Ann. Inst. Statist. Math.
Vol. 44, No. 2, 261–279 (1992)

MINIMUM f-DIVERGENCE ESTIMATORS AND QUASI-LIKELIHOOD FUNCTIONS*

PAUL W. VOS

Department of Mathematics, University of Oregon, Eugene, OR 97403, U.S.A.

(Received September 4, 1989; revised September 14, 1990)

Abstract. Maximum quasi-likelihood estimators have several nice asymptotic properties. We show that, in many situations, a family of estimators, called the minimum f-divergence estimators, can be defined such that each estimator has the same asymptotic properties as the maximum quasi-likelihood estimator. The family of minimum f-divergence estimators include the maximum quasi-likelihood estimators as a special case. When a quasi-likelihood is the log likelihood from some exponential family, Amari's dual geometries can be used to study the maximum likelihood estimator. A dual geometric structure can also be defined for more general quasi-likelihood functions as well as for the larger family of minimum f-divergence estimators. The relationship between the f-divergence and the quasi-likelihood function and the relationship between the f-divergence and the power divergence is discussed.

Key words and phrases: Quasi-likelihood, f-divergence, minimum divergence estimator, minimum chi-square estimator, dual geometries, generalized linear models.

1. Introduction

Generalized linear models (glms) are a natural extension of the linear regression model and the geometry used to describe glms, known as Amari's α-geometry, is an extension of the Euclidean geometry used in the classical linear model (Amari (1985), Vos (1987)). Allowing the error distribution to belong to some exponential family such as the normal, binomial, multinomial, Poisson, gamma and inverse Gaussian distributions, is one extension of glms. In the classical linear model, many important properties of the maximum likelihood estimator under the assumption of normal errors are preserved by the least squares estimator with the weaker assumption of constant variance. An analogous situation holds in glms. Wedderburn (1974) and McCullagh (1983) have shown that many important properties of the maximum likelihood estimator hold with the weaker assumption involving a functional relationship between the mean and variance. The likelihood

* This work was supported by National Science Foundation grant DMS 88-03584.

function provided by exact distributional assumptions is replaced with what is called the quasi-likelihood function. The estimator obtained from this function, the maximum quasi-likelihood (mql) estimator, has many desirable asymptotic properties. Further, just as the constant variance assumption determined the Euclidean geometric structure for the classical linear model, we show here that the quasi-likelihood function defines a pair of dual geometries. These geometries describe varying aspects of the mql estimators and play an analogous role to that of Amari's ± 1 geometries in maximum likelihood estimation. Vos (1991*a*) shows how the dual geometries can be used to define invariant measures of influence.

Under exact distributional assumptions, it is difficult to find an estimator with as many desirable properties as the maximum likelihood estimator. (Even with full distributional assumptions, not every one considers the ml estimator as the clear choice, see e.g. Berkson (1980).) Under the weaker assumptions on the first two moments of a distribution, however, there is no ml estimator and no single estimator dominates. The superiority of one estimator over an other cannot be based on higher order asymptotic calculations, such as second order efficiency (Rao (1962)), without making further distributional assumptions. We define a family of estimators, called minimum f-divergence estimators, that contain the mql estimate as a special case. These minimum f-divergence estimators have the same (first order) asymptotic properties as the ml estimator and they admit a dual geometric structure similar to that of the mql estimator. The family of f-divergences is related to the α-divergences of Amari (1985) and each defines a geometric structure using the f-connections and α-connections, respectively.

In the next section, we describe the relationship between quasi-likelihood functions and divergence measures. Before considering the minimum f-divergence estimators, we consider minimum χ^2 estimators in Section 3. The relationship between mql estimators and minimum χ^2 estimators will illustrate how to construct minimum f-divergence estimators. In Section 4, we define the minimum f-divergence estimators and describe their asymptotic properties. In Section 5, we discuss the relationship between a particular family of minimum f-divergence estimators and the minimum power divergence estimators of Read and Cressie (1988). The role of the minimum f-divergence estimators in applications is considered through an example.

2. Quasi-likelihood and divergence

Quasi-likelihood functions and divergence functions typically appear in different contexts. Amari (1985, 1987) uses divergence functions in studying the higher order asymptotic properties of estimators in exponential families, while Wedderburn (1974) and McCullagh (1983) use quasi-likelihood functions to study estimators in which distributional assumptions have been weakened to assumptions about the first two moments. Higher order asymptotic calculations are not possible without making further assumptions on the distribution represented by the quasi-likelihood function. Nevertheless, these functions are closely related. In many situations, for a given quasi-likelihood function there is a corresponding divergence function and for a given divergence there is a corresponding quasi-likelihood function. The quasi-likelihood and divergence are related in such a

manner that maximizing the quasi-likelihood is equivalent to minimizing the corresponding divergence function. This statement is made precise in the proposition and corollary. We begin with two definitions.

DEFINITION 2.1. Let $\mathcal{M}$ and $\mathcal{Y}$ be subsets of $\mathbb{R}^n$. A *quasi-likelihood* is a function $l : \mathcal{M} \times \mathcal{Y} \mapsto \mathbb{R}^1$ such that

$$\frac{\partial l(\mu; y)}{\partial \mu} = V^-(\mu)(y - \mu)$$

where $V^-(\mu)$ is nonnegative definite for all $\mu \in \mathcal{M}$.

This definition arises from McCullagh and Nelder ((1983), pp. 168, 169). McCullagh and Nelder ((1989), p. 325) motivate the quasi-likelihood in terms of a function that behaves like the score vector. As the notation suggests, the quasi-likelihood is typically used when $\mathcal{M}$ is the mean of a random variable Y, $\mathcal{Y}$ is the convex hull of the support of Y, and $V^-(\mu)$ is a generalized inverse of the variance matrix for Y. The solution of the system of differential equations given in Definition 2.1 can be written as

$$l(\mu; y) = \theta \cdot y - \psi(\theta) + K(y) \tag{2.1}$$

where $\theta = \theta(\mu)$ is a function of μ into $\mathbb{R}^n$ such that $\partial\theta/\partial\mu = V^-(\mu)$, $\partial\psi(\theta)/\partial\theta = \mu$, and $K(y)$ is not a function of μ. The image of the map θ will be denoted by $\Theta = \{\theta(\mu) : \mu \in \mathcal{M}\}$.

DEFINITION 2.2. Let $\mathcal{N}$ be an open subset of $\mathbb{R}^n$ and let $\eta_1, \eta_2 \in \mathcal{N}$. The *divergence* from η_1 to η_2, denoted by $D(\eta_1, \eta_2)$, is the function taking values in $[0, \infty]$ such that for any $\eta_1, \eta_2 \in \mathcal{N}$

(1) $D(\eta_1, \eta_2) \geq 0$ with equality holding if and only if $\eta_1 = \eta_2$

(2) The metric matrix $G(\eta) = (g_{ij}(\eta))$ is positive definite where $g_{ij}(\eta_1) = \dfrac{\partial^2}{\partial\eta_1^i \partial\eta_1^j} D(\eta_1, \eta_2)$ is a smooth function of η_1 alone.

Notice that the smoothness of D and (1) imply $\partial_{1i} D(\eta_1, \eta_2) = 0 = \partial_{2i} D(\eta_1, \eta_2)$ when $\eta_1 = \eta_2$, $\partial_{1i} = \partial/\partial\eta_1^i$ and $\partial_{2i} = \partial/\partial\eta_2^i$. For a given $G(\eta)$, the divergence is defined uniquely on $\mathcal{N}$. Amari (1985) defines the divergence on a Riemannian manifold S. We define the divergence on $\mathcal{N}$ to avoid the introduction of differential geometric concepts. In order to make our definition agree with Amari's we shall also call $\bar{D}(\xi_1, \xi_2) = D(h^{-1}(\xi_1), h^{-1}(\xi_2))$ a divergence when h is a diffeomorphism on $\mathcal{N}$. Notice that the ξ parameterization will generally not satisfy the three properties of a divergence. To distinguish between these parameters we call η the divergence parameter and reserve the notation D for the divergence expressed in this parameterization and use $\bar{D}$ for other parameterizations.

