$; $(\cdot)_N$ and $(\cdot)_T$ signify normal and tangential components with respect to the information matrix (g_{ij}); and the summations are over all indices.

Certain special cases, and combinations, of these measures have important statistical interpretations [Kass (1989), Section 3.5]. For instance, γ_E is a generalization of what Efron (1975) called statistical curvature and is a measure of the insufficiency of the MLE. The quantity $m^2\overline{\omega}_M^2/4$ is the relative bias of the MLE, and $\gamma_E^2 + \omega_M^2$ is the second-order risk of the MLE using a quadratic loss function defined by the information matrix. Also, when the curvatures here are defined in terms of any α-connection, in the special case of nonlinear regression they will reduce to curvatures defined in Section 2.1.

We now introduce generalizations of the root-mean-squared curvatures. We begin with a lifted line in the ζ-parameter space $c_v(t) = \zeta(\widehat{\theta} + tv)$ at $c(0) = \zeta(\widehat{\theta})$, where v is a vector in R^m, and define

$$\kappa_{X,N}(v) = \|c_v'(0)\|_{i(\zeta(\hat{\theta}))}^{-2} \left\| \left(c_v''(0)\right)_N \right\|_{i(\zeta(\hat{\theta}))}$$

and

$$\kappa_{X,T}(v) = \|c_v'(0)\|_{i(\zeta(\hat{\theta}))}^{-2} \left\| \left(c_v''(0)\right)_T \right\|_{i(\zeta(\hat{\theta}))}$$

exactly as in (2.1) and (2.9), except that the information inner product [defined following (2.19)] is used to define the norm $\|\cdot\|_{i(\zeta(\hat{\theta}))}$. We then have

$$\gamma_{X,\mathrm{RMS}}^2 = \frac{1}{A_m} \int_S \left(\kappa_{X,N}(v)\right)^2 dS$$

and

$$\omega^2_{X,\,\mathrm{RMS}} = \frac{1}{A_m}\int_S \left(\kappa_{X,\,T}(v)\right)^2 dS,$$

where the integrals are over the sphere $\{v\colon \|c_v'(0)\|_{i(\zeta(\hat\theta))} = 1\}$.

All of these curvature measures may be easily computed by virtue of the following two propositions.

PROPOSITION 3. *Reparameterizing Ω_0 by ϕ defined by* (2.3) *via* (2.15), *the expressions for $\gamma_X, \overline{\gamma}_X, \omega_X$ and $\overline{\omega}_X$ are given by* (2.6), (2.7), (2.11) *and* (2.12), *where the subscript X determines the choice of ζ in* (2.15) *and where A_ζ array is defined in* (2.16) *and* (2.17).

PROOF. Applying (2.24)–(2.27) in terms of ϕ, this follows immediately from the simplification $g^{ab} = 1$ if $a = b$ and $g^{ab} = 0$ otherwise, together with definitions (2.16) and (2.17). □

PROPOSITION 4. *For any choice of ζ, indicated by X,*

$$m(m+2)\omega^2_{X,\,\mathrm{RMS}} = m^2\overline{\omega}^2_X + 2\omega^2_X,$$
$$m(m+2)\gamma^2_{X,\,\mathrm{RMS}} = m^2\overline{\gamma}^2_X + 2\gamma^2_X.$$

PROOF. Applying Proposition 3, the argument follows that of Bates and Watts (1980), leading to their equation (2.29). □

Although it is simplest, and may be most desirable, to use the curvature measures $\gamma_X, \overline{\gamma}_X, \omega_X$ and $\overline{\omega}_X$ (and the resulting root-mean-squared curvatures), another possibility would be to follow Bates and Watts (1980) and others by computing maximal curvatures over all directions in the parameter space. The method described by Bates and Watts (1980) is applicable here with obvious modifications.

3. Summaries of the observed third-derivative array. In Section 2 we emphasized curvature measures that would be zero if the third derivative of the loglikelihood function vanished in expectation. These should, in many problems, provide useful indications of whether the loglikelihood function is approximately quadratic. In this section, we instead summarize directly the "observed" third derivatives, that is, the third derivatives of the loglikelihood function $\ell(\theta)$ evaluated at the MLE $\hat\theta$ and the log posterior $\tilde\ell(\theta)$ evaluated at the mode $\tilde\theta$. For reasons given earlier, in Section 2.1, it is desirable to obtain one-number summaries that are invariant to affine transformations of the parameter space. We do so; then we note an interesting property of one of these measures of nonnormality, which provides an easy method of computing it.

