Ann. Inst. Statist. Math.
Vol. 47, No. 1, 49–64 (1995)

QUASI-LIKELIHOOD OR EXTENDED QUASI-LIKELIHOOD? AN INFORMATION-GEOMETRIC APPROACH

PAUL W. VOS

*Department of Biostatistics, East Carolina University,
Greenville, North Carolina 27858, U.S.A.*

(Received November 30, 1992; revised May 24, 1994)

Abstract. When the variance is a known function of the mean, as in quasi-likelihood applications, the sample variance also contains information about the mean and extensions of quasi-likelihood functions have been suggested that incorporate this additional information. In order to be sure these extensions are an improvement, further assumptions are made typically on the higher moments of the data so that there is a trade-off between the greater robustness of the quasi-likelihood estimates and the potentially improved estimates based on the extended quasi-likelihood functions. Improvement is often measured by relative efficiency but more insight can be gained by considering optimality of estimating functions, information loss, and sufficiency. All these measures can be described using the dual geometries of the quasi- and extended quasi-likelihood estimators. For a substantial range of models, the extended estimates offer little improvement when the coefficient of variation is small.

Key words and phrases: Information, sufficiency, efficiency, extended quasi-likelihood, generalized linear model, dual geometries, angle, curvature.

1. Introduction

We compare estimation using a quasi-likelihood (ql) function (Wedderburn (1974), McCullagh (1983)) to estimation using an extended quasi-likelihood (eql) function (Firth (1987), Crowder (1987), Godambe and Thompson (1989)) in terms of their geometric properties, certain curvatures and angles, which measure how well the quasi-likelihood estimators summarize the information contained in the data. Efron (1982) discusses this distinction for the maximum likelihood statistic which is both an optimal *estimator* and a superior *summary* of the data. When the extended quasi-likelihood is the log likelihood of an exponential family, the geometry is related to the ideas of Fisher information and sufficiency. We suggest extensions of these terms to quasi-likelihood and extended quasi-likelihood functions.

We shall use the following quasi-likelihood model for observations Y_{ij} where

$i = 1, \ldots, N$ and $j = 1, \ldots, n_i$:

$$\text{(1.1)} \qquad \mathrm{E}(Y_{ij}) = \mu_i, \qquad \mathrm{Var}(Y_{ij}) = \phi V(\mu_i),$$

$$\text{(1.2)} \qquad \mu_i = h(\beta; X_i)$$

where β is a p dimensional vector and X_i is a vector of k covariates. Often, the argument of the function h is $X_i'\beta$ but we shall not require this. In (1.1) we have assumed there are replications since this will be useful in motivating our discussion. In Appendix 1 we show what changes are required for unreplicated data. Notice that (1.1) defines a quasi-likelihood, we include (1.2) since this is commonly how quasi-likelihoods are used and because it illustrates how information agruments can be used when asymptotic methods are not applicable.

The quasi-likelihood estimates for β are functions of the data only through the vector of sample means $(\bar{Y}_1, \ldots, \bar{Y}_N)'$ where $\bar{Y}_i = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}$. When the quasi-likelihood corresponds to the log likelihood of a one parameter exponential family this vector is sufficient for $\mu = (\mu_1, \ldots, \mu_N)'$.

The extended quasi-likelihood allows for 'sufficient' statistics beyond $(\bar{Y}_1, \ldots, \bar{Y}_N)'$; for example, $(\bar{Y}_1, \ldots, \bar{Y}_N, S_1^2, \ldots, S_N^2)'$ where $S_i^2 = (n_i-1)^{-1}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$. Since S_i^2 is an unbiased estimator for $\sigma_i^2 = \phi V(\mu_i)$, a function of μ_i, it is reasonable to suspect that S_i^2 might contain information or evidence for μ_i. The eql estimator is a function of $(\bar{Y}_1, \ldots, \bar{Y}_N, S_1^2, \ldots, S_N^2)'$. In order to choose this function so that the eql estimator is an improvement of the ql estimator, assumptions beyond the second moment are made. Hence, there is a tradeoff between the greater robustness of the ql estimators and the potentially better eql estimators. We shall see that in many cases eql estimators offer little improvement. In particular, when the coefficient of variation c is small (say, $c < 0.5$) the two estimates and their estimated standard errors are very similar as are other inferential procedures based on the ql and eql.

The next section states the main results for the geometric properties of the estimators. The remaining sections relate the geometric features to statistical properties. Proofs and technical details appear in Appendix 2.

2. Geometric results

Before giving the main results we emphasize that we are comparing the estimators for β by considering how well $(\bar{Y}_1, \ldots, \bar{Y}_N)'$ summarizes the evidence or information in the data compared to $(\bar{Y}_1, \ldots, \bar{Y}_N, S_1^2, \ldots, S_N^2)'$. We are, in effect, comparing the ql and eql estimators for the mean vector $\mu = (\mu_1, \ldots, \mu_N)'$ rather than β. If the vector of sample means is a poor summary for μ, it may still do an adequate job for β—this will depend on h and X. The advantage of considering μ is that it can be done componentwise. That is, $\bar{Y}_i$ and $(\bar{Y}_i, S_i^2)$ can be compared as summaries for μ_i.

Since we study the ql and eql estimators for β in terms of the components for μ, we fix a single point of replication and for notational economy drop the i. So, for the rest of this paper, μ is a real number, the mean of the random sample $Y_1, \ldots, Y_n$ where n was previously n_i.

Let $Y_1, \dots, Y_n$ be a random sample from a distribution where for $j = 1, \dots, n$

$$(2.1) \qquad \mu = \mathrm{E}(Y_j), \qquad \sigma^2 = \mathrm{Var}(Y_j), \qquad \gamma_1 = \frac{\kappa_3(Y_j)}{\sigma^3}, \qquad \gamma_2 = \frac{\kappa_4(Y_j)}{\sigma^4}$$

and

$$(2.2) \qquad \sigma^2 = \phi V(\mu)$$

where κ_3 and κ_4 are the third and fourth cumulants. We assume ϕ, γ_1 and γ_2 are known and that μ is the single parameter of interest. The ql estimate is the root of the estimating equation $(\bar{y} - \mu)/(\phi V(\mu)) = 0$. That is, the ql estimate $\hat{\mu}$ is simply $\bar{y}$.