We note that other authors have defined divergence differently than Definition 2.2. Many definitions do not require G to be a function of η_1 alone. Both

definitions are closely related to the terms contrast functional and yoke. Further discussion can be found in Eguchi (1983, 1985), Barndorff-Nielsen (1987) and Rao (1987).

The relationship between quasi-likelihoods and divergence measures is given in the following proposition.

PROPOSITION 2.1. *If $D(\eta_1, \eta_2)$ is a divergence on $\mathcal{N}$ with metric matrix $G(\eta)$, then we can define a quasi-likelihood such that $\mu = \eta$ and $V^-(\mu) = G(\eta)$. Conversely, if $l(\mu; y)$ is a quasi-likelihood with $V^-(\mu)$ positive definite for all μ and a smooth function in μ, $\mathcal{M} \subset \mathcal{Y}$, and $\mathcal{M}$ open, then there is a divergence $D(\eta_1, \eta_2)$ with $\eta = \mu$ and metric matrix $G(\eta) = V^-(\mu)$.*

PROOF. Given the divergence $D(\eta_1, \eta_2)$, we fix $\eta_0 \in \mathcal{N}$ and define $\phi(\eta) = D(\eta, \eta_0)$ and $\theta^i(\eta) = \partial\phi(\eta)/\partial\eta^i$. Since $G(\eta) = (\partial\theta/\partial\eta)$ is full rank, we can write $\eta = (\eta^1, \ldots, \eta^n)'$ as a function of $\theta = (\theta^1, \ldots, \theta^n)'$, $\eta = \eta(\theta)$. Since $G^{-1}(\eta) = (\partial\eta/\partial\theta)$ is symmetric, $\partial\eta^i/\partial\theta^j = \partial\eta^j/\partial\theta^i$ so there exists a function $\psi(\theta)$ such that $\partial\psi(\theta)/\partial\theta^i = \eta^i$. If we take $\mathcal{M} = \mathcal{Y}$ and $\mu = \eta$, then it is easily verified that $l(\mu; y) = \theta(\eta) \cdot y - \psi(\theta(\mu))$ is a quasi-likelihood with inverse variance matrix $V^-(\mu) = G(\mu)$. Conversely, let $l(\mu; y)$ be a quasi-likelihood defined on $\mathcal{M} \times \mathcal{Y}$ and take $\eta = \mu$. Let θ be the function given in (2.1), so that $\theta(\eta) = \theta(\mu)$ is a function of η. It is now easily verified that

$$D(\eta_1, \eta_2) = l(\eta_1; \eta_1) - l(\eta_2; \eta_1) \tag{2.2}$$

is a divergence with metric matrix $G(\eta) = V^-(\eta)$.

By the uniqueness of $D(\eta_1, \eta_2)$ on $\mathcal{N}$, we know (2.2) holds for all likelihoods and divergence functions for which $\eta = \mu$ and $G(\eta) = V^-(\mu)$. The following corollary is an immediate consequence of (2.2).

COROLLARY 2.1. *Let $D(\eta_1, \eta_2)$ and $l(\mu; y)$ be the divergence and a likelihood described in Proposition 2.1. If $M \subset \mathcal{M} \subset \mathcal{Y}$, $y \in \mathcal{M}$ and $\hat{\mu} \in M$, then $l(\hat{\mu}; y) = \max_{\mu \in M} l(\mu; y)$ if and only if $D(y, \hat{\mu}) = \min_{\mu \in M} D(y, \mu)$.*

This corollary is important because the dual geometries can be used to describe when the divergence is minimized and thus when the quasi-likelihood is maximized.

Suppose now that $\mu(\beta)$ is some smooth function of the m-dimensional parameter $\beta = (\beta^1, \ldots, \beta^m)' \in \mathcal{B}$ where $\mathcal{B} = \{\beta : \mu(\beta) \in \mathcal{M}\}$; for a glm, this function takes the following form $\mu(\beta) = H^{-1}(X\beta)$ where H is a bijection with domain $\mathcal{M}$ called the link function and X is the $N \times m$ matrix of covariates. The mql estimator $\hat{\beta}$ has several desirable asymptotic properties that we now consider. First, $\hat{\beta}$ is asymptotically unbiased; that is, $E(\hat{\beta} - \beta) = o(N^{-1/2})$. Second, $\sqrt{N}(\hat{\beta} - \beta)$ is asymptotically normal with mean zero and variance matrix $\Sigma = [(\partial\mu/\partial\beta)'V^-(\partial\mu/\partial\beta)]^-$ where $(\partial\mu/\partial\beta)$ is the $N \times m$ matrix with components $\partial\mu^i/\partial\beta^a$ for $i = 1, \ldots, N$ and $a = 1, \ldots, m$. Hence, $\hat{\beta}$ is $O(\sqrt{N})$-consistent. Third, among all estimators $\tilde{\beta}$ satisfying

$$\tilde{\beta} - \beta = L_\mu(Y - \mu) + o_p(N^{-1/2}) \tag{2.3}$$

where $L_\mu = \Sigma(\partial\mu/\partial\beta)'V^-$ is an $m \times N$ matrix evaluated at μ, the mql $\hat{\beta}$ estimator has the minimum asymptotic variance. The matrix A is smaller than the matrix B, if $B - A$ is positive semi-definite. Finally, the quasi-likelihood difference statistic $2l(\hat{\beta}; Y) - 2l(\beta; Y)$ is asymptotically χ^2_m. These properties are considered in greater detail in McCullagh (1983). Each of the estimators we consider will have these four asymptotic properties.

Besides the asymptotic properties mentioned above, the mql estimator has other desirable characteristics as well. In order to calculate $\hat{\beta}$, the quasi-likelihood function is treated just as an ordinary log likelihood function from an exponential family so that in many respects the quasi-likelihood offers a natural extension of ml estimators. In particular, the estimation algorithm in both instances is a fairly simple one, consisting of iterative weighted least squares in which the weights depend only on the current values of the parameters. Furthermore, the dual geometry for log-likelihood functions from an exponential family can be extended to quasi-likelihood functions. We leave the details to the Appendix.

3. Minimum chi-square estimates

Before studying minimum divergence estimators it will be useful to consider minimum chi-square estimators and their relationship to mql estimators. In the preceding section the asymptotic properties of the mql estimator were expressed in terms of N, the length of the data vector y. In other applications of quasi-likelihoods each of N observations appears in exactly one of n cells. The data vector now becomes an n-dimensional column y where the i-th element of y equals the number of observations in the i-th cell. The quasi-likelihood most commonly used in this situation is the log likelihood of the Poisson or multinomial distribution. It can be shown that as $N \to \infty$, the mql estimator $\hat{\beta}$ is asymptotically unbiased, normal, and is optimal in a context similar to that expressed in (2.3). Furthermore, the quasi-likelihood difference statistic is again asymptotically chi-square. In Section 4, we show that these four properties hold in a class of estimators.