Since the third derivatives form three-way arrays, we work by analogy with the affine-invariant reduction of the three-way array of second derivatives of

$\zeta(\theta)$ employed in Section 2.4, and we produce scalars that are analogous to ω and $\overline{\omega}$. Letting τ_{ab}^{c} be the c-component of $(\partial_{ab}\zeta)_T$ [formally, the component of $(\partial_{ab}\zeta)_T$ multiplying ∂_c in the $\partial_1, \ldots, \partial_m$ basis of θ-coordinate tangent vectors],

$$m^2\overline{\omega}^2 = \sum_{a,b,c,d,e,f} g^{ab}g^{de}g_{cf}\tau_{ab}^{c}\tau_{de}^{f}$$

and, using $\sum_b g^{ab}g_{bc} = 1$ if $a = c$ and 0 otherwise, and $\tau_{ab}^{c} = \sum_d g^{cd}\tau_{abd}$,

$$m^2\overline{\omega}^2 = \sum_{a,b,c,d,e,f} g^{ab}g^{de}g^{cf}\tau_{abc}\tau_{def}.$$

Letting the second and third partial derivatives be denoted by $\tilde{g}_{ab} = -\partial_{ab}\tilde{\ell}(\tilde{\theta})$ and $\partial_{abc}\tilde{\ell} = \partial_{abc}\tilde{\ell}(\tilde{\theta})$ we substitute $\partial_{abc}\tilde{\ell}$ and $\partial_{def}\tilde{\ell}$ for τ_{abc} and τ_{def}, and also $\tilde{g}_{ab}$ for g_{ab}, and so forth. We thereby obtain the summary we will define by

$$\overline{B}^2 = m^{-2} \sum_{a,b,c,d,e,f} \tilde{g}^{ab}\tilde{g}^{de}\tilde{g}^{cf}\,\partial_{abc}\tilde{\ell}\,\partial_{def}\tilde{\ell}.$$

Similarly, we define

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

If we take the prior to be uniform, then the above quantities are based on the MLE, the observed information and the observed third derivatives of the loglikelihood. In this case, both B and $\overline{B}$ may be considered generalizations of the normalized third derivative used by Sprott (1973) to measure nonnormality of the MLE in one-parameter problems.

The quantity $\overline{B}$ may be given an additional interpretation in terms of the *posterior bias* of the mode, defined as $\tilde{\theta} - \overline{\theta}$, where $\overline{\theta}$ is the posterior mean. Since the posterior bias is a vector, we consider the *relative posterior bias* defined by

$$R = (\tilde{\theta} - \overline{\theta})^T\tilde{G}(\tilde{\theta} - \overline{\theta}),$$

where $\tilde{G}$ is the matrix having components $\tilde{g}_{ab}$. This is an affine-invariant scalar. Now, from Kass, Tierney and Kadane [(1990), equation (2.6)],

$$\tilde{\theta}_r - \overline{\theta}_r \doteq \tfrac{1}{2}\sum_{s,t,u} \tilde{g}^{rs}\tilde{g}^{tu}\,\partial_{tus}\tilde{\ell}$$

(where s, t and u are summed from 1 to m), with an error of order $O(n^{-2})$ for any sequence of observations $y_1, y_2, \ldots$ satisfying regularity conditions specified by those authors (as in that paper, the statement here concerns a specific sequence and is not probabilistic, although under certain assumptions such data sequences will occur with probability 1). Thus, replacing t, u, s with a, b, c and

using v, w to index the quadratic-form multiplication in R (again summed from 1 to m),

$$R \doteq \tfrac{1}{4} \sum_{a,b,c,d,e,f,v,w} \tilde{g}^{vc}\tilde{g}^{ab}\,\partial_{abc}\tilde{\ell}\,\tilde{g}_{vw}\tilde{g}^{wf}\tilde{g}^{de}\partial_{def}\tilde{\ell}.$$

Since $\sum_{v,w}\tilde{g}^{vc}\tilde{g}^{wf} = \tilde{g}^{cf}$,

$$R \doteq \tfrac{1}{4}m^2\overline{B}^2,$$

again with an error of order $O(n^{-2})$.

Thus, not only was the quantity $\overline{B}^2$ defined here via its formal analogy with $\overline{\omega}$, but in fact it plays the same role in the leading term of the relative posterior bias of the mode as does $\overline{\omega}$ in the leading term of the relative bias of the MLE. [This remark was made in Kass (1989).] Furthermore, the result not only provides additional interpretation of $\overline{B}$, it also gives a fast and convenient method of computing it, approximately. Letting $\overline{\theta}^*$ be a second-order approximation to the posterior mean $\overline{\theta}$, such as discussed in Tierney and Kadane (1986) and Tierney, Kass and Kadane (1989), we may define $R^* = (\tilde{\theta} - \overline{\theta}^*)^T\tilde{G}(\tilde{\theta} - \overline{\theta}^*)$ and obtain

$$m^2\overline{B}^2 \doteq 4R^*, \tag{3.1}$$

which once again holds with error of order $O(n^{-2})$. We emphasize that this relationship may be used in either the Bayesian or non-Bayesian approach: by requiring only second derivatives it provides a relatively easy method of obtaining a general diagnostic based on the third derivatives of either the log-likelihood or the log posterior density. Existing computer code in LISP-STAT [Tierney (1990)] allows the user to avoid explicit calculation of derivatives and to specify only the loglikelihood function.