There are several ways to derive estimating equations from (2.1) and (2.2); we shall do so by considering the ordinary quasi-likelihood function for the two dimensional statistic $(\bar{Y}, S^2)$. That is, the quasi-likelihood with mean vector and covariance matrix given by

$$(2.3) \qquad \begin{aligned} \xi &= \mathrm{E}\begin{pmatrix} \bar{Y} \\ S^2 \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}; \\ \Sigma &= n\,\mathrm{Cov}\begin{pmatrix} \bar{Y} \\ S^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \gamma_1\sigma \\ \gamma_1\sigma & \sigma^2\left(\gamma_2 + 2\dfrac{n}{n-1}\right) \end{pmatrix}. \end{aligned}$$

The covariance structure of $(\bar{Y}, S^2)'$ and its relationship to μ is given by (2.1) and (2.2). The estimating equations are obtained from the fact that the residual vector for the (maximum) ql estimator is orthogonal to the derivative of the expectation vector in the inner product defined by the inverse of the covariance matrix; i.e., the eql for μ is a root of

$$(2.4) \qquad (\bar{y} - \mu, s^2 - \sigma^2(\mu))\Sigma^{-1}(\mu)\dot{\xi}(\mu) = 0$$

where $\xi = (\mu, \sigma^2)'$, $\dot{\xi} = d\xi/d\mu$, and $\dot{\xi}$ and Σ are functions of μ. Notice $\hat{\hat{\mu}}$ is the ordinary ql estimate based on $(\bar{Y}, S^2)$ but we shall call $\hat{\hat{\mu}}$ the eql to distinguish it from the ql estimate $\hat{\mu}$ based on $\bar{Y}$ alone.

We shall compare the quasi-likelihood estimator $\hat{\mu} = \bar{Y}$ and the extended quasi-likelihood estimator $\hat{\hat{\mu}} = \hat{\hat{\mu}}(\bar{Y}, S^2)$ geometrically. We begin with the subset of the real plane $\Xi = \{(\xi_1, \xi_2)' : \xi_1 \in \Xi_1, \xi_2 \in \Xi_2\}$ where Ξ_1 is the space of allowable means (for the ql function) and Ξ_2 is the space of allowable variances. Notice that each point in $\xi \in \Xi$ represents the distribution (or class of distributions) having mean equal to the first component ξ_1 and variance equal to the second component ξ_2. The quasi-likelihood assumption that the variance is a function of the mean is represented by the submanifold $\mathcal{M} = \{(\xi_1, \xi_2)' : \xi_2 = \phi V(\xi_1)\}$ and the quasi-likelihood estimator $\hat{\mu}$ by the auxiliary submanifolds $\hat{A}_\mu = \{(\xi_1, \xi_2)' : \xi_1 = \hat{\mu}\}$. See Fig. 1. Since $\hat{\mu} = \bar{y}$ does not depend on $\xi_2 = s^2$, A_μ is a 'vertical' manifold for each μ. In the same way, the extended quasi-likelihood estimator is represented

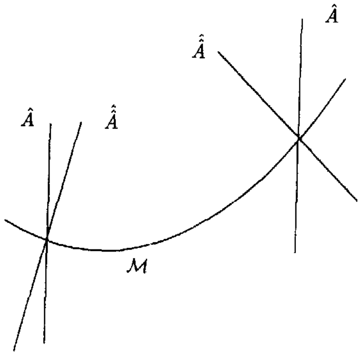

Fig. 1. Ql and extended ql estimation.

by its auxiliary manifold $\hat{\hat{A}}_\mu = \{(\xi_1, \xi_2)' : \hat{\hat{\mu}}(\xi_1, \xi_2) = \mu\}$. From the extended quasi-likelihood estimating equations (2.4) we see that

$$\hat{\hat{A}}_\mu = \{\xi \in \Xi : (\xi_1 - \mu, \xi_2 - \sigma^2(\mu))\Sigma^{-1}\dot{\xi} = 0\}$$

where Σ^{-1} and $\dot{\xi}$ are both evaluated at μ so that $\hat{\hat{A}}_\mu$ is linear in ξ.

The geometric comparison of the quasi-likelihood and extended quasi-likelihood estimators is summarized in the following two theorems.

THEOREM 2.1. *Suppose $Y_1, \ldots, Y_n$ is a random sample from a distribution having moments specified by* (2.1) *and let $\hat{A}_\mu$ and $\hat{\hat{A}}_\mu$ be the auxiliary submanifolds for $\hat{\mu}$ and $\hat{\hat{\mu}}$, respectively. Then the tangent of the angle between $\hat{A}_\mu$ and $\hat{\hat{A}}_\mu$ is*

$$m = \sqrt{\frac{(2\dot{\sigma} - \gamma_1)^2}{(\gamma_2 - \gamma_1^2) + 2\dfrac{n}{n-1}}} \tag{2.5}$$

where $\dot{\sigma} = d\sigma/d\mu$.

Under the following extended quasi-likelihood structure

$$\mu = \mathrm{E}(Y_j), \qquad \sigma^2 = \mathrm{Var}(Y_j), \qquad \gamma_1 = k_1 c, \qquad \gamma_2 = k_2 c^2, \tag{2.6}$$

$$\sigma^2 = \phi\mu^d \tag{2.7}$$

where $c = \sigma/\mu$ is the coefficient of variation, we have the following result.

THEOREM 2.2. *Let $Y_1, \ldots, Y_n$ be a random sample from the quasi-likelihood specified by* (2.6) *and* (2.7). *Then the tangent of the angle between $\hat{A}_\mu$ and $\hat{\hat{A}}_\mu$ is*

$$m = c \cdot \sqrt{1/2}|d - k_1| \frac{1}{\sqrt{\dfrac{n}{n-1} + \dfrac{1}{2}c^2(k_2 - k_1^2)}} \tag{2.8}$$

and the statistical curvature of $\mathcal{M}$ is

$$\gamma = \frac{c^2}{\sqrt{n}} \cdot \sqrt{1/2}|d - k_1| \sqrt{\frac{\left\{\frac{n}{n-1} + \frac{1}{2}c^2(d-1)(k_2 - k_1 d)\right\}^2}{\left\{\frac{n}{n-1} + \frac{1}{2}c^2[(d-k_1)^2 + (k_2 - k_1^2)]\right\}^3}}. \tag{2.9}$$

Notice that the term under the radical in (2.8) and (2.9) is approximately $\frac{n}{n-1}$ and 1, respectively, for small c. Although the angle and curvature depend on d, k_1 and k_2, in general, when the coefficient of variation c is small so are the curvature and angle.