We begin by defining the divergence

$$D(\eta_1, \eta_2) = \frac{1}{2}\sum_1^n (\eta_1^i - \eta_2^i)^2/\eta_1^i(\eta_2^i)^2 \tag{3.1}$$

for $\eta_1, \eta_2 \in (0, \infty)^n$. $D(\eta_1, \eta_2)$ provides a divergence on multinomial distributions by taking $\eta = 1/\pi$ where π is the vector of cell probabilities and division of vectors is done componentwise. Writing D in terms of the cell probabilities we have obtained the more familiar goodness-of-fit measure

$$\bar{D}(\pi_1, \pi_2) = \sum_1^n \frac{(\pi_1^i - \pi_2^i)^2}{\pi_1^i} = \frac{1}{N}\sum_1^n \frac{(\mu_1^i - \mu_2^i)^2}{\mu_1^i}. \tag{3.2}$$

If the model provides a smooth $\mathbb{R}^n$ valued function $\pi(\beta)$ from $\mathcal{B}$, then from (3.2) we see that the minimum Pearson chi-square (χ^2_P) estimate $\tilde{\beta}$ minimizes $\bar{D}(\pi(\beta), y/N)$ and the minimum Neyman chi-square (χ^2_N) estimate $\hat{\beta}$ minimizes $\bar{D}(y/N, \pi(\beta))$.

To better understand the relationship between maximum likelihood estimators and minimum chi-square estimators we use the fact that $D(\eta_1, \eta_2)$ is the Kullback information (1968) between two inverse Gaussian random variables with means η_1 and η_2. If Z is an n-dimensional vector of independent inverse Gaussian random variables, z is the vector of realizations, and $\eta(\beta)$ is an imbedding of $\mathcal{B} \subset \mathbb{R}^m$ in the mean parameter space for Z, then the value of β that minimizes $D(z, \eta(\beta))$ is the maximum likelihood estimate which we denote by $\hat{\beta}$. On the other hand, the value of β that minimizes $D(\eta(\beta), z)$ is the dual maximum likelihood estimate $\tilde{\beta}$. Hence, the minimum χ^2_P estimate can be found by transforming y to $z = 1/y$, $\mu(\beta)$ to $\eta(\beta) = (\mu(\beta))^{-1}$, treating z as the realization of an inverse Gaussian random vector and finding the dual maximum likelihood estimate for β. Likewise, the minimum χ^2_N estimate $\hat{\beta}$ is found by making the same transformations and finding the maximum likelihood estimate for β. Clearly, the transformation $z^i = 1/y^i$ is not possible when $y^i = 0$. But when $y^i = 0$, the minimum χ^2_N estimate can not be found since then $d(\cdot, y) = \infty$. In this case, a small positive constant c is sometimes added to each observation so that y^i is replaced with $y^i + c$, for $i = 1, \ldots, n$. Now the transformation becomes $z^i = 1/(y^i + c)$ and the estimate can be obtained as before.

This relationship between minimum χ^2 estimates and ml (and dual ml) estimates for the inverse Gaussian distribution shows that the estimation algorithm used to find ml estimates (iterative reweighted least squares) can also be used to find minimum χ^2 estimates. Hence, software designed to find ml estimates, such as GLIM, can be used to find minimum χ^2 estimates as well. The main point, however, is to show that sensible estimators can be obtained using divergence measures with divergence parameter different from the mean parameterization. Certainly, not all transformations of the mean will lead to useful estimators. The reciprocal transformation used for the Poisson distribution is too strong for other situations. For example, the reciprocal of a normal random variable does not have finite mean so it is not sensible to model the variance as a function of it. To avoid such problems restrictions must be placed on the distributions and/or transformations. We shall consider these restrictions in a later section and turn to an example next.

Ship Damage Example. The maximum quasi-likelihood estimates and minimum χ^2 estimate have similar large sample properties and the algorithm for computing these estimates and their variances are both special cases of iterative weighted least squares. Which of these estimates to use will depend on several factors. To investigate the behavior of these estimates we consider an example found in McCullagh and Nelder (1983). The data for this example is given in Table 1 where y is the number of wave induced damage incidences to cargo ships. McCullagh and Nelder (1983) propose a model for the risk of damage based on the ship type, period of construction, period of operation, and the aggregate months of service. These authors fit the following model

$$\begin{aligned} (3.3) \qquad &\log(\text{expected number of damage incidences}) \\ &\quad = \beta_0 + \log(\text{aggregate months of service}) \\ &\qquad + (\text{effect due to ship type}) \end{aligned}$$

$$+ \text{(effect due to construction period)}$$
$$+ \text{(effect due to service period)}.$$

The coefficient of log(aggregate months of service) is assumed known and equal to 1 corresponding to the assumption that the number of accidents should be proportional to the length of risk. The data support this assumption. To allow for over-dispersion of the data a Poisson distribution is not assumed; the only distributional assumptions are that $\text{Var}(Y) = \sigma^2 E(Y)$ where $\sigma^2 > 1$. The dispersion estimate $\tilde{\sigma}^2$ is defined to be

$$\tilde{\sigma}^2 = \frac{1}{n-m}\sum_{i=1}^{n}\frac{(y^i - \hat{\mu}^i)^2}{\hat{\mu}^i} = \frac{N}{n-m}\sum_{i=1}^{n}\frac{(y^i/N - \hat{\pi}^i)^2}{\hat{\pi}^i}, \tag{3.4}$$

where $\hat{\mu}^i = \mu^i(\hat{\beta})$ and for this example $\tilde{\sigma}^2 = 1.69$. Notice that $n = 34$, and not 40, since there are 6 necessarily zero observations in Table 1. The mql estimate $\hat{\beta}$ and approximate standard errors are given in Table 2. The main effects model (3.3) fits the data reasonably well although the 21st observation is an outlier with standardized residual 2.87. The i-th standardized residual is defined to be $(y^i - \hat{\mu}^i)/\tilde{\sigma}\sqrt{\hat{\mu}^i}$. This residual remains high even with the inclusion of the interaction term between ship type and period of construction. McCullagh and Nelder (1983) draw the following conclusions: There is evidence for over-dispersion ($\tilde{\sigma}^2 = 1.69$), after 1974 the rate of ship damage increased by 47% with a 95% confidence interval given by (8%, 100%), ship types B and C have the lowest risk of damage while E has the highest, and the ships constructed between 1960 and 1964 appear to be the safest.

Table 2 also lists the minimum χ^2_P estimates $\tilde{\beta}$ and their approximate standard errors. The standard errors are computed using the estimate $\tilde{\sigma}^2_P$ which is defined by replacing $\hat{\mu}$ with $\tilde{\mu} = \mu(\tilde{\beta})$ in (3.4) and takes the value 1.46 in this example. Using the minimum χ^2_P estimate, there appears to be less evidence for ship type C being safer than A while there is more evidence that ship A is safer than E. The other minimum χ^2_P estimates are similar to the mql estimates. Using $\tilde{\beta}$, the risk of ship damage increased by 44% and has a 95% confidence interval (9%, 91%). Since $\tilde{\sigma}^2_P < \tilde{\sigma}^2$, the standard errors for the $\tilde{\beta}^i$'s are all slightly smaller than the corresponding standard errors for the $\hat{\beta}^i$'s. The major conclusions do not change by using $\tilde{\beta}$, although the estimate for over-dispersion is smaller ($\tilde{\sigma}^2_P = 1.46$). Since both $\tilde{\beta}$ and $\hat{\beta}$ have optimal asymptotic properties under mild distributional assumptions and each algorithm is equally easy to implement, we need some other criteria for deciding between using $\tilde{\beta}$ or $\hat{\beta}$. For this example we note two advantages that favor the minimum χ^2_P estimate. First, minimization of $\sum(y^i - \mu^i)^2/\mu^i$ is easier to motivate and interpret than the maximization of the quasi-likelihood function. Second, the minimum χ^2 estimator also improves the fit by decreasing the standardized residual for the 21st observation to 1.98.

Table 1. Ship damage data.