To judge how small the quantity $\overline{B}$ should be in order to consider the normal approximation adequate, Slate (1991) carried out detailed investigations in one-parameter problems and determined posterior distributions to be adequately normal according to a tail-probability criterion when $\overline{B} < 0.36$. From the structure of R^* it seems reasonable to use the cutoff value $m^2\overline{B}^2 < m\cdot(0.36)^2$ (or, roughly, $0.15m$) in m-parameter problems (since, for m identical independent posteriors, the joint value of R^* would be m times its value for each of the m marginals evaluated separately). This seems to us conservative in the sense that substantially larger values of $m^2\overline{B}^2$ would be worrisome but distributions having only somewhat larger values might still be adequately normal for many purposes.

An additional result connecting the present section to the previous one comes from viewing third derivatives as second derivatives of first derivatives: a new third-derivative summary may be derived as a curvature-like second-derivative summary of the first-derivative vector. For $h \in R^m$ we may consider the "lifted-line"

$$c_h(t) = \left(\partial_1\tilde{\ell}(\tilde{\theta} + th), \ldots, \partial_m\tilde{\ell}(\tilde{\theta} + th)\right)^T$$

and compute its curvature. In doing so, we will use the inner product defined by the modal covariance matrix $G^{-1} = \widehat{\Sigma}$. Intuitively, the matrix G is here replaced by G^{-1} because G used previously was the inverse of the approximate covariance matrix of θ and here G^{-1} is the approximate covariance matrix of $(\partial_1\widetilde{\ell}, \ldots, \partial_m\widetilde{\ell})$. Thus, we compute $\kappa(h) = \|c_h''(0)\|_{G^{-1}}/\|c_h'(0)\|_{G^{-1}}^2$ and then take the spherical average to define

$$B_{\text{RMS}}^2 = \frac{1}{A_m}\int_S \kappa(h)^2\, dS,$$

where the integral is over the sphere $\{h: \|c_h'(0)\|_{G^{-1}} = 1\}$. This quantity is analogous to $\omega_{Q,\text{RMS}}^2$ and might be considered a Bayesian version of the nonlinearity measure proposed by Beale (1960). By calculations like those leading to Proposition 4, we obtain

$$B_{\text{RMS}}^2 = \frac{m^2\overline{B}^2 + 2B^2}{m(m+2)}.$$

This result emphasizes the formal similarity of the third derivative summaries proposed here to those examined in Section 2.

4. Examples

EXAMPLE 1 (Continued). We return to the leukemia data set and model of Feigl and Zelen (1965), introduced in Section 1, and consider the 17 AG-positive patients. Maximization of the loglikelihood produces the MLE $\widehat{\theta}$ =(51, 1.11). Using a uniform prior on (θ_1, θ_2), this becomes the posterior mode as well. From the Hessian of the loglikelihood evaluated at the MLE we obtain approximate standard deviations for $\widehat{\theta}_1$ and $\widehat{\theta}_2$ (or posterior standard deviations for θ_1 and θ_2) of 12 and 0.41, with an approximate correlation of 0.00. These quantities determine the large-sample normal approximations to the distribution of the MLE and the posterior distribution.

In addition to θ, we also consider the parameterizations β and λ given, respectively, by $\mu_i = (\beta_1 x_i^*)^{\beta_2}$, where $x_i^* = \exp(-x_i)$, and $\mu_i = \exp\{\lambda_1 - \lambda_2 x_i\}$.

We begin by interpreting the parameter-effects arrays A_ζ^T based on three choices of ζ, shown in Table 1. We emphasize the array computed using the quadratic loglikelihood parameterization, labeled Q, although we include the arrays based on the natural parameterization (E) and the mean-value parameterization (M) for completeness. First, note that all elements of the A_ζ^T array for λ are small while, for $\theta, a_{111} = 0.32$ and, for $\beta, a_{111} = 0.30$, $a_{112} = -1.6$ and $a_{122} = 11.8$. We may interpret these as in Bates and Watts (1981), noting that a substantial a_{111} *compansion* element indicates that the ϕ_2-parameter curves will be nonuniformly spaced, a large a_{122} *arcing* element indicates that the ϕ_2-parameter curves will bend and a large negative a_{112} *fanning* element will cause convergence of the ϕ_2-parameter curves. We verify this interpretation by drawing tangent plane representations of the ϕ-parameterization curves, again by analogy with Bates and Watts (1981). The pictures appear in Figure 2. It