One parameter quasi-likelihood functions share important properties—such as one dimensional 'sufficient' statistics—with one parameter exponential families and we shall see that extended quasi-likelihood functions have similar properties with respect to two parameter exponential families.

3. Geometry and curved exponential subfamilies

To relate the geometric results given in the previous section to statistical ideas we consider one parameter (i.e., μ) quasi-likelihoods as subfamilies of several two parameter exponential families where the relationship between geometric quantities (such as angle and curvature) and statistical quantities (efficiency and sufficiency) is well understood (see e.g., Efron (1975), Amari (1985), and Kass (1989)). For each subfamily, we discuss the adequacy of $\hat{\mu} = \bar{Y}$ as an estimator in terms of efficiency and as an information summarizer in terms of sufficiency.

An estimator $\hat{\theta}$ is *efficient* if its variance attains the Cramér-Rao lower bound $CR = (nI_\theta)^{-1}$ where $I_\theta = -E(\partial_\theta^2 \ell)$ and ℓ is the log likelihood obtained from a single observation. Departures from fully efficient estimators can be measured with the efficiency which is simply the ratio of CR over the variance. In many cases the variance is difficult to calculate exactly and the (first order) *asymptotic efficiency* is defined by replacing the exact variance with the asymptotic variance; we need not do this here, since the ql estimator for μ is simply the sample mean. A statistic T is *sufficient* if the log likelihood can be recovered up to an additive constant from a function depending on the parameter and T alone. That is, there is a function h depending on the data w only through T and a K which is not a function of the parameter such that

$$\ell(\theta; w) = h(\theta; T(w)) + K(w). \tag{3.1}$$

The 1-imbedding curvature γ can be used to measure departures from sufficiency for reasons given following Summary 1.

We show that the ql estimator has efficiency near one and is nearly sufficient when the coefficient of variation c is small. The results for the three exponential families are similar so we present them together.

Summary 1. Let $Y_1, \ldots, Y_n$ be a random sample from either the normal, inverse Gaussian, or gamma distributions such that the variance σ^2 is a known function of the mean μ. Then the efficiency of $\hat{\mu} = \bar{Y}$ is given by

$$\text{Eff}(\hat{\mu}) = \frac{1}{1 + k(k_1 c - 2\dot{\sigma})^2}. \tag{3.2}$$

In (3.2), $k_1 = 0, 2, 3$, for the normal, gamma, and inverse Gaussian distributions, respectively, and $k = 1/2$ for the normal and inverse Gaussian, while for the gamma distribution $k = k(c) = c^{-4}G'(c^{-2}) - c^{-2}$ is a function of the coefficient of variation c satisfying $k(c) > \frac{c^2}{1+c^2}$ and $k(c) < (1 + \frac{1}{1+c^2})\frac{c^2}{1+c^2} < 2\frac{c^2}{1+c^2}$. If $\sigma^2 = \phi\mu^d$ then

$$\text{Eff}(\hat{\mu}) = \frac{1}{1 + m^2} \tag{3.3}$$

where $m = \sqrt{k}c|d - k_1|$ is the tangent of the angle between the auxiliary submanifolds associated with $\hat{\mu}$ and the mle.

Note that the efficiency of $\hat{\mu}$ is near one for each of the three exponential families and for a reasonable range of d provided the coefficient of variation is small. In particular, for $0 \leq d \leq 4$ and $c < 1/9$ the efficiency is better than 90%. The relationship between the angle and efficiency described in (3.3) is a special case of the same description for asymptotic relative efficiency discussed by Kass (1989) and others.

One motivation for the geometric approach is the failure of asymptotic methods in small samples. Since this failure led Fisher (see Hinkley (1980)) to define sufficiency (and its relationship to information loss), we consider the geometric description of sufficiency. Efron (1975) noticed that γ is identically zero for all μ when $\mathcal{M}$ is an exponential family. Since the mle $\tilde{\mu}$ is a sufficient statistic for exponential families, it is reasonable to expect that $\tilde{\mu}$ is approximately sufficient when the curvature γ is small. Following Efron, we say there is a one dimensional statistic that is approximately sufficient for μ when the curvature is small (over the appropriate region).

Again, we can summarize the results for the three exponential families.

Summary 2. Let $Y_1, \ldots, Y_n$ be a random sample from a distribution in $\mathcal{M}$ lying in either the normal, inverse Gaussian, or gamma family of distributions. We assume further that $\sigma^2 = \phi\mu^d$. Then the tangent of the angle between the auxiliary submanifolds $\hat{A}_\mu$ and $\tilde{A}_\mu$ is

$$m = c\sqrt{k}|d - k_1| \tag{3.4}$$

and the statistical curvature of $\mathcal{M}$ is

$$\begin{aligned} \gamma &= \frac{c^2}{\sqrt{n}} \cdot \sqrt{k}|d - k_1| \sqrt{\frac{\{d-1\}^2}{\{1 + kc^2(d - k_1)^2\}^3}} \\ &\leq \frac{c^2}{\sqrt{n}} \frac{|d-1|}{\sqrt{3}} \qquad \text{for} \quad c^2 \leq 6 \end{aligned} \tag{3.5}$$

where k and k_1 depend on the exponential family and are given in Summary 1.

There are obvious similarities between Theorem 2.2 and Summary 2; that is, between eql functions and two parameter exponential families. In particular, m and γ both tend to zero as $c \to 0$.

Comment A. The expression for m and γ in the summary above differ from Theorem 2.2 because $(\bar{Y}, S^2)$ is not sufficient for the gamma or inverse Gaussian distribution. The expression for m differs by the factor $(\frac{n}{(n-1)} + \frac{1}{2}c^2(k_2 - k_1^2))^{-1/2}$. When c is not small, say $c = 1$, then the angle between $\hat{A}$ and $\hat{\hat{A}}$ is different from the angle between $\hat{A}$ and $\tilde{A}$. For example, for the inverse Gaussian distribution $(k_2 - k_1^2) = 6$ so that the tangent of the angle between $\hat{A}$ and $\hat{\hat{A}}$ is about half that between $\hat{A}$ and $\tilde{A}$. In other words, $\hat{\mu} = \bar{y}$ may be a reasonable estimator compared to the eql estimator $\hat{\hat{\mu}}$ where only functions of $(\sum y_i, \sum y_i^2)$ are considered but not compared to the ml estimator $\tilde{\mu}$ which in this case is a function of $(\sum y_i, \sum y_i^{-1})$. Therefore, when $(k_2 - k_1^2)$ is large it may be worthwhile considering nonquadratic extended estimators.