Ship type	Period of construction	Period of operation	Aggregate months service	Number of damage incidents
A	1960–64	1960–74	127	0
A	1960–64	1975–79	63	0
A	1965–69	1960–74	1095	3
A	1965–69	1975–79	1095	4
A	1970–74	1960–74	1512	6
A	1970–74	1975–79	3353	18
A	1975–79	1960–74	0	0*
A	1975–79	1975–79	2244	11
B	1960–64	1960–74	44882	39
B	1960–64	1975–79	17176	29
B	1965–69	1960–74	28609	58
B	1965–69	1975–79	20370	53
B	1970–74	1960–74	7064	12
B	1970–74	1975–79	13099	44
B	1975–79	1960–74	0	0*
B	1975–79	1975–79	7117	18
C	1960–64	1960–74	1179	1
C	1960–64	1975–79	552	1
C	1965–69	1960–74	781	0
C	1965–69	1975–79	676	1
C	1970–74	1960–74	783	6
C	1970–74	1975–79	1948	2
C	1975–79	1960–74	0	0*
C	1975–79	1975–79	274	1
D	1960–64	1960–74	251	0
D	1960–64	1975–79	105	0
D	1965–69	1960–74	288	0
D	1965–69	1975–79	192	0
D	1970–74	1960–74	349	2
D	1970–74	1975–79	1208	11
D	1975–79	1960–74	0	0*
D	1975–79	1975–79	2051	4
E	1960–64	1960–74	45	0
E	1960–64	1975–79	0	0*
E	1965–69	1960–74	789	7
E	1965–69	1975–79	437	7
E	1970–74	1960–74	1157	5
E	1970–74	1975–79	2161	12
E	1975–79	1960–74	0	0*
E	1975–79	1975–79	542	1

*Necessarily empty cells.

Table 2. Parameter estimates.

Parameter		$\hat{\beta}$(s.e.)	$\tilde{\beta}$(s.e.)
Intercept		−6.41	−6.37
Ship type	A	0.00	0.00
	B	−0.54 (0.23)	−0.57 (0.21)
	C	−0.69 (0.43)	−0.25 (0.34)
	D	−0.08 (0.38)	0.11 (0.33)
	E	0.33 (0.31)	0.45 (0.27)
Period of construction	1960–64	0.00	0.00
	1965–69	0.70 (0.19)	0.71 (0.18)
	1970–74	0.82 (0.22)	0.81 (0.20)
	1975–79	0.45 (0.30)	0.46 (0.28)
Period of service	1960–74	0.00	0.00
	1975–79	0.38 (0.15)	0.37 (0.14)

4. Minimum f-divergence estimators

In the example of the previous section, we found a divergence whose minimization provided an estimator with good properties. In this section we show that it is often possible to use an entire family of divergences to define optimal estimators. Some divergences together with their variance functions are listed in Table 3. The variance matrix of a divergence $D(\eta_1, \eta_2)$ is the matrix inverse of $G(\eta_1)$ and when $\eta_1 \in \mathbb{R}^1$, the variance matrix is called the variance function. If $\eta = (\eta^1, \ldots, \eta^n)'$ and the variance matrix is diagonal, then $D(\eta_1, \eta_2) = \sum_1^n D^1(\eta_1^i, \eta_2^i)$ where $D^1(\cdot, \cdot)$ is the appropriate one-dimensional divergence function. The parameter $\theta(\eta)$ is the dual parameter to η and plays an important role in the dual geometric structure. The divergence with variance function given by η^d for $d = 0$, 1, 2 and 3 is the Kullback information for the normal, Poisson, gamma and inverse Gaussian distributions, respectively, when η is the expectation parameter. The divergences with variance $\eta(1-\eta)$ and $\eta + \eta^2/k$ correspond to the Kullback information for the binomial and negative binomial distribution, respectively, when $\eta = E(Y)$. In these special cases, $\theta(\eta)$ is the natural parameter for these exponential families. Table 3 is closely related to McCullagh and Nelder's table for quasi-likelihoods (1983). This is not too surprising in the light of the relationship between quasi-likelihoods and divergence functions.

We have seen that minimum chi-square estimates can be obtained by transforming the data and finding a value for β that minimized a particular divergence between the transformed data and the transformed fitted values. Clearly, we need not restrict ourselves to the reciprocal transformation and the Kullback divergence for the inverse Gaussian distribution. Under suitable regularity conditions we can use any transformation with any divergence function. However, to ensure that our estimates are optimal, for a given transformation we must choose the appropriate divergence.

Table 3. Divergences.

Variance	$D^1(\eta_1, \eta_2)$	$\theta(\eta)$	Comments
1	$\frac{1}{2}(\eta_1 - \eta_2)^2$	η	
η	$\eta_1 \log\left(\frac{\eta_1}{\eta_2}\right) - (\eta_1 - \eta_2)$	$\log(\eta)$	$\eta > 0$
η^2	$\log\left(\frac{\eta_2}{\eta_1}\right) + \frac{\eta_1 - \eta_2}{\eta_2}$	$-\eta^{-1}$	$\eta > 0$
η^3	$\frac{1}{2}\eta_1^{-1}\eta_2^{-2}(\eta_1 - \eta_2)^2$	$-\frac{1}{2}\eta^{-2}$	$\eta > 0$
η^p	$\frac{\eta_1^{2-p} - \eta_2^{1-p}\{(2-p)\eta_1 - (1-p)\eta_2\}}{(1-p)(2-p)}$	$\frac{\eta^{1-p}}{1-p}$	$\eta > 0$; $p \neq 0, 1, 2$
$\eta(1-\eta)$	$\eta_1\left\{\log\left(\frac{\eta_1}{\eta_2}\right) - \log\left(\frac{1-\eta_1}{1-\eta_2}\right)\right\} + \log\left(\frac{1-\eta_1}{1-\eta_2}\right)$	$\log\left(\frac{\eta}{1-\eta}\right)$	$\eta > 0$

Let $l(\mu; y)$ be a quasi-likelihood function defined on $\mathcal{M} \times \mathcal{Y}$ with variance matrix $V(\mu)$. Suppose $V(\mu)$ has a nonsingular generalized inverse $V^-(\mu)$ and define $V^* = V^*(\mu) = (V^-)^{-1}$. The construction of V^* allows us to define a divergence even when V is not full rank, as is the case when Y is multinomial. Suppose further that there is a divergence function $D(\mu_1, \mu_2)$ on $\mathcal{M} \times \mathcal{M}$ with variance matrix V^*. Notice that μ is the divergence parameter for this particular divergence. Let $f(\cdot)$ be any diffeomorphism between $\mathcal{M}$ and $\mathcal{N}$ such that there exists a divergence $D_f(\eta_1, \eta_2)$ on $\mathcal{N} \times \mathcal{N}$ with variance matrix $(\partial f/\partial\mu)'V^*(\partial f/\partial\mu)$ where $\mu = f^{-1}(\eta)$ is a function of η. In terms of the original expectation parameter, we have $\bar{D}_f(\mu_1, \mu_2) = D_f(f(\mu_1), f(\mu_2))$. The divergence $\bar{D}_f(\mu_1, \mu_2)$ is called the f-divergence for $D(\mu_1, \mu_2)$, the f-divergence for $l(\mu; y)$, or simply the f-divergence. The minimum f-divergence estimate $\hat{\beta}_f$ is the value for β that minimizes $\bar{D}_f(y, \mu(\beta))$.

From the glms perspective, the minimum f-divergence estimators can be understood in the following way. We know that $\mathrm{Var}(Y) = V(\mu) + o(1)$ is some function of the expectation parameter μ. The mql estimate is found by maximizing the quasi-likelihood associated with $V(\mu)$. When f is the identity map the mql estimate and the minimum f-divergence estimate are the same. For the other minimum f-divergence estimates, we also consider the quasi-likelihood defined by the variance matrix $V_f(\eta) = (\partial f/\partial\mu)'V^*(\mu)(\partial f/\partial\mu)$ where μ is a function of η. This variance function is chosen to ensure that the minimum divergence estimators have the same asymptotic properties as the mql estimator. As in Section 3, we are assuming that each value in the data vector represents a sum of N observations. We note that the minimum f-divergence estimates are not simply estimates based on a transformation of the data. Not only is the data transformed, but the model $\{\mu(\beta) : \beta \in \mathcal{B}\}$ is also transformed. To this point, finding minimum f-divergence estimators is like using the transform both sides model discussed in Ruppert and Aldershof (1989). There is, however, an important difference. In transformation models, the transformed data is assumed to have a particular error distribution

or variance structure; often normal errors or homoscedasticity is assumed. For minimum f-estimators, the variance structure $V(\mu)$ is assumed up to first order for the untransformed data and the variance structure of the transformed data is defined in terms of $V(\mu)$ and f.