TABLE 1

The A_ζ^T arrays corresponding to the Q ($\alpha = \frac{1}{3}$), M ($\alpha = -1$) and E ($\alpha = 1$) choices of ζ for Example 1

	Parameterization					
A^T	λ		θ		β	
	0.0808	0.0000	0.3234	0.0000	0.2995	−1.6382
	0.0000	0.0808	0.0000	0.0808	−1.6382	11.7627
Q						
	0.0000	0.0808	0.0000	0.0808	0.0000	0.0808
	0.0808	0.0024	0.0808	0.0024	0.0808	0.0024
	−0.2425	0.0000	0.0000	0.0000	−0.0239	−1.6382
	0.0000	−0.2425	0.0000	−0.2425	−1.6382	11.4393
M						
	0.0000	−0.2425	0.0000	−0.2425	0.0000	−0.2425
	−0.2425	−0.0072	−0.2425	−0.0072	−0.2425	−0.0072
	0.2425	0.0000	0.4851	0.0000	0.4611	−1.6382
	0.0000	0.2425	0.0000	0.2425	−1.6382	11.9244
E						
	0.0000	0.2425	0.0000	0.2425	0.0000	0.2425
	0.2425	0.0072	0.2425	0.0072	0.2425	0.0072

may be seen that the ϕ_2-parameter curves for β (holding ϕ_1 constant) do indeed show extreme bending. This bending to some extent conceals the strong tendency of the ϕ_2-parameter curves to converge as ϕ_2 increases, and it also conceals the much smaller yet noticeable tendency for the ϕ_2-parameter curves to be unevenly spaced: they are somewhat closer together for ϕ_1 negative and get further apart as ϕ_1 increases. This compansion may be seen more clearly in the θ parameterization. It may be contrasted with the λ parameterization, for which the grid is fairly uniform.

To check the relevance of the overall conclusions, we display the contours of the likelihood in Figure 3, together with elliptical regions based on observed information that have approximate 95% confidence (and approximate 95% posterior probability). It may be seen that λ does provide a quite good parameterization, as predicted from the A_ζ^T array, while θ is worse and β is terrible.

Turning to the curvature measures, in Table 2, the most important conclusions would be based on ω_Q and $\overline{\omega}_Q$ (of Section 2.4) and on B and $\overline{B}$. These indicate, appropriately, that λ improves on θ while β is disastrous. A similar conclusion is also obtained from $\omega_{Q,\,\mathrm{RMS}}$. In addition to verification of these conclusions by direct examination of the likelihood contours, we may also evaluate posterior probabilities for the approximate inference regions. We did so by taking the prior to be uniform on the parameterization λ and using Monte Carlo importance sampling for the computations. We found that the ellipsoids having putative 95% probability actually had probability 0.930 ($\pm$0.003), 0.903 ($\pm$0.004) and 0.56 ($\pm$0.04) for the parameterizations λ, θ and β, respectively. Once again, these values are consistent with the indications of the diagnostics.