4. Ql-information and ql-sufficiency

In this section we discuss how the geometry of quasi-likelihood functions is related to statistical ideas. In Section 2 we saw that the extended quasi-likelihood could be viewed as a two dimensional ordinary quasi-likelihood so that this discussion also holds for eql functions.

For obtaining point estimators and establishing properties such as asymptotic normality and asymptotic relative efficiency one needs to assume little more than the functional relationship between mean and variance. In this case the quasi-likelihood function is equivalent to a special class of estimating functions. To discuss ql-information we also assume that the ql is a good approximation to the true log likelihood function. With this assumption the ql-information can be interpreted as describing the local behavior of the true log likelihood function. The ql-information is simply the expectation of the second order derivative of the ql function so that ql-information is a natural extension of Fisher information. In particular, the ql-information for the eql is

$$I_{\mu}^{(\bar{y},s^2)} = (1, 2\sigma\dot{\sigma})\Sigma^{-1}(1, 2\sigma\dot{\sigma})' \tag{4.1}$$

while the ql-information for the ql is

$$I_{\mu}^{(\bar{y})} = \frac{1}{\sigma^2}. \tag{4.2}$$

Using (2.3), (2.5), and (4.1), the relationship between these information functions is

$$I_{\mu}^{(\bar{y},s^2)} = (1 + m^2)I_{\mu}^{(\bar{y})} \tag{4.3}$$

where m is the tangent of the angle between $\hat{A}$ and $\hat{\hat{A}}$ and is a function of μ. Calculating the ql-information for β we see

$$(4.4) \qquad I_\beta^{(1)} = \sum_{i=1}^{n} I_\mu^{(\bar{y}_i)} \dot{\mu}_i \dot{\mu}_i', \qquad I_\beta^{(2)} = \sum_{i=1}^{n} I_\mu^{(\bar{y}_i, s_i^2)} \dot{\mu}_i \dot{\mu}_i'$$

where $\dot{\mu}_i' = (\partial\mu_i/\partial\beta_1, \ldots, \partial\mu_i/\partial\beta_p)$. Clearly, if m_i^2 is small for each case i, (4.3) shows $I_\mu^{(\bar{y}_i)}$ and $I_\mu^{(\bar{y}_i, s_i^2)}$ are close so that $I_\beta^{(1)}$ and $I_\beta^{(2)}$ will also be close.

Comment B. The role of $I_\mu^{(\bar{y})}$ and $I_\mu^{(\bar{y}, s^2)}$ can also be described in terms estimating functions. An unbiased estimating function $g(\mu, w)$ is a function of the data and the parameter μ such that $\mathrm{E}_\mu(g(\mu, W)) = 0$ and $0 < \mathrm{E}_\mu(\dot{g}(\mu, W)) < \infty$ for all μ where $\dot{g} = dg/d\mu$. We only consider $g(\mu, w)$'s that are linear in w. Godambe and Kale (1991) call an estimating function g^* optimal in a class of estimating equations G if

$$\mathrm{Var}(g_S^*) \le \mathrm{Var}(g_S) \qquad \text{for all} \quad g \in G$$

where $g_S = g/\mathrm{E}(\dot{g})$. The ql estimating function having minimal variance $1/I_\mu^{(\bar{y})}$ is optimal in the class of estimating functions depending on the data only through $\bar{y}$. The eql estimating function has variance $1/I_\mu^{(\bar{y}, s^2)}$ and is optimal in the larger class of estimating functions that depend on the data through both $\bar{y}$ and s^2. The optimality condition of Godambe is related to asymptotic optimality of the resulting estimator but can also be motivated without asymptotics (see Godambe and Kale (1991)). In the larger class of estimating functions the eql is optimal while the ql generally is not; however, when m is small equation (4.3) shows little is gained by allowing the estimating function to depend on s^2. Bhapkar (1972) calls $(\mathrm{Var}\, g_S)^{-1}$ the information of g. See also Bhapkar (1991).

Equation (4.4) and the comment above show the amount of information lost by the ql *statistic* and the suboptimality of the ql *estimating functions* can both be measured using the angle between the auxiliary submanifolds for the ql and eql estimators.

The geometry is also useful for measuring departures from ql-sufficiency where ql-sufficiency is defined as follows. Let $w = (w_1, \ldots, w_J)'$ be observations having quasi-likelihood function $\ell(\xi; w)$. A statistic $T(w)$ is ql-sufficient if

$$(4.5) \qquad \ell(\xi; w) = \ell_T(\xi; T) + K(w)$$

where $K(w)$ is not a function of ξ. Comparing (3.1) and (4.5) we see ql-sufficiency is an extension of ordinary sufficiency. Heuristically, a ql-sufficient statistics contains all the information (for the quasi-likelihood) since it recovers the quasi-likelihood function. The results for the curvature γ and sufficiency in exponential families also holds for ql-sufficiency. When the curvature of $\mathcal{M}$ is zero, the eql function $\ell(\xi; \bar{y}, s^2)$ with $\xi = (\mu, \sigma^2)'$ satisfies (4.5) and so $\hat{\mu}$ is ql-sufficient. When the curvature γ is small, $\ell(\xi \mid \xi_0; \bar{y}, s^2)$ can be approximated by a function that depends on the data only through $\hat{\mu}$. We discuss this in more detail in the next section.

5. Quasi-likelihood decomposition

The ql-information for the ql and eql provide local quadratic approximations to the ql and eql, respectively; that is, $I_{\mu_0}^{(\bar{y},s^2)}(\mu-\mu_0)^2$ is a quadratic approximation to twice the difference of the eql evaluated at μ and μ_0 while $I_{\mu_0}^{(\bar{y})}(\mu-\mu_0)^2$ is an approximation to twice the difference of the corresponding ql functions. Equation (4.3) shows there is a simple relationship between these approximations and we show now that there is a simple geometric relationship between the exact ql and eql functions provided γ is small.