This relationship between the minimum divergence estimators and quasi-likelihood functions makes it clear that the same estimation algorithm used for mql estimates can be used for minimum divergence estimates. Mql estimates are found using an iterative weighted least squares algorithm that is Fisher's scoring algorithm when the quasi-likelihood function is the log likelihood from some exponential family (McCullagh and Nelder ((1983)), pp. 31–34). This same algorithm can be used for minimum divergence estimates provided the transformed data $z = f(y)$ and $\eta = f(\mu)$ parameterization is used. If β_0 is an initial estimate for $\hat{\beta}_f$, then the one step estimate β_1 is found from

$$(\beta_1 - \beta_0) = W^{-1}\left(\frac{\partial \eta}{\partial \beta}\right)' V_f^{-1}(z - \eta(\beta_0))$$

where

$$W = \left(\frac{\partial \eta}{\partial \beta}\right)' V_f^{-1}\left(\frac{\partial \eta}{\partial \beta}\right) = \left(\frac{\partial \mu}{\partial \beta}\right)' V^{-}\left(\frac{\partial \mu}{\partial \beta}\right)$$

and each matrix is evaluated at β_0. Notice that W is same for all transformations $f(\cdot)$; we show later that W is the asymptotic covariance matrix for $\hat{\beta}_f$.

To make our discussion more explicit, we consider the family of power transformations $f(\cdot;\lambda)$

$$f(y;\lambda) = \begin{cases} y^{\lambda} & \text{for } \lambda \neq 0 \\ \log y & \text{for } \lambda = 0. \end{cases}$$

Let $E(Y) = \mu = (\mu^1, \ldots, \mu^n)'$ with $\mu^i > 0$, for $i = 1, \ldots, n$. The family of divergence measures associated with the power family of transformation will depend on the variance matrix for Y. Let V and V^* be defined as above with the added restriction that V^* is a diagonal matrix with diagonal μ^d, the mean vector raised to the power d. For each real d, we can define the following family of divergences

$$\bar{D}_{\lambda,d}(\mu_1, \mu_2) = \sum \frac{\mu_1^{2-d} - \mu_2^{2-d} + \dfrac{2-d}{\lambda}\mu_2^{2-d}\{1 - (\mu_1/\mu_2)^{\lambda}\}}{(2-d-\lambda)(2-d)} \tag{4.1}$$

provided $\lambda \neq 0$, $d \neq 2$ and $\lambda + d \neq 2$. In (4.1), raising a vector to a power and division, μ_1/μ_2, are done componentwise so that μ_1/μ_2 is the vector with i-th component μ_1^i/μ_2^i. The function $\sum : \mathcal{M} \mapsto \mathbb{R}$ is defined by $\sum v = \sum_1^n v^i$ where $v = (v^1, \ldots, v^n)' \in \mathcal{M}$. If $\lambda \neq 0$ and $d = 2 - \lambda$, then

$$\bar{D}_{\lambda,2-\lambda}(\mu_1, \mu_2) = \sum \frac{\mu_2^{\lambda} - \mu_1^{\lambda} + \lambda \mu_1^{\lambda} \log(\mu_1/\mu_2)}{\lambda^2} \tag{4.2}$$

where $\log(\mu_1/\mu_2)$ is the vector with components $\log(\mu_1^i/\mu_2^i)$. If $\lambda \neq 0$ and $d = 2$, then

$$\bar{D}_{\lambda,2}(\mu_1, \mu_2) = \sum \frac{\lambda \log(\mu_2/\mu_1) + (\mu_1/\mu_2)^{\lambda} - 1}{\lambda^2}. \tag{4.3}$$

If $\lambda = 0$ and $d \neq 2$, then

$$\bar{D}_{0,d}(\mu_1, \mu_2) = \sum \frac{\mu_1^{2-d} - \mu_2^{2-d} + (2-d)\mu_2^{2-d}\log(\mu_2/\mu_1)}{(2-d)^2}. \tag{4.4}$$

If $\lambda = 0$ and $d = 2$, then

$$\bar{D}_{0,2}(\mu_1, \mu_2) = \sum \frac{(\log \mu_1 - \log \mu_2)^2}{2}. \tag{4.5}$$

Equations (4.1) to (4.5) are obtained by writing $\eta = f(\mu; \lambda)$ as a function of μ in the appropriate divergence function from Table 3. In particular, equation (4.1) is obtained by replacing η_1 (η_2) with $f(\mu_1; \lambda)$ ($f(\mu_2; \lambda)$) in the divergence with variance matrix equal to $\lambda^2 \operatorname{diag}(\eta^p)$ where $p = 2+(d-2)/\lambda$. For a fixed d, a family of minimum λ-divergence estimators $\hat{\beta}_\lambda$ is defined by minimizing $\bar{D}_{\lambda,d}(y, \mu(\beta))$. If some of the components of y are zero, then adjustments must be made to y (as is done for minimum χ^2_N estimators) when $\lambda \leq 0$ or $d = 2$.

Now we consider the asymptotic properties of the minimum f-divergence estimator $\hat{\beta}_f$. Once again, it will be convenient to use the f-divergence parameter η rather than the mean parameter μ. Let $\hat{\beta}_f$ be the minimum f-divergence estimate for y; i.e., $\hat{\beta}_f$ minimizes $D_f(z, \eta(\beta))$ where $z = f(y)$ for all $\beta \in \mathcal{B}$. We have seen that there exists a quasi-likelihood function $l(\eta; z)$ whose maximization corresponds to minimizing $D_f(z, \eta)$. If $\hat{\beta}_f$ maximizes $l(\eta(\beta); z)$, then

$$\frac{\partial}{\partial \beta} l(\eta(\beta); z)|_{\beta=\hat{\beta}_f} = 0. \tag{4.6}$$

Since the metric matrix for D_f is the inverse of the variance matrix for $l(\eta; z)$, we can use (2.1) to rewrite (4.6) as

$$\left(\frac{\partial \eta}{\partial \beta}\right)' V_f^{-1}(\beta)(z - \eta(\beta))|_{\beta=\hat{\beta}_f} = 0. \tag{4.7}$$

Equation (4.6) shows that $\hat{\beta}_f$ is also a mql estimator and we can expect it to have the corresponding asymptotic properties for mql estimators. However, the asymptotic assumptions are slightly different from those given in McCullagh (1983). Our data is a vector of fixed length n and has components of order N. McCullagh (1983) considers data of length N where $N \to \infty$.

One approach to establishing the asymptotic properties of $\hat{\beta}_f$ is to check that the argument outlined in McCullagh (1983) can be extended to data vectors of fixed length. Instead, we shall use two results from Ferguson (1958) for estimators that satisfy

$$Q(\beta)(z - \eta(\beta)) = 0 \tag{4.8}$$

for some $m \times n$ matrix $Q(\beta)$. Notice that $\hat{\beta}_f$ satisfies (4.8) with $Q(\beta) = (\partial\eta/\partial\beta)' V_f^{-1}$. The strongest assumption we need is that $\sqrt{N}(Y - \mu(\beta))$ be asymptotically normal with mean 0 and variance matrix $V(\beta)$. Although this assumption

is not valid for all generalized linear models, it applies to repeated sampling situations and to quasi-likelihoods proportional to that of the Poisson or multinomial distribution where y^i is now the number of observations in the i-th category divided by the total number of observations. Notice that y^i is defined differently here than in the previous section. We shall also assume the following:

ASSUMPTION 1. $\mathcal{B}$ is an open subset of $\mathbb{R}^m$.