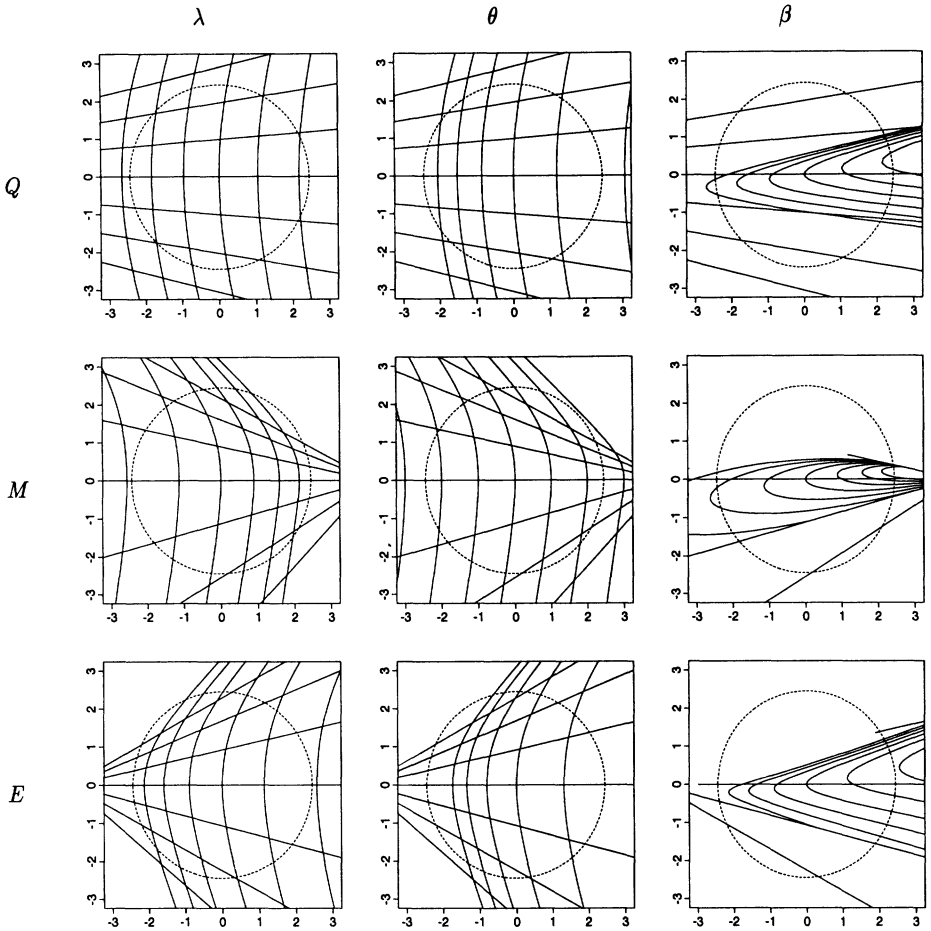

FIG. 2. *Coordinate curves for the expectation surface projected onto the tangent plane for Example* 1. *The curves are shown for the three parameterizations and choices of* ζ *of interest. The coordinates have been linearly tranformed so that their orthogonality with respect to the information metric appears in the plot as orthogonality in the usual Euclidean sense. The circle encloses the* 95% *region based on the linear approximation to the expectation surface.*

There are a few further observations we can make about this example. First, concerning the diagnostics of Section 2, although we have emphasized the quadratic loglikelihood-based curvatures, which may be viewed as having been computed using the $\alpha = \frac{1}{3}$ connection, we also evaluated the skewness curvatures based on the $\alpha = -\frac{1}{3}$ connection and found them to have values similiar to ω_Q and $\overline{\omega}_Q$. Second, the equality of certain elements of Table 2 is not coincidental: analytical calculations show that the $m\overline{\gamma}_E = \gamma_E = m\overline{\gamma}_M = \gamma_M$ and $m\overline{\gamma}_Q = \gamma_Q$ as a consequence of the partial linearity of the exponential model $\mu_i = \theta_1 \exp(-\theta_2 x_i)$. Finally, it is of interest to notice that the "intrinsic" curvatures γ_X (= $\overline{\gamma}_X$) are small. This is consistent with the general experience in