We use the following notation:

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \bar{y} \\ s^2 \end{pmatrix} \quad \text{and} \quad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mathrm{E}(\bar{y}) \\ \mathrm{E}(s^2) \end{pmatrix}.$$

Notice that $(\bar{y}, s^2)'$ is defined on the sample for the i-th case and that when we consider all cases at once we shall write $(\bar{y}_i, s_i^2)'$. Under mild assumptions on the ql-function there exists a dual parameter $\theta = (\theta_1, \theta_2)'$ and a function $\psi(\theta)$ such that

$$\frac{\partial \psi}{\partial \theta} = \xi, \qquad \frac{\partial \xi}{\partial \theta} = \mathrm{Cov}(Z) = \Sigma$$

and

$$\ell(\theta; z) = \theta' z - \psi(\theta) + h(z). \tag{5.1}$$

In other words, z is ql-sufficient, ξ is the expectation parameter and θ is its dual. This duality exists because the eql function can be characterized as a divergence function which has the required duality properties as found in Amari ((1985), pp. 80–81). Details can be found in Vos (1992).

The eql for β is $\ell(\beta; z) = \ell(\theta(\beta); z)$ and we write

$$\ell(\beta \mid \beta_0; z) = \ell(\beta; z) - \ell(\beta_0; z).$$

Next, we assume that the relationship between mean and variance is such that $\mathcal{M} = \{\xi : \xi_2 = \phi V(\xi_1)\}$ has vanishing exponential curvature γ. In this case the eql estimator $\hat{\hat{\mu}}$ is ql-sufficient since $\ell(\beta \mid \beta_0; z)$ depends on z only through $\hat{\hat{\mu}}$. Using (5.1) we see that

$$\begin{aligned} \ell(\beta \mid \beta_0; z) - \ell(\beta \mid \beta_0; \hat{\hat{\xi}}) &= (z - \hat{\hat{\xi}})'(\theta(\beta) - \theta(\beta_0)) \\ &= 0. \end{aligned} \tag{5.2}$$

The last equality holds because $\hat{\hat{\mu}}$ is defined so that the residual $z - \hat{\hat{\xi}}$ is orthogonal to $\mathcal{M}$.

We evaluate to what extent $\hat{\mu}$ is ql-sufficient by seeing how well $\ell(\beta \mid \beta_0; z)$ can be approximated by a function of $\hat{\mu}$ alone. Replacing $\hat{\hat{\xi}}$ with $\hat{\xi} = (\hat{\mu}, \sigma^2(\hat{\mu}))'$ in (5.2) gives

$$\ell(\beta \mid \beta_0; z) - \ell(\beta \mid \beta_0; \hat{\xi}) = (z - \hat{\xi})'(\theta(\beta) - \theta(\beta_0)). \tag{5.3}$$

If we let C be the angle between $\hat{A}$ and $\mathcal{M}$ at $\hat{\mu}$, then the right hand side of (5.3) becomes

$$\cos C \cdot \|z - \hat{\xi}\| \cdot \ \|(\theta(\beta) - \theta(\beta_0)\|$$

where

$$\|z - \hat{\xi}\|^2 = (z - \hat{\xi})'\Sigma^{-1}(z - \hat{\xi}) = \Sigma^{22}(s^2 - \sigma(\hat{\mu}))^2,$$
$$\|\theta(\beta) - \theta(\beta_0)\|^2 = (\theta(\beta) - \theta(\beta_0))'\Sigma(\theta(\beta) - \theta(\beta_0)),$$

Σ^{22} is the second diagonal element of Σ^{-1}, and Σ and Σ^{-1} are evaluated at the same μ. Combining the above results together with $\cos C = m/\sqrt{1+m^2}$ we see that (5.3) becomes

$$\ell(\beta \mid \beta_0; z) - \ell(\beta \mid \beta_0; \hat{\xi}) = \frac{m}{\sqrt{1+m^2}} \cdot \|z - \hat{\xi}\| \cdot \|\theta(\beta) - \theta(\beta_0)\|. \tag{5.4}$$

The left hand side of (5.4) is the difference in the eql and ql function for the i-th sample and since this is zero when $\hat{\xi}$ is ql-sufficient departures from ql-sufficiency can be measured by (5.4). The last factor is a function of β and β_0 that tends to zero as $\beta \to \beta_0$ just like the difference in quasi-likelihood functions. The other two factors do not depend on β or β_0 and the size of these can be used to describe the magnitude of the difference between the eql and the ql. The factor $\|z - \hat{\eta}\|$ is proportional to $|s^2 - \sigma^2(\hat{\mu})|$ and describes how well the data fit the assumed relationship between mean and variance. The value of this factor will vary from sample to sample but should not be too large or else we would suspect the assumption about the variance and mean. The first factor $m/\sqrt{1+m^2}$ describes the relationship between inference based on the ql and the eql.

The decomposition in (5.4) is exact but assumes $\gamma^2 = 0$; when γ^2 is small the decomposition holds approximately. When the coefficient of variation c is small so is the angle m (except for extreme values of d or extreme kurtosis or skewness). Writing the curvature given in (2.9) in terms of c and m, we find

$$\gamma^2 = n^{-1}c^2m^2 \times \frac{\dfrac{n}{n-1} + \dfrac{1}{2}c^2(k_2 - k_1^2)}{\dfrac{n}{n-1} + \dfrac{1}{2}c^2[k_2 - k_1^2 + (d - k_1)^2]}$$
$$\times \left(\frac{\dfrac{n}{n-1} + \dfrac{1}{2}c^2(d-1)(k_2 - k_1 d)}{\dfrac{n}{n-1} + \dfrac{1}{2}c^2[k_2 - k_1^2 + (d - k_1)^2]} \right)^2 .$$

Notice that the last fraction is near one for small c and the first fraction must be less than one. Thus, for small c, $\gamma \doteq n^{-1}c^2m^2$ and in particular if we take $c \leq 1/2$ and $m \leq 1/2$, then

$$\gamma^2 \doteq \frac{1}{16n}$$

which is clearly less than $1/8$, the value Efron (1975) suggests that should be considered small enough to make linear approximations adequate. Following this guideline, we can use (5.4) as an approximate decomposition of the difference between the quasi-likelihoods when c is small.

6. Discussion

The dual geometries described above were introduced by Efron (1978) and developed by Amari (1985). An important feature of the dual geometries is that statistical ideas such as sufficiency and efficiency can be characterized in terms of the flatness (linearity) and orthogonality of certain manifolds. Departures from sufficiency and efficiency are then measured by departures from flatness using curvatures and from orthogonality using the angle. Amari (1985) and others exploit these useful properties for exponential families; in this paper, we have tried to do the same for the theory of quasi-likelihood but have placed a greater emphasis on the information and sufficiency aspects of the ql and eql estimators.