ASSUMPTION 2. The map $\eta(\beta) = f(\mu(\beta)) : \mathcal{B} \mapsto \mathbb{R}^n$ $(m < n)$ is smooth and homeomorphic on its image.

ASSUMPTION 3. $Q(\beta)(\partial\eta/\partial\beta)$ is nonsingular for each $\beta \in \mathcal{B}$.

ASSUMPTION 4. $Q(\beta)$ and $\partial Q(\beta)/\partial\beta^a$, $a = 1, 2, \ldots, m$, are continuous.

For $\hat{\beta}_f$, Assumption 3 says that W is nonsingular on $\mathcal{B}$.

The following two results will be used.

THEOREM 4.1. *If Assumptions* 1–4 *hold and* $\sqrt{N}(Z - \eta(\beta))$ *is asymptotically normal with mean zero and variance matrix* $\Sigma(\beta)$, *then there exists a neighborhood,* $\mathcal{O}$, *of the set* $\{\eta(\beta) : \beta \in \mathcal{B}\}$ *and a unique function* $\hat{\beta}(z)$ *from* $\mathbb{R}^n$ *to* $\mathbb{R}^m$ *continuous in* $\mathcal{O}$ *such that* $\hat{\beta}(\eta(\beta)) = \beta$ *for all* $\beta \in \mathcal{B}$ *and* $Q(\hat{\beta}(z))(z - \eta(\hat{\beta}(z))) = 0$. *Furthermore,* $\sqrt{N}(\hat{\beta} - \beta)$ *is asymptotically normal with mean zero and covariance matrix*

$$\left[Q\left(\frac{\partial\eta}{\partial\beta}\right)\right]^{-1} Q\Sigma Q' \left[\left(\frac{\partial\eta}{\partial\beta}\right)' Q'\right]^{-1}.$$

The above theorem is actually a slightly weaker version of Theorem 1 of Ferguson (1958). The second theorem shows in which cases the estimator satisfying the linear form in (4.6) has minimum variance.

THEOREM 4.2. *If the assumptions of Theorem* 4.1 *hold and if there exists an* $n \times n$ *nonsingular matrix* Σ^- *such that* $\Sigma\Sigma^-(\partial\eta/\partial\beta) = (\partial\eta/\partial\beta)$, *then the asymptotic variance matrix of* $\hat{\beta}(Z)$ *is minimized when* $Q(\beta) = (\partial\eta/\partial\beta)\Sigma^-$. *The minimum is* $[(\partial\eta/\partial\beta)'\Sigma^-(\partial\eta/\partial\beta)]^{-1}$.

The proof of Theorem 4.2 is given in Ferguson (1958).

It is now easy to prove that the minimum divergence estimators have optimal asymptotic properties analogous to those of the maximum quasi-likelihood estimates.

PROPOSITION 4.1. *Suppose* $\sqrt{N}(Y - \mu(\beta))$ *is asymptotically normal with mean zero and covariance matrix* $V(\beta)$ *and* V^- *is a nonsingular generalized inverse of* V. *We also assume that* f *is a diffeomorphism on* $\mathcal{Y}$ *and is defined so*

that $\eta(\beta) = f(\mu(\beta))$ *satisfies Assumptions* 1–4 *and* $\hat{\beta}_f$ *is the minimum divergence estimator satisfying* (4.7) *with* $z = f(y)$. *Then*

(1) $E(\hat{\beta}_f - \beta) = o(N^{-1/2})$.

(2) $\sqrt{N}(\hat{\beta}_f - \beta)$ *is asymptotically normal with mean zero and covariance matrix* $W^{-1} = \left[\left(\frac{\partial\mu}{\partial\beta}\right)' V^- \left(\frac{\partial\mu}{\partial\beta}\right)\right]^{-1}$.

(3) *If* $VV^-(\partial\mu/\partial\beta) = (\partial\mu/\partial\beta)$, *variance then* $\hat{\beta}_f$ *has minimum asymptotic among all estimators satisfying* (4.8).

(4) *If* $D_f(\cdot,\cdot)$ *exists, then* $2ND_f(\theta(\hat{\beta}_f), \theta(\beta))$ *is asymptotically* χ^2_m.

PROOF. Since $f(\cdot)$ is smooth and Y is asymptotically normal, a Taylor series expansion shows that $Z = f(Y)$ has mean $\eta = f(\mu)$ and $\sqrt{N}(Z - \eta)$ is asymptotically normal with mean zero and covariance matrix $\Sigma = (\partial\eta/\partial\mu)V(\partial\eta/\partial\mu)'$. Statements (1) and (2) now follow from Theorem 4.1. An easy calculation shows that $VV^-(\partial\mu/\partial\beta) = (\partial\mu/\partial\beta)$ implies $\Sigma\Sigma^-(\partial\eta/\partial\beta) = (\partial\eta/\partial\beta)$. Now, Theorem 4.2 is invoked to prove (3). A Taylor series expansion of $D_f(\eta(\hat{\beta}_f), \eta(\beta))$ shows,

$$
\begin{aligned}
(4.9) \qquad & 2ND_f(\eta(\hat{\beta}_f), \eta(\beta)) \\
& = N(\hat{\beta}_f - \beta)' \left(\frac{\partial\eta}{\partial\beta}\right)' V_f^{-1} \left(\frac{\partial\eta}{\partial\beta}\right)(\hat{\beta}_f - \beta) + O_p(N^{-1/2}) \\
& = N(\hat{\beta}_f - \beta)' W (\hat{\beta}_f - \beta) + O_p(N^{-1/2}).
\end{aligned}
$$

Statement (4) now follows from the asymptotic normality of $\sqrt{N}(\hat{\beta}_f - \beta)$.

Note A. Statement (3) of this proposition does not compare the minimum f-divergence estimate with the mql estimate. Statement (3) says that for a given transformation $f(\cdot)$, we cannot do any better by using a different $Q(\beta)$, i.e., a different divergence function. The fact that the variance matrix W^{-1} for the optimal choice of $Q(\beta)$ does not depend on $f(\cdot)$ shows that we cannot do any better (or worse) by transforming the data. Hence, all $\hat{\beta}_f$ have the same first order asymptotic properties. In many situations V is nonsingular, so that the hypothesis of (3) is automatically satisfied. If Y is multinomial, V has ij-th component $-\mu^i\mu^j$ for $i \neq j$ and ii-th component $\mu^i(1 - \mu^i)$ and is singular. If V^- is the diagonal matrix with diagonal μ^{-1}, then it is easily checked that V^- satisfies the hypothesis of (3) (Ferguson (1958)).

Note B. The class of estimators for which $\hat{\beta}_f$ is optimal expressed in (4.8) is different than that for mql estimators given in (2.3). The following calculations show that the two are essentially the same. We shall use the fact that $Z - \eta(\beta)$ is $O_p(N^{-1/2})$, $Q(\tilde{\beta}) = Q(\beta) + O_p(N^{-1/2})$ and

$$
\eta(\tilde{\beta}) = \eta(\beta) + \left(\frac{\partial\eta}{\partial\beta}\right)(\tilde{\beta} - \beta) + O_p(N^{-1}).
$$

Replacing β with $\tilde{\beta}$, z with Z in (4.8), and then making the above substitutions gives

$$Q(\tilde{\beta})(Z-\eta(\tilde{\beta}))=Q(\beta)(Z-\eta(\beta))-Q(\beta)\left(\frac{\partial\eta}{\partial\beta}\right)(\tilde{\beta}-\beta)+O_p(N^{-1}). \tag{4.10}$$

If $Q_1(\beta)=Q(\beta)(\partial\eta/\partial\beta)$ is invertible then (4.8) is equivalent to

$$\tilde{\beta}-\beta=L_1(\beta)(Z-\eta(\beta))+O_p(N^{-1}) \tag{4.11}$$

where $L_1(\beta)=Q_1^{-1}(\beta)Q(\beta)$. Under repeated sampling $Z-\eta=(\partial f/\partial\mu)(Y-\mu)+O_p(N^{-1})$, so that (4.11) is equivalent to

$$\tilde{\beta}-\beta=L(\beta)(Y-\mu)+O_p(N^{-1})$$

where $L(\beta)=L_1(\beta)(\partial f/\partial\mu)$.