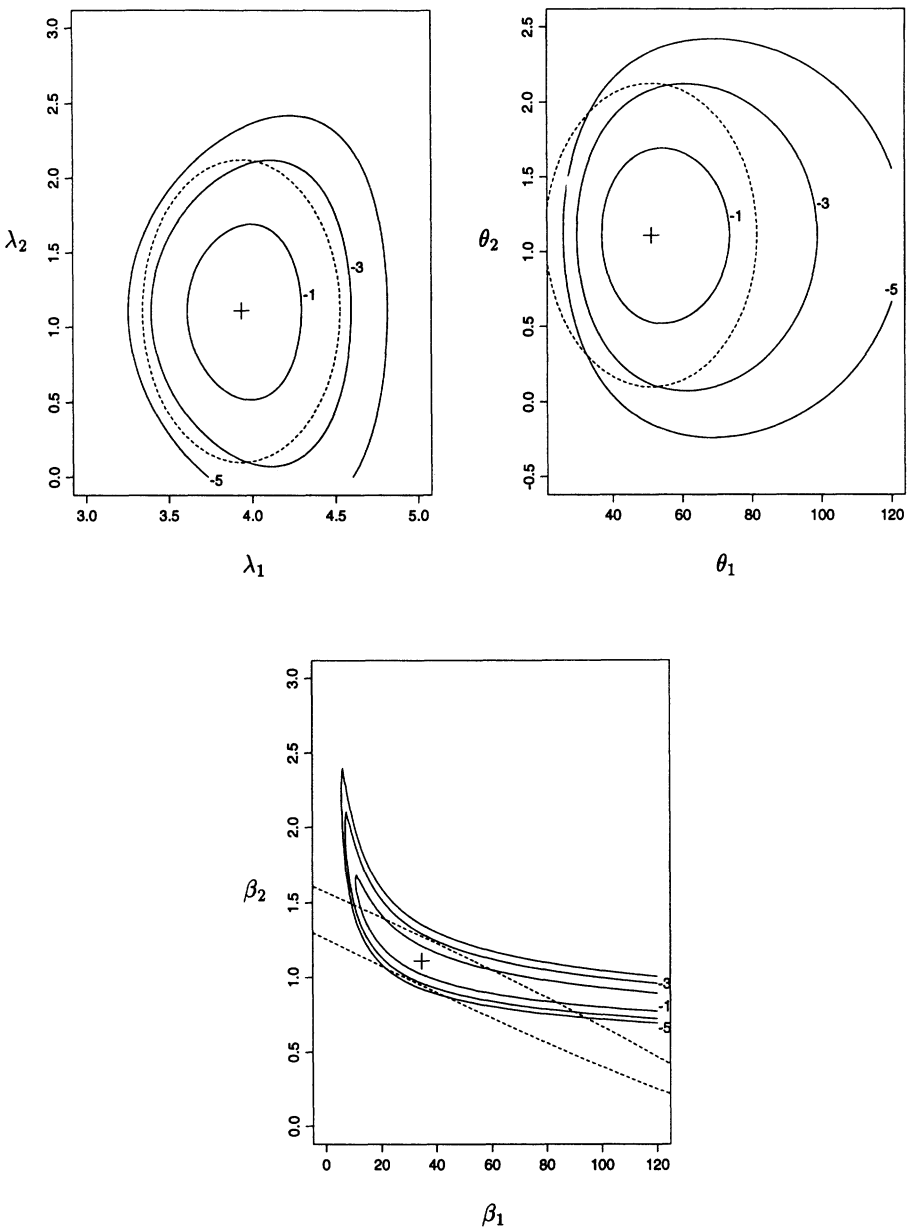

FIG. 3. *Loglikelihood contours for the three parameterizations studied in Example* 1. *The contours are labelled according to their difference from the maximum, the MLE's are marked by "+" and the dotted line is approximate* 95% *region.*

TABLE 2
Curvature diagnostics for Example 1

	Parameterization		
Diagnostic	λ	θ	β
$m^2\bar{\omega}_Q^2$	0.0261	0.1634	145.4948
ω_Q^2	0.0261	0.1242	143.8304
$m^2\bar{\gamma}_Q^2$	0.0077	0.0077	0.0077
γ_Q^2	0.0077	0.0077	0.0077
$m^2\bar{\omega}_M^2$	0.2353	0.0589	130.3106
ω_M^2	0.2353	0.1765	136.3428
$m^2\bar{\gamma}_M^2$	0.0695	0.0695	0.0695
γ_M^2	0.0695	0.0695	0.0695
$m^2\bar{\omega}_E^2$	0.2353	0.5295	153.4008
ω_E^2	0.2353	0.4118	147.8880
$m^2\bar{\gamma}_E^2$	0.0695	0.0695	0.0695
γ_E^2	0.0695	0.0695	0.0695
$\omega_{Q,\mathrm{RMS}}^2$	0.0098	0.0515	54.1445
$\omega_{M,\mathrm{RMS}}^2$	0.0883	0.0515	50.3745
$\omega_{E,\mathrm{RMS}}^2$	0.0883	0.1691	56.1471
$\gamma_{Q,\mathrm{RMS}}^2$	0.0029	0.0029	0.0029
$\gamma_{M,\mathrm{RMS}}^2$	0.0261	0.0261	0.0261
$\gamma_{E,\mathrm{RMS}}^2$	0.0261	0.0261	0.0261
$m^2\bar{B}^2$	0.2213	1.4332	174.9869
B^2	0.2140	1.0964	458.4332

nonlinear regression that parameter effects tend to be more worrisome than "intrinsic" nonlinearity.

EXAMPLE 2. We consider now a nonlinear binary response model taken from McCullagh and Nelder [(1989), page 384]. The response is the number of grasshoppers killed when exposed to various dosages of the insecticide carbofuran and the synergist piperonal butoxide, which enhances the toxicity of the insecticide. Out of a sample of size m_i, the number of grasshoppers killed under the dosages x_{1i} of insecticide and x_{2i} of synergist is modeled as binomial with the probability that a grasshopper is killed π_i satisfying

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 \log(x_{1i} - \theta) + \frac{\beta_2 x_{2i}}{\delta + x_{2i}},$$

independently for $i = 1, \ldots, 15$. The parameter θ represents a threshold value for the insecticide.

In addition to the original parameterization $\rho_1 = (\alpha, \beta_1, \beta_2, \delta, \theta)$ we investi-

TABLE 3
Generalized curvature measures and approximate third-derivative summaries for Example 2