Dual geometries have more to offer than an interpretation of the efficiency of the ql-estimator and this is illustrated best by ql-sufficiency. If the ql and eql estimators have relative efficiency, asymptotic or non-asymptotic, near one then these estimators have nearly the same variance. That is, the ql and eql estimators have similar *distributions* but the individual estimates could be quite different. Geometrically, a first order efficient estimator has auxiliary families that are orthogonal and -1-flat at their point of intersection with $\mathcal{M}$ (Amari (1985)). The ql and eql have more structure than this: their auxiliary manifolds are -1-flat everywhere and it often happens that the 1-imbedding curvature of $\mathcal{M}$ is small. This additional structure cannot be described by relative efficiency. The ql estimate is approximately ql-sufficient when the eql can be approximated by a function depending on the data only through the ql estimate. Equation (5.4) shows that when $\gamma = 0$, then the difference between the eql and this approximation is small. Since the eql and this approximation are close, so are their maxima—the eql estimate and ql estimate, respectively. Thus, the geometry shows when the *estimates* themselves are similar not just the variance of their distributions.

The estimating equations in (2.4) are very similar to other quasi-likelihood extensions but are not exactly the same. Crowder (1987), Firth (1987), and Godambe and Thompson (1989) replace $\bar{y}$ in s^2 with μ and discuss the optimality of the resulting estimators. The optimality properties of the estimators obtained from (2.4) are those of the ordinary quasi-likelihood for the observation $(\bar{Y}, S^2)$. In this respect, (2.4) seems to be a more natural extension of the quasi-likelihood function for $\bar{Y}$ alone. We note that Crowder (Section 5) gives a more general discussion of estimating equations and (2.4) is a special case and is optimal in the sense described there.

The motivation for considering the extended quasi-likelihood expressed in (2.6) and (2.7) is that for several common two parameter exponential families we have

$$\gamma_r = \frac{\kappa_{r+2}}{\sigma^{r+2}} = k_r c^r \qquad r = 1, 2, \ldots \tag{6.1}$$

where κ_{r+2} is the $(r+2)$-th cumulant and $\sigma^{r+2} = (\sigma^2)^{(r+2)/2}$. For the normal distribution $k_r = 0$ for all r, for the gamma distribution $k_r = (r+1)!$, and for the inverse Gaussian $k_r = (2r+1) \cdot (2r-1) \cdots 1$. Equation (6.1) is also satisfied by the Poisson distribution. The reason (6.1) is satisfied by these distributions is that the variance is proportional to some power of the mean. Notice that k_1 and

k_2, like ϕ, are estimated using information from other points of replication, not just $Y_1, \ldots, Y_n$. The quasi-likelihood extension in (2.6) and (2.7) differs from other extensions (Crowder (1987), Firth (1987), and Godambe and Thompson (1989)) in that κ_3 and κ_4 are not assumed fixed nor are they assumed to vary freely from one set of replications to another. We note that Nelder (1989) predicted that higher moments would be a function of the mean for exponential quasi-likelihoods.

The size of the angle between $\hat{A}$ and $\hat{\hat{A}}$ depends on several factors but in many cases it will be small when the coefficient of variation c is small. This is consistent with the asymptotic results obtained by others. For the special case where the coefficient of variation is constant, Firth (1987) shows that for various distributions the asymptotic relative efficiency is close to one when c is small. For distributions from the log normal and inverse Gaussian families, the asymptotic relative efficiencies are greater than .90 for $c = \sqrt{.2} = .45$. McCullagh (1984) compares $\hat{\beta}_1$, the ql for β_1 and the ml estimator $\tilde{\beta}_1$ for the special case where $Y_i \sim N(\mu_i, \phi\mu_i^2)$. When the normality assumption fails to hold, McCullagh shows that the asymptotic relative efficiency tends to one as the coefficient of variation $\sqrt{\phi}$ becomes small. Lee and Nelder (1992) compare ql and extended ql estimators in terms of robustness and MSE ratios. They provide numerical results for the negative binomial and inverse Gaussian families and these show that the MSE ratios approach one as the coefficient of variation becomes small. For the inverse Gaussian, the ratios are greater than .90 for $c = \sqrt{.2} = .45$.

Appendix 1

We shall now assume there are no replications and once again use the subscript i to distinguish between observations. The geometric argument can again be used with some modification. First, the motivation now is the incorporation of the information about the third and fourth moments, not the information in the sample variance. Second, $S^2 = S_i^2$ is replaced with Y_i^2 and the information about the higher moments is used to construct the covariance matrix Σ_2 for (Y_i, Y_i^2). We drop the subscript i on the covariance matrix, but of course Σ_2 depends on i. Third, the estimating equations are obtained using Σ_2 instead of the covariance matrix for $(\bar{Y}_i, S_i^2)'$

$$(y_i - \mu_i, y_i^2 - \sigma_i^2 - \mu_i^2)\Sigma_2^{-1}\begin{pmatrix} 1 \\ 2\sigma_i\dot{\sigma}_i + 2\mu_i \end{pmatrix} = 0. \tag{A1.1}$$

Finally, we replace $\mathcal{N}$, the expectation space for $(Y_i, S_i^2)'$, with the expectation space for $(Y_i, Y_i^2)'$, call it $\mathcal{N}_1$.

