Ann. Inst. Statist. Math.
Vol. 43, No. 3, 515–537 (1991)

GEOMETRY OF f-DIVERGENCE*

PAUL W. VOS

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

(Received December 26, 1989; revised August 7, 1990)

Abstract. Amari's ± 1-divergences and geometries provide an important description of statistical inference. The ± 1-divergences are constructed so that they are compatible with a metric that is defined by the Fisher information. In many cases, the ± 1-divergences are but two in a family of divergences, called the f-divergences, that are compatible with the metric. We study the geometries induced by these divergences. Minimizing the f-divergence provides geometric estimators that are naturally described using certain curvatures. These curvatures are related to asymptotic bias and efficiency loss. Under special but important restrictions, the geometry of f-divergence is closely related to the α-geometry, Amari's extension of the ± 1-geometries. One application of these results is illustrated in an example.

Key words and phrases: Divergence, contrast functional, yoke, minimum divergence estimator, geometric estimator, curvature, dual geometries, statistical manifold.

1. Introduction

The f-divergence can be viewed as an extension of the divergence measures described by Read and Cressie (1988) who show that many of the functions commonly minimized in analyzing contingency tables belong to a family of divergence measures. Vos (1991) shows the close relationship between divergence measures and quasi-likelihood functions. The divergence associated with a quasi-likelihood is one member of a family of divergence measures, called the f-divergences. A given quasi-likelihood function l allows us to model the error distribution as a function of the mean while the collection of f-divergences for l allows us to model the skewness of the error distribution. One way to describe the f-divergence is via its relationship to quasi-likelihood functions and generalized linear models. This is done in Vos (1991). Further discussion on generalized linear models and quasi-likelihood functions is given in McCullagh and Nelder (1989).

The f-divergence can also be studied geometrically. Under weak regularity conditions, modeling the variance of an error distribution as a function of the

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

mean allows one to construct a Riemannian manifold. The quasi-likelihood gives one divergence that is compatible with this geometric structure. The collection of all divergences compatible with a given Riemannian manifold is just the set of f-divergences. In this paper, we consider the geometric properties of the f-divergence. The family of f-divergences allows us to define a family of geometric estimators, called the minimum f-divergence estimators. We show that a natural way to compare these estimators is in terms of a curvature. Along with a geometric interpretation, this curvature appears in higher order asymptotic calculations and describes second order bias and efficiency loss. We also consider an important special class of quasi-likelihood functions where the geometry induced by a subset of the f-divergences is the same as Amari's α-geometries (Amari (1985, 1987)).

The quasi-likelihood functions, together with the associated f-divergence measures, provide a powerful family for modeling the variance and skewness of a distribution. The role of the f-divergences in applications is illustrated through an example.

2. Divergence and geometry

Divergence measures are typically defined on a differentiable manifold S. For us, S will be a smooth n-dimensional Riemannian manifold with metric $\langle \cdot, \cdot \rangle$. A divergence is a smooth mapping $D(\cdot, \cdot) : S \times S \mapsto \mathbb{R}$ for which at each point of S there exists a coordinate system $\eta = (\eta^1, \ldots, \eta^n)'$, called the divergence parameterization, with image $\mathcal{N} = \eta(S)$ such that $D(\eta_1, \eta_2) = D(\eta^{-1}(\eta_1), \eta^{-1}(\eta_2))$ satisfying the conditions

(2.1)
$$\begin{aligned} &\text{a)} \quad D(\eta_1, \eta_2) \geq 0 \text{ for all } \eta_1, \eta_2 \in \mathcal{N}, \\ &\qquad\qquad \text{equality holding if and only if } \eta_1 = \eta_2, \\ &\text{b)} \quad g_{rs}(\eta_1) = \frac{\partial}{\partial \eta_1^r} \frac{\partial}{\partial \eta_1^s} D(\eta_1, \eta_2) \text{ does not depend on } \eta_2 \\ &\qquad\qquad \text{and the matrix } (g_{rs}) \text{ is positive definite for all } \eta_1 \in \mathcal{N}. \end{aligned}$$

Notice that the smoothness of D and a) imply $\partial_{1r} D(\eta_1, \eta_2) = 0 = \partial_{2r} D(\eta_1, \eta_2)$ when $\eta_1 = \eta_2$ where $\partial_{1r} = \partial / \partial \eta_1^r$ and $\partial_{2r} = \partial / \partial \eta_2^r$. The matrix function $g_{rs}(\eta_1)$ given in condition b), called the metric matrix, defines a metric on the manifold so that S need not be Riemannian. In many applications, however, there will be an obvious given metric (defined by the second order moments) which need not be the same as the metric determined by a divergence. When the metrics are the same the divergence is called compatible. That is, D is compatible with $\langle \cdot, \cdot \rangle$ if $g_{rs} = \langle \partial_r, \partial_s \rangle$ where $\partial_r = \partial / \partial \eta^r$ and $\partial_s = \partial / \partial \eta^s$. We shall only consider divergences compatible with the metric.