	Parameterization		
Diagnostic	ρ_1	ρ_2	ρ_3
$m^2\overline{\omega}_Q^2$	0.198	0.094	0.198
ω_Q^2	0.282	0.143	0.287
$\omega_{Q,\mathrm{RMS}}^2$	0.0218	0.0109	0.0221
$m^2\overline{\gamma}_Q^2$		0.00354	
γ_Q^2		0.00473	
$\gamma_{Q,\mathrm{RMS}}^2$		0.00037	
$4R^*$	1.682	0.821	1.653

gate two alternatives $\rho_2 = (\alpha, \beta_1, \phi_1, \phi_2, \theta)$ and $\rho_3 = (\alpha, \beta_1, \gamma_1, \gamma_2, \theta)$, where $\phi_1 = \beta_2 x_2^*/(\delta + x_2^*)$ and $\phi_2 = \beta_2 x_2^\circ/(\delta + x_2^\circ)$, $\gamma_1 = \beta_2^{-1}$ and $\gamma_2 = \delta\beta_2^{-1}$. The values of x_2^* and x_2° are design-dependent and are here taken to be 3.9 and 19.5, respectively. The parameterizations ρ_2 and ρ_3 only transform β_2 and δ of ρ_1, which appear exclusively in the last summand of the expression for η. The parameters ϕ_1 and ϕ_2 have the form of "expected value" or "stable" parameters recommended by Ross (1970) [treating $\beta_2 x_{2i}/(\delta + x_{2i})$ as an expectation function], while γ_1 and γ_2 were chosen so that the parameters of this abbreviated expectation function both appear in the denominator. We note that the MLE of ρ_1 is $\hat{\rho}_1 = (-2.90, 1.35, 1.71, 2.06, 1.67)$, with approximate standard deviations of $(0.25, 0.10, 0.18, 1.09, 0.11)$.

To compute the curvature measures ω_Q and $\overline{\omega}_Q$, we use the product parameterization ζ defined from the quadratic loglikelihood parameterization τ. For the binomial model, τ is defined in terms of the success probability π by the incomplete beta function $\tau = \int_0^\pi [u(1-u)]^{-2/3}\, du$. The resulting curvatures are given in Table 3. The intrinsic curvatures (γ's) are once again quite small. The parameterization ρ_2 yields markedly smaller parameter-effects curvatures (ω's) than ρ_1, while ρ_3 offers no improvement at all.

We also calculate the approximate third-derivative summary $4R^*$ assuming the prior of Jeffreys' general rule, which has the form $p(\rho) \propto |i(\rho)|^{1/2}$, where $i(\rho)$ is the Fisher information matrix for ρ. Computation of this prior is straightforward for generalized linear models [see Ibrahim and Laud (1991) for discussion]; it was computed exactly for ρ_1 and approximately for the other parameterizations, using the numerical derivative capabilities in LISP-STAT. The $4R^*$ measures, also given in Table 3, support the conclusion that ρ_2 is better than ρ_1 and is probably adequate according to our rough guideline. Of course, its interpretation is less immediate.

5. Discussion. Our purpose here was to construct diagnostics that would indicate poor performance of large-sample normal approximations. The diagnostics we sought had to be easily calculated so they could be computed on a

routine basis, for commonly used models of modest dimensionality, within an interactive statistical computing environment. The work in Section 2 provides a direct extension of nonlinear regression methodology to exponential family nonlinear models. On the other hand, in Section 3 we defined diagnostics that arise from a pure likelihood or Bayesian point of view. Particularly with our emphasis on the quadratic-loglikelihood parameterization in Section 2, we do not see the approaches in Sections 2 and 3 as incompatible. Indeed, we would expect the methods of both sections to be of use regardless of whether a frequentist or Bayesian inferential framework were adopted.

Our emphasis in Section 2 on diagnostics based on the quadratic-loglikelihood parameterization of exponential families is supported by work of Slate (1994), which shows the quadratic-loglikelihood parameterization to be very effective for NEF-QVF families (as was illustrated in our Figure 1). Furthermore, the close analogy of the methods for exponential family nonlinear models with those for nonlinear regression [simply substituting the A_ζ array (2.17) for the A array (2.4)] allows extension of various computational and methodological enhancements. [It appears straightforward, for instance, to extend the methods of Bates and Watts (1981) and Cook and Goldberg (1986).] On the other hand, the foundation supplied in Section 2.3 shows that our approach is actually much more general.

An important problem not discussed here is that of examining parameter subsets, that is, examining breakdown of normal approximations for inferences about a vector component θ_1, where $\theta = (\theta_1, \theta_2)$. One possibility would be to extend the methods of Cook and Goldberg (1986). It is, however, easy to solve this problem from a Bayesian point of view by modifying the quantity R^* of Section 3 to assess the marginal posterior distribution of θ_1: we simply partition $\widetilde{G}^{-1}$ and then use $R_1^* = (\widetilde{\theta}_1 - \overline{\theta}_1^*)^T(\widetilde{G}^{-1})_{11}^{-1}(\widetilde{\theta}_1 - \overline{\theta}_1^*)$. The case of one parameter of interest [or more generally a function $g(\theta)$ of interest, such as a survival probability in Example 1] was treated by Kass, Tierney and Kadane (1989) using a one-dimensional version of the quantity R^*, and also using an approximate Pearson skewness. Kass and Slate (1992) apply a formula for posterior tail probabilities due to DiCiccio, Field and Fraser (1990) as an additional diagnostic. Examples in these references [and in Cook and Tsai (1990)] show the importance of being able to examine parameter subsets. Indeed, further analysis of Example 2 in Section 4 indicates that much of the nonnormality in the parameterization ρ_2 comes from skewness in the component θ. Transformation from θ to θ^2 reduces the squared curvature measures by nearly 50%: we obtain $m^2\overline{\omega}_Q^2 = 0.047, \omega^2 = 0.08$ and $\omega_{Q,\mathrm{RMS}}^2 = 0.0059$, while $4R^* = 0.535$.