From (A1.1) we see that the auxiliary submanifold for the extended estimator is a line in $\mathcal{N}_1$ and so it is reasonable to compare this estimator and Y_i using the angle between their auxiliary submanifolds. Since

$$\begin{pmatrix} y_i - \mu_i \\ y_i^2 - \sigma_i^2 - \mu_i^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2\mu_i & 1 \end{pmatrix} \begin{pmatrix} y_i - \mu_i \\ (y_i - \mu_i)^2 - \sigma_i^2 \end{pmatrix}$$

(A1.1) is equivalent to

$$(y_i - \xi_{1i}, (y_i - \mu_i)^2 - \xi_{2i})\Sigma_1^{-1}\dot{\xi}_i = 0 \tag{A1.2}$$

where $\xi_i = (\mu_i, \sigma_i^2)'$ and

$$\text{(A1.3)} \qquad \Sigma_1 = \sigma_i^2 \begin{pmatrix} 1 & \gamma_1\sigma_i \\ \gamma_1\sigma_i & \sigma_i^2(\gamma_2+2) \end{pmatrix}.$$

The estimating equations for $\beta = (\beta_1, \dots, \beta_p)'$ are obtained by replacing $\dot{\xi}_i$ with the two dimensional vector $\partial\xi_i/\partial\beta_a$ for $a = 1, \dots, p$ in (A1.2) and summing over i

$$\text{(A1.4)} \qquad \sum_{i=1}^{n} (y_i - \xi_{1i}, (y_i - \mu_i)^2 - \xi_{2i}) \Sigma_1^{-1} \frac{\partial \xi_i}{\partial \beta_a} = 0.$$

The estimating equations (A1.4) are those given in McCullagh and Nelder ((1989), Section 10.6). Comparing (A1.2) and (2.4), we see that the modifications given above result in replacing $\bar{y}$ with y_i, s^2 with $(y_i - \mu_i)^2$, and Σ with Σ_1. From (2.3) we see that Σ and Σ_1 are the same if $n/(n-1)$ and n^{-1} in Σ are set equal to 1. With these modifications the angles in Theorem 2.1 and Theorem 2.2 are the same as the ones expressed in (2.5) and (2.8) after $n/(n-1)$ is replaced by 1. The curvature in (2.9) becomes

$$\text{(A1.5)} \qquad \gamma = c^2\sqrt{1/2}|d - k_1| \sqrt{\frac{\left\{1 + \frac{1}{2}c^2(d-1)(k_2 - k_1 d) + (k_1 - d)\right\}^2}{\left\{1 + \frac{1}{2}c^2[(d-k_1)^2 + (k_2 - k_1^2)]\right\}^3}}.$$

Notice that for the normal distribution $k_1 = k_2 = 0$ and (A1.5) and (3.5) are identical as they should be.

Appendix 2

Proof of Theorem 2.1. The angle between $\hat{A}$ and $\hat{\hat{A}}$ is defined to be the angle between their tangent vectors, or, equivalently, the angle between their normal vectors. Clearly, $\dot{\xi} = (1, 2\sigma\dot{\sigma})'$ is normal to $\hat{\hat{A}}$ and an easy calculation shows $(1, \gamma_1\sigma)'$ is normal to $\hat{A}$. We use the following identity which holds for any two vectors $\vec{v}$ and $\vec{w}$

$$\text{(A2.1)} \qquad \tan^2 a = \left(\frac{\|\vec{v}\| \cdot \|\vec{w}\|}{\langle \vec{v}, \vec{w} \rangle}\right)^2 - 1$$

where a is the angle between $\vec{v}$ and $\vec{w}$. Calculating the right hand side with $\vec{v} = (1, 2\sigma\dot{\sigma})'$ and $\vec{w} = (1, \gamma_1\sigma)'$ and using the inner product defined by Σ^{-1}, we find

$$\tan^2 a = \frac{(2\dot{\sigma} - \gamma_1)^2}{(\gamma_2 - \gamma_1^2) + 2\frac{n}{n-1}}. \qquad \square$$

Proof of Theorem 2.2. Efron's statistical curvature (also called the 1-imbedding curvature Amari (1982)) is defined by

$$\gamma = \sqrt{\frac{|nM(\mu)|}{(nM^{11}(\mu))^3}} = n^{-1/2}\sqrt{\frac{|M(\mu)|}{(M^{11}(\mu))^3}} \tag{A2.2}$$

where

$$M(\mu) = \begin{pmatrix} M^{11}(\mu) & M^{12}(\mu) \\ M^{21}(\mu) & M^{22}(\mu) \end{pmatrix} = \begin{pmatrix} \dot\theta'\Sigma\dot\theta & \dot\theta'\Sigma\ddot\theta \\ \ddot\theta'\Sigma\dot\theta & \ddot\theta'\Sigma\ddot\theta \end{pmatrix} \tag{A2.3}$$

and $\ddot\theta = d\dot\theta(\mu)/d\mu$. Equations (A2.2) and (A2.3) can be found in Efron (1975).

Equation (2.8) follows from (A2.1) upon making substitutions for $\gamma_1 = k_1 c$, $\gamma_2 = k_2 c^2$, and $2\dot\sigma = dc$. Writing $\dot\theta = \Sigma^{-1}\dot\xi$ where Σ is given in (2.3) and $\dot\xi = (1, d\phi\mu^{d-1})'$ in terms of μ and differentiating we find

$$\dot\theta = D^{-1}\begin{pmatrix} A\mu^{-d} + B\mu^{2-2d} \\ C\mu^{1-2d} \end{pmatrix} \quad \text{and}$$

$$\ddot\theta = D^{-1}\begin{pmatrix} [-2+E]A\mu^{-d-1} + [-d+E]B\mu^{1-2d} \\ [-d-1+E]C\mu^{-2d} \end{pmatrix}$$

where $A = (k_2 - k_1 d)\phi^{-1}$, $B = 2\frac{n}{n-1}\phi^{-2}$, $C = (d-k_1)\phi^{-2}$, $D = (k_2 - k_1^2) + 2\frac{n}{n-1}\phi^{-1}\mu^{2-d}$, and $E = \frac{(2-d)(k_2-k_1^2)}{D}$. Notice $d/d\mu(D^{-1}) = D^{-1}(d-2+E)$. Using the fact that $|M| = |\Sigma|(\dot\theta^1\ddot\theta^2 - \dot\theta^2\ddot\theta^1)^2$ and $|\Sigma| = \phi^4\mu^{4d-2}D$, we find

$$|M| = \phi^{-4}\mu^{2-4d}D^{-3}[(d-1)(k_2-k_1 d)(d-k_1)\phi\mu^{d-2} + 2\frac{n}{n-1}(d-k_1)]^2.$$

Notice that $M^{11} = \dot\theta'\dot\xi = D^{-1}[(A + d\phi C)\mu^{-d} + B\mu^{2-2d}]$; evaluating M^{11} and substituting the above expression for $|M|$ into (A2.2) gives (2.9). □

We provide a few details for the calculations of efficiency and statistical curvature given in Section 3. We notice that the normal, gamma, and Gaussian distributions are two parameter exponential families with natural parameter $\theta = (\theta_1, \theta_2)'$. Each of these families can be parameterized by the mean μ and variance σ^2 by taking θ equal to

$$\sigma^{-2}\begin{pmatrix} \mu \\ -1/2 \end{pmatrix}, \quad \sigma^{-2}\begin{pmatrix} -\mu \\ \mu^2 \end{pmatrix}, \quad \sigma^{-2}\begin{pmatrix} -\mu/2 \\ -\mu^3/2 \end{pmatrix}$$

respectively. Since $\sigma^2 = \phi V(\mu)$ is a function of μ and ϕ is known, the variance structure of the quasi-likelihood function defines a one parameter curved exponential family. Since the details are similar, we just consider the gamma family.