The definition (2.1) for divergence is the same as the one given in Amari (1985). It is important to note that divergence has been defined differently by other authors. Eguchi (1985) defines a contrast functional ρ on a pair of probability measures which is positive except when the measures agree. Eguchi (1985) also calls ρ a divergence. Rao (1987) gives yet another definition for a divergence. A related term is a yoke. Barndorff-Nielsen (1987) defines a yoke on a smooth manifold S

with coordinate chart $\omega = (\omega^1, \ldots, \omega^n)$. A smooth function $g : S \times S \mapsto \mathbb{R}$ is a yoke if for all ω

$$\text{(2.2)} \quad \begin{aligned} &\text{a)} \quad \frac{\partial}{\partial \omega_1^j} g(\omega_1, \omega_2)|_{\omega_1=\omega_2} = 0 \text{ for } j = 1, \ldots, n, \\ &\text{b)} \quad -\frac{\partial^2}{\partial \omega_1^j \partial \omega_2^k} g(\omega_1, \omega_2)|_{\omega_1=\omega_2=\omega} \text{ is a positive definite matrix.} \end{aligned}$$

It should be clear that Amari's definition of a divergence is a special case of the more general concepts of a contrast functional and yoke. Further information on contrast functionals can be found in Pfanzagl (1973) and Eguchi (1983) while yokes are discussed in Blæsild (1987).

The divergence parameter η allows us to define a connection on S. All connections considered will be smooth, affine and torsion-free. It will be convenient to use a special set of vector fields to define the *divergence connection* on S. The set of vectors $\{\partial_{rp} = \partial/\partial\eta^r|_p\}$ is called the η-natural basis for T_pS; let ∂_r be the corresponding vector field defined on a neighborhood of p. The (primal) divergence connection can be defined in terms of these vector fields by $\nabla_{\partial_r}\partial_s = 0$. The components of the divergence connection must also be zero in the η-coordinate system $\Gamma_{rst} = \langle \nabla_{\partial_r}\partial_s, \partial_t \rangle = 0$. If $\vec{v}$ is the tangent vector field to a curve $c(t)$ such that $\nabla_{\vec{v}}\vec{v} = 0$ on $c(t)$, then $c(t)$ is called a geodesic. Hence, the coordinate curves for the η-parameter are geodesic for ∇. Two connections ∇ and ∇^* are dual if

$$A\langle B, C\rangle = \langle \nabla_A B, C\rangle + \langle B, \nabla_A^* C\rangle$$

for all vector fields A, B, C. Hence, the components of the connection dual to ∇ must be $\Gamma^*_{rst} = \partial_r g_{st}$. In order to compare different connections, it will be useful to write the components of ∇ in terms of a common parameter μ. If $\Gamma_{ijk} = \langle \nabla_{\partial_i}\partial_j, \partial_k \rangle$ where $\partial_i = \partial/\partial\mu^i$, then

$$\text{(2.3)} \quad \begin{aligned} \Gamma_{ijk} &= \partial_i\eta^r\partial_j\eta^s\partial_k\eta^t\Gamma_{rst} + g_{rs}\partial_i\partial_j\eta^r\partial_k\eta^s \\ &= g_{rs}\partial_i\partial_j\eta^r\partial_k\eta^s. \end{aligned}$$

In (2.3) and the following, the Einstein summation convention is used.

For a given connection ∇ on a Riemannian manifold, we can define a tensor on vector fields A, B, C, D by

$$R(A, B, C, D) = \langle \nabla_A\nabla_B C - \nabla_B\nabla_A C - \nabla_{AB}C - \nabla_{BA}C, D\rangle$$

called the Riemannian curvature tensor for ∇. Using the divergence parameter η, it is clear that the Riemannian curvature for ∇ is zero everywhere. When the Riemannian curvature vanishes, the manifold is called flat. Hence, whenever a divergence can be defined on S, S must be flat in the divergence connection. By the properties of dual connections, if S is flat in the connection ∇ then S is also flat in the dual connection ∇^*. Further details concerning ∇ and ∇^* can be found in Amari (1985).