Note C. Mql estimators and minimum f-divergence estimators have different higher order asymptotic properties and which estimator is optimal will depend on the higher order moment structure of the data. The higher order asymptotic properties are naturally studied using Amari's dual geometries (1985, 1987). Vos (1991*b*) considers the second and third order asymptotic properties of minimum f-divergence estimators using the dual geometries.

Ship Damage Example (continued). We have already seen how the minimum χ^2_P estimate improves the fit of model (3.3). Certainly the other minimum f-divergence estimators $\hat{\beta}_\lambda$ cannot improve on the fit when lack of fit is measured by $\bar{D}_{2,1}(y,\mu(\hat{\beta}_\lambda))$. If we measure lack of fit using $\bar{D}_{\lambda,1}(y,\mu(\hat{\beta}_\lambda))$, then it is no longer clear which $\hat{\beta}_\lambda$ provides the best fit. The trouble here is that it is difficult to know how to compare $\bar{D}_{\lambda,1}(y,\mu(\hat{\beta}_\lambda))$ for different values of λ, especially when σ^2 must be estimated. Another way to measure lack of fit is by the size of the largest standardized residual. This is of particular interest when the data may contain outliers. For the ship data, the estimate with the smallest standardized residual is $\hat{\beta}_\lambda$ with $\lambda=2.12$. The largest standardized residual is 1.89 (for the 21st observation). When $\lambda=2$, the largest standardized residual is very similar, 1.98, and there is no evidence suggesting that we use a different minimum divergence estimator. Since we are using the data to estimate λ, the degrees of freedom should be reduced by one, so that the standardized residuals are increased by $\sqrt{25/24}$. This is only one method of allowing the data to choose λ. In other applications where other problems present themselves, different criteria could be used to find minimum divergence estimators. In other situations, one might want to use the more general minimum f-divergence estimators.

Whenever conclusions rely on asymptotic approximations we must assume the sample is large enough for these approximations to be valid. How large is large enough is often difficult to say. One way to get insight into this problem is to use several first order efficient estimators. If the sample is large enough and the other assumptions hold one can expect each estimator to give roughly the same

conclusion. For this example the particular values for the parameter estimates change as λ varies, but the major conclusions remain the same for $\lambda \in [.33, 2.5]$. For larger values of λ, ship type C is no longer safer than type A. For smaller values of λ, type B is no longer safer than type D. The other assumptions remain unchanged over a wide range for λ.

5. Relationship to power divergence statistics

The minimum divergence estimators are related to the power divergence statistics of Cressie and Read (1984). See also Read and Cressie (1988). Before considering the similarities, we mention two important differences. First, Read and Cressie assume that the data comes from a multinomial distribution. We have only assumed a particular known functional relationship between the mean and variance. When the variance is proportional to the mean the divergence discussed by Read and Cressie is very similar to ours. Second, Read and Cressie only consider estimators for which the sum of the fitted values equals the sum of the observations. The conditionality principle requires this since the sum of the observations is ancillary for the parameters β. Since we make no assumptions about the exact distributional form of the data, the concept of ancillarity is not defined. Particularly when the dispersion parameter needs to be estimated, it is not clear that one should condition on the total number of observations. Following the approach taken for glms by McCullagh and Nelder (1983), we do not require that the sum of the fitted values equal the sum of the observations.

Read and Cressie ((1988), p. 94) define the power divergence between an n-tuple $X = (X_1, \ldots, X_n)$ of random variables and a multinomial distribution with parameter $m = (m_1, \ldots, m_n)$ where $\sum m_i = N$, the total number of observations. The power divergence statistic is written $I^\lambda(X : m)$ and defined by

$$I^\lambda(X : m) = \sum_{i=1}^{n} \frac{1}{\lambda(\lambda+1)} \left\{ X_i \left[\left(\frac{X_i}{m_i} \right)^\lambda - 1 \right] + \lambda(m_i - X_i) \right\}. \tag{5.1}$$

For $\lambda = 0$ and $\lambda = -1$, the power divergence I^λ is defined by its limiting value:

$$I^0(X : m) = \sum_{i=1}^{n} X_i \log(X_i/m_i) + (m_i - X_i), \tag{5.2}$$

$$I^{-1}(X : m) = \sum_{i=1}^{n} m_i \log(m_i/X_i) + (X_i - m_i). \tag{5.3}$$

Comparing (5.1)–(5.3) to (4.1), (4.2) and (4.4) with $d = 1$ shows that

$$\bar{D}_{\lambda,1}(\mu_1, \mu_2) = N^{-1} I^{-\lambda}(N\mu_2 : N\mu_1) \tag{5.4}$$

for any real λ. We see then a close relationship between I^λ and $D_{\lambda,1}$, and that the divergence measures $D_{\lambda,d}$ and D_f offer extensions to the power divergence statistic I^λ of Read and Cressie (1988).

6. Conclusion and related research

When a random vector is described using a quasi-likelihood function, the mql estimation is but one method of estimating parameters. In several situations, it is possible to define a family of minimum f-divergence estimators for a given quasi-likelihood function. These estimators have the same first order asymptotic properties as the mql estimator, can be obtained using the same estimation algorithm, and admit a dual geometric structure similar to that for mql estimators.

The added flexibility of using a family of estimators often allows one to improve the fit of a given model. We can also view the minimum f-divergence estimators as varying, in an indirect way, the higher order moment structure of the model while leaving the first two moments unchanged. To prefer one minimum f-divergence estimator over another because of its asymptotic properties must result from a difference in the moment structure beyond the first two moments because the first order asymptotic properties only depend on the first two moments.

Although the minimum f-divergence estimators are used in different contexts and under different assumptions, there is a formal similarity between the power divergences discussed in Read and Cressie (1988) and a special subfamily of the f-divergences. We have discussed the minimum f-divergence estimators in terms of higher moment assumptions. Lindsay (1991) uses robustness considerations in comparing estimators from the family of estimators described by Read and Cressie (1988). Except for the Appendix, the geometric structure of the minimum f-divergence estimator has not been treated. As it turns out, the α-geometries of Amari (1985) are closely related to the geometry generated by the f-divergence in special but important cases. It is also possible to use the geometry to describe minimum f-divergence estimators and how they depend on higher order moment assumptions. The geometric considerations are discussed in Vos (1991*b*).

Acknowledgement

The example showing the difficulties encountered with the normal distribution and the reciprocal transformation was suggested by a referee.

Appendix

Amari (1985) shows how a divergence allows one to construct a dual geometric structure on a smooth manifold. We give this construction explicitly here for the manifold corresponding to the quasi-likelihood functions.

In order to place a dual geometric structure on $\mathcal{M}$ we require that $\mathcal{M}$ be a smooth manifold. It is customary to define a geometric structure on the manifold itself, rather than on the range of one of its parameterizations (coordinate charts). Hence, in defining a geometric structure on an exponential family, the manifold that we consider is the set of densities S_1 rather than the natural parameter space. For quasi-likelihoods, the corresponding manifold S consists of a set of equivalence classes of n-dimensional distributions for a random variable Y such that each distribution has the same support, $E(Y) = \mu$ for some $\mu \in \mathcal{M}$, and $\text{Var}(Y) = V(\mu)$ where $V(\mu)$ is a known function of μ. Random variables Y_1 and

Y_2 are equivalent if $E(Y_1) = E(Y_2)$. Rather than use S, it is notationally simpler to define the dual geometries directly on $\mathcal{M}$.