Details for Summary 1. Let $Y_1, \ldots, Y_n$ be a random sample from the gamma distribution

$$f(y_1, \ldots, y_n; \theta_1, \theta_2) = \left[\prod y_i\right]^{-1} \exp\{\theta_1 u_1 + \theta_2 u_2 - n\psi(\theta_1, \theta_2)\}$$

where $u_1 = \sum y_i$, $u_2 = \sum \log(y_i)$, and

$$\psi(\theta_1, \theta_2) = -\theta_2 \log(-\theta_1) + \log \Gamma(\theta_2).$$

The Fisher information for θ is

$$nI_\theta = n \begin{pmatrix} \theta_2(-\theta_1)^{-2} & (-\theta_1)^{-1} \\ (-\theta_1)^{-1} & G'(\theta_2) \end{pmatrix} = nc^2 \begin{pmatrix} \mu^2 & \mu \\ \mu & c^{-2}G'(c^{-2}) \end{pmatrix}$$

where $G(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function and $G'(x)$ is its derivative. From this we find the Fisher information for μ, $nI_\mu = n\sigma^{-2}\{1 + 4(c^{-2}G'(c^{-2}) - 1)(1 - \dot{\sigma}/c)^2\}$, so that

$$\text{(A2.4)} \qquad \text{Eff}(\bar{Y}) = \frac{1}{1 + 4(c^{-2}G'(c^{-2}) - 1)(1 - \dot{\sigma}/c)^2}.$$

We obtain bounds for the efficiency in the gamma model by using the following identity found in Spiegel (1968)

$$\text{(A2.5)} \quad G(x) = \Gamma'(1) + \left(\frac{1}{1} - \frac{1}{x}\right) + \left(\frac{1}{2} - \frac{1}{x+1}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{x+n-1}\right) + \cdots.$$

Differentiating (A2.5) the following bounds are easily established

$$\text{(A2.6)} \qquad x^{-1} - (x+1)^{-1} \leq xG'(x) - 1 \leq x^{-1} - (x+1)^{-1} + (x+1)^{-2}.$$

Upon substitution of (A2.6) into (A2.4) we find

$$\frac{1}{1 + \left(1 + \frac{1}{1+c^2}\right)\frac{c^2}{1+c^2}(2c - 2\dot{\sigma})^2} \leq \text{Eff}(\bar{Y}) \leq \frac{1}{1 + \frac{c^2}{1+c^2}(2c - 2\dot{\sigma})^2}.$$

Details of Summary 2. Equation (3.4) follows from the relationship $\text{Eff}(\bar{Y}) = (1 + m^2)^{-1}$ and (3.3). Equation (3.5) follows from (A2.2) and (A2.3) with $\dot{\theta}$ and $\ddot{\theta}$ obtained from $\theta' = \sigma^{-2}(-\mu, \mu^2)$ and with $\Sigma = I_\theta$. The bound in (3.5) follows from the fact that

$$\frac{k(d - k_1)^2}{(1 + c^2k(d - k_1)^2)^3}$$

maximized over all values of $k(d - k_1)^2$ is less than or equal to $(3 - c^2)^2/27$ which is less than $1/3$ provided $c^2 < 6$.

References

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

Amari, S. (1985). Differential-geometrical methods in statistics, *Lecture Notes in Statist.*, **28**, Springer, New York.

Bhapkar, V. P. (1972). On a measure of efficiency of an estimating equation, *Sankyā Ser. A*, **34**, 467–472.

Bhapkar, V. P. (1991). Sufficiency, ancillarity, and information in estimating functions, *Estimating Functions*, (ed. V. P. Godambe), 239–254, Oxford Science Publications, Clarendon Press, Oxford.

Crowder, M. J. (1987). On linear and quadratic estimating functions, *Biometrika*, **74**, 591–597.

Efron, B. (1975). Defining the curvature of a statistical problem (with discussion), *Ann. Statist.*, **3**, 1189–1242.

Efron, B. (1978). The geometry of exponential families, *Ann. Statist.*, **6**, 362–376.

Efron, B. (1982). Maximum likelihood and decision theory, *Ann. Statist.*, **10**, 340–356.

Firth, D. (1987). On the efficiency of quasi-likelihood estimation, *Biometrika*, **74**, 233–245.

Fisher, R. A. (1925). Theory of statistical estimation, *Proc. Camb. Phil. Soc.*, **22**, 700–725.

Godambe, V. P. and Kale, B. K. (1991). Estimating functions: an overview, *Estimating Functions*, (ed. V. P. Godambe), 1–20, Oxford Science Publications, Clarendon Press, Oxford.

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with discussion), *J. Statist. Plann. Inference*, **22**, 137–172.

Hinkley, D. (1980). Theory of statistical estimation: the 1925 paper in R. A. Fisher: An Appreciation, *Lecture Notes in Statist.*, **1**, 85–94, Springer, New York.

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

Lee, Y. and Nelder, J. A. (1992). The robustness of estimators from quasi likelihood and quadratic estimating functions (unpublished manuscript).

McCullagh, P. (1983). Quasi-likelihood functions, *Ann. Statist.*, **11**, 59–67.

McCullagh, P. (1984). Generalized linear models, *European J. Oper. Res.*, **16**, 285–292.

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

Nelder, J. A. (1989). Comments C to Godambe and Thompson's paper, *J. Statist. Plann. Inference*, **22**, 155–158.

Spiegel, M. R. (1968). *Mathematical Handbook of Formulas and Tables*, Schaum's Outline Series, McGraw-Hill, New York.

Vos, P. W. (1992). Minimum f-divergence estimators and quasi-likelihood functions, *Ann. Inst. Statist. Math.*, **44**, 261–279.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method, *Biometrika*, **61**, 439–447.