Finally, we note two related outstanding problems. First, it would be useful to characterize parameterizations that tend to be preferable to alternatives and situations in which a particular parameterization would tend to work well. In particular, it would be good to know what sample sizes are needed for valid normal approximations in standard models. We believe these could be determined using the work of Slate (1994). Second, it would be extremely valuable to have general procedures that produce useful reparameterization. One idea due to

Ross (1970) was illustrated in our Example 2, and others may be found in Hills and Smith (1992) and the reply to discussants by Kass and Slate (1994), but the complexity of miltidimensional distributions makes this problem difficult.

Acknowledgments. We are grateful to Paul Vos, a referee and an Associate Editor for many useful comments.

REFERENCES

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

AMARI, S.-I. (1985). *Differential-Geometric 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, CA.

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

BATES, D. M. and WATTS, D. G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares. *Ann. Statist.* **9** 1152–1167.

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

COOK, R. D. and GOLDBERG, M. L. (1986). Curvatures for parameter subsets in nonlinear regression. *Ann. Statist.* **14** 1399–1418.

COOK, R. D. and TSAI, C. L. (1990). Diagnostics for assessing the accuracy of normal approximations in exponential family nonlinear models. *J. Amer. Statist. Assoc.* **85** 770–777.

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

DICICCIO, T. J. (1984). On parameter transformations and interval estimation. *Biometrika* **71** 477–485.

DICICCIO, T. J., FIELD, C. A. and FRASER, D. A. S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77** 77–95.

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

FEIGL, P. and ZELEN, M. (1965). Estimation of exponential survival probabilities with concomittant information. *Biometrics* **21** 826–838.

HILLS, S. E. and SMITH, A. F. M. (1992). Parameterization issues in Bayesian inference (with discussion). In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting* (J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith, eds.) 227–246. Oxford Univ. Press.

HODGES, J. (1987). Assessing the accuracy of normal approximations. *J. Amer. Statist. Assoc.* **82** 149–154.

HOUGAARD, P. (1982). Parameterizations of non-linear models. *J. Roy. Statist. Soc. Ser. B* **44** 244–252.

IBRAHIM, J. G. and LAUD, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *J. Amer. Statist. Assoc.* **86** 981–986.

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

JORGENSEN, B. (1987). Small dispersion asymptotics. *The Brazilian Journal of Probability and Statistics* **1** 59–90.

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

KASS, R. E. (1989). The geometry of asymptotic inference (with discussion). *Statist. Sci.* **4** 188–234.

KASS, R. E. and SLATE, E. H. (1992). Reparameterization and diagnostics of posterior non-

normality. In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting* (J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith, eds.) 289–305. Oxford Univ. Press.

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76** 663–674.

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *Essays in Honor of George Barnard* (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.) 473–488. North-Holland, Amsterdam.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, New York.

RATKOWSKY, D. A. (1983) *Nonlinear Regression Modeling.* Dekker, New York.

RATKOWSKY, D. A. (1990). *Handbook of Nonlinear Regression Models.* Dekker, New York.

ROSS, G. J. S. (1970). The efficient use of function minimization in non-linear maximum likelihood estimation. *J. Roy. Statist. Soc. Ser. C* **19** 205–221.

SEBER, G. A. F and WILD, C. J. (1989). *Nonlinear Regression.* Wiley, New York.

SLATE, E. H. (1991). Reparameterizing statistical models, Ph.D. dissertation, Dept. Statistics, Carnegie Mellon Univ.

SLATE, E. H. (1994). Parameterizations for natural exponential families with quadratic variance functions. *J. Amer. Statist. Assoc.* To appear.

SPROTT, D. A. (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika* **60** 457–465.

TIBSHIRANI, R. and WASSERMAN, L. (1994). Some aspects of the reparameterization of statistical models. *Canad. J. Statist.* **22** 163–173.

TIERNEY, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics.* Wiley, New York.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84** 710–716.

WILSON, E. B. and HILFERTY, M. M. (1931). The distribution of chi-square. *Proc. Nat. Acad. Sci. U.S.A.* **17** 684–688.

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURG, PENNSYLVANIA 15213

SCHOOL OF ORIE
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853