Since $\mathcal{M}$ is a smooth manifold, there is a tangent space $T_\mu\mathcal{M}$ at each $\mu \in \mathcal{M}$. One possible basis for $T_\mu\mathcal{M}$ is the set of derivative operators $\{\partial/\partial\mu^r; r = 1, \ldots, n\}$. To define a dual geometric structure on $\mathcal{M}$, however, it will be convenient to work with the natural basis for the η parameter; that is, $\{\partial_i = \partial/\partial\eta^i; i = 1, \ldots, n\}$ where $\eta = f(\mu) = (f^1(\mu), \ldots, f^n(\mu))'$ and f is a diffeomorphism. These bases are related by

$$\frac{\partial}{\partial\mu^r} = \sum_{j=1}^{n} \frac{\partial f^j}{\partial\mu^r}\frac{\partial}{\partial\eta^j}. \tag{A.1}$$

We use the convention that $r, s, t, \ldots$ are used with the μ parameterization while $i, j, k, \ldots$ are used with the η parameterization. The first step in placing a dual geometric structure on $\mathcal{M}$ is to make $\mathcal{M}$ into a Riemannian manifold. We do this by defining a metric $\langle\cdot,\cdot\rangle_\eta$ using the metric matrix $G(\eta) = (g_{ij}(\eta))$ of the divergence function; that is, $\langle\partial_i, \partial_j\rangle_\eta = g_{ij}(\eta)$. Next, we define a pair of dual affine connections ∇ and ∇^*. Let $\Gamma_{ijk} = \langle\nabla_{\partial_i}\partial_j, \partial_k\rangle$ and $\Gamma^*_{ijk} = \langle\nabla^*_{\partial_i}\partial_j, \partial_k\rangle$ be the components of these connections in terms of the natural basis $\{\partial_i; i = 1, \ldots, n\}$ for η. Then $\Gamma_{ijk} = 0$ for all $i, j, k = 1, \ldots, n$ and $\Gamma^*_{ijk} = \partial_i g_{jk}$ for all $i, j, k = 1, \ldots, n$. It is easily verified that these definitions satisfy the definition for an affine connection and that they are dual; i.e., $A\langle B, C\rangle = \langle\nabla^*_A B, C\rangle + \langle B, \nabla_A C\rangle$ for any vector fields A, B, C. Furthermore, since $\Gamma_{ijk} = 0$, $\mathcal{M}$ is flat in the connection ∇ and therefore $\mathcal{M}$ is flat in the dual connection ∇^* (Amari ((1985), p. 72)).

Now we consider the f-divergences defined in Section 4. Suppose $l(\mu; y)$ is a quasi-likelihood for which we can define a divergence $D(\mu_1, \mu_2)$. Let $D_f(\eta_1, \eta_2)$ with $\eta = f(\mu)$ be the f-divergence associated with $D(\mu_1, \mu_2)$. The geometries defined by these divergences are closely related. Let $\langle\cdot,\cdot\rangle$ be the metric defined by $D_f(\eta_1, \eta_2)$ so that $\langle\partial/\partial\eta^i, \partial/\partial\eta^j\rangle_\eta = g_{ij}(\eta)$ where $(g_{ij}) = V_f^{-1}(\eta)$. From (A.1) we see that

$$\left\langle \frac{\partial}{\partial\mu^r}, \frac{\partial}{\partial\mu^s} \right\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial f^i}{\partial\mu^r} g_{ij} \frac{\partial f^j}{\partial\mu^s}. \tag{A.2}$$

Since $(\partial f/\partial\mu)' V_f^{-1}(\partial f/\partial\mu) = V^-(\mu)$, the right-hand side of (A.2) is just $v^-_{rs}(\mu)$ where $(v^-_{rs}(\mu)) = V^-(\mu)$. Hence, $\langle\cdot,\cdot\rangle$ is also the metric defined by $V^-(\mu)$. Since $D(\mu_1, \mu_2)$ and $D_f(\eta_1, \eta_2)$ define the same metric, as Riemannian manifolds the geometry is the same under either divergence for any diffeomorphism f.

The difference comes in how the connections are defined. Let ∇ and ∇^* be the dual connections defined by $D(\mu_1, \mu_2)$; in particular, the components of ∇ with respect to $\{\partial_r = \partial/\partial\mu^r; r = 1, \ldots, n\}$ are zero for all μ. Amari ((1985), p. 47) calls such a coordinate system affine for the connection ∇. If $\overset{f}{\nabla}$ and $\overset{f}{\nabla}{}^*$ are the dual connections for $D_f(\eta_1, \eta_2)$, then the connection $\overset{f}{\nabla}$ is defined to ensure that the η parameterization is affine. In studying exponential families we see that the dual geometries are always defined so that the expectation parameter is affine for one

of the connections (and the natural parameter is affine for the dual connection). The generalization of the f-divergence is that we define a connection that has affine parameter $\eta = f(\mu)$ where f is any diffeomorphism relating μ and η. In general, the f-connections are different from the α-connections of Amari ((1985), p. 39). The α-connections are defined using a linear combination of the primal connection and its dual and for most linear exponential families the manifold is flat only in the 1- and -1-connections. A notable exception is the multinomial distribution which is flat in each α-connection. In this case the α-connection is the same as the primal connection defined by the power divergence $D_{1,\lambda}(\eta_1, \eta_2)$ with $\lambda = (1 + \alpha)/2$.

REFERENCES

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

Amari, S. (1987). Differential geometrical theory of statistics, *Differential Geometry in Statistical Inference*, IMS Lecture Notes—Monograph Series, Vol. 10, 19–94, Hayward, California.

Barndorff-Nielsen, O. E. (1987). Differential geometry and statistics: some mathematical aspects, *Indian J. Math.*, **29**, 335–350.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood!, *Ann. Statist.*, **8**, 457–487.

Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests, *J. Roy. Statist. Soc. Ser. B*, **46**, 440–464.

Eguchi, S. (1983). Second-order efficiency of minimum contrast estimators in a curved exponential family, *Ann. Statist.*, **11**, 793–803.

Eguchi, S. (1985). A differential geometric approach to statistical inference on the basis of contrast functionals, *Hiroshima Math. J.*, **15**, 341–391.

Ferguson, T. (1958). A method of generating best asymptotically normal estimates with applications to the estimation of bacterial densities, *Ann. Math. Statist.*, **29**, 1046–1062.

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

Lindsay, B. (1991). Residual adjustment functions: efficiency, robustness and curvature in minimum distance estimators, *Ann. Statist.* (to appear).

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

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

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

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples, *J. Roy. Statist. Soc. Ser. B*, **24**, 46–72.

Rao, C. R. (1987). Differential metrics in probability spaces, *Differential Geometry in Statistical Inference*, IMS Lecture Notes—Monograph Series, Vol. 10, 217–240, Hayward, California.

Read, N. and Cressie, T. R. C. (1988). Goodness-of-fit statistics for discrete multivariate data, *Springer Ser. Statist.*, Springer, New York.

Ruppert, D. and Aldershof, B. (1989). Transformations to symmetry and homoscedasticity, *J. Amer. Statist. Assoc.*, **84**, 437–446.

Vos, P. W. (1987). Dual geometries and their applications to generalized linear models, Ph.D. Dissertation, University of Chicago.

Vos, P. W. (1991*a*). A geometric approach to detecting influential cases, *Ann. Statist.*, **19**, 1570–1581.

Vos, P. W. (1991*b*). The geometry of f-divergence, *Ann. Inst. Statist. Math.*, **43**, 515–537.

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