We note that a yoke provides a metric and a pair of dual connections. The yoke ρ defines a metric by

$$g_{ij}^{(\rho)} = \frac{\partial}{\partial\theta_1^i}\frac{\partial}{\partial\theta_1^j}\rho(\theta_1,\theta_2)|_{\theta_1=\theta_2=\theta} \tag{2.4}$$

and dual connections by

$$\begin{aligned}\Gamma_{ijk}^{(\rho)} &= -\frac{\partial}{\partial\theta_1^i}\frac{\partial}{\partial\theta_1^j}\frac{\partial}{\partial\theta_2^k}\rho(\theta_1,\theta_2)|_{\theta_1=\theta_2=\theta},\\ \Gamma_{ijk}^{(\rho)*} &= -\frac{\partial}{\partial\theta_2^i}\frac{\partial}{\partial\theta_2^j}\frac{\partial}{\partial\theta_1^k}\rho(\theta_1,\theta_2)|_{\theta_1=\theta_2=\theta}\end{aligned} \tag{2.5}$$

where $\theta = (\theta^1,\ldots,\theta^n)$ is any coordinate chart on S. When $\rho = D$ and θ is the divergence parameterization, it is easily verified that $g_{ij}^{(\rho)} = g_{ij}$, $\Gamma_{ijk}^{(\rho)} = \Gamma_{ijk}$ and $\Gamma_{ijk}^{(\rho)*} = \Gamma_{ijk}^*$. The additional structure of Amari's divergence ensures that S is flat in the ∇ and ∇^* connections.

For a given connection ∇, there are two functions relating elements of the tangent space to the manifold S. At each point η_0 the exponential map $\exp_{\eta_0}(\cdot)$ takes a neighborhood around the origin in $T_{\eta_0}S$ into S and its inverse $\exp_{\eta_0}^{-1}(\cdot)$ takes a neighborhood of η_0 into $T_{\eta_0}S$. We shall write $\vec{v}(\eta_0,\eta_1)$ for $\exp_{\eta_0}^{-1}(\eta_1)$ so that $\vec{v}(\eta_0,\eta_1)$ is the vector in $T_{\eta_0}S$ "connecting" (via the ∇ connection) the points η_0 and η_1. When the manifold is geodesically complete, $\vec{v}(\cdot,\cdot)$ can, in fact, be defined on all $S\times S$. When ∇ is defined from a divergence, then exp and $\vec{v}(\cdot,\cdot)$ can easily be defined in terms of the divergence parameter η. It is enough to define the exponential map on the basis vectors $\epsilon\partial_{r0} = \epsilon\partial/\partial\eta^r|_{\eta_0}$, as

$$\exp_{\eta_0}(\epsilon\partial_{r0}) = \eta_0 + \epsilon e_r \tag{2.6}$$

where $\epsilon > 0$ is chosen to ensure that $\epsilon\partial_{r0}$ is in the domain of $\exp_{\eta_0}$, $e_r = (e_r^i)$ and $e_r^i = 1$ if $i = r$ and is zero otherwise. The image of exp lies in S, but the right-hand side of (2.6) lies in $\mathcal{N}$. Strictly speaking, the right-hand side of (2.6) is the η coordinate expression of the exponential map of the left-hand side. From time to time it will be convenient to refer to η, or sometimes μ, as points in S; this should cause no confusion. In terms of the η parameter, $\vec{v}(\eta_0,\eta_1)$ is defined by $\vec{v}(\eta_0,\eta_1) = (\eta_1^r - \eta_0^r)\partial_{r0}$. For the dual divergence connection ∇^*, we can define the corresponding functions $\exp^*$ and $\vec{v}^*(\cdot,\cdot)$. Notice $\vec{v}^*(p_1,p_2)$ is not a vector dual to $\vec{v}(p_1,p_2)$. Both $\vec{v}(p_1,\cdot)$ and $\vec{v}^*(p_1,\cdot)$ have their range in $T_{p_1}S$; the dual terminology and the $*$ notation refer to the connection used to define this map.

We have shown that a divergence defines a connection on a manifold. To understand the relationship between this connection and the divergence, we study the divergence more closely. Let η be a divergence parameter for $D(p_1,p_2)$. Fix η_0 and define $\psi(\eta) = D(\eta,\eta_0)$. Since $\partial_r\partial_s\psi(\eta) = \langle\partial_r,\partial_s\rangle$, $\psi(\eta)$ is called a potential function for η. By Theorem 3.4 of Amari ((1985), p. 80), there is a dual parameter to η defined by $\eta_r^* = \partial_r\psi(\eta)$ and a dual potential function $\phi(\eta^*)$ for η^*. This dual potential function satisfies

$$\partial^r\partial^s\phi(\eta^*) = \langle\partial^r,\partial^s\rangle \quad\text{and}\quad \nabla^*_{\partial^r}\partial^s = 0 \tag{2.7}$$

where $\partial^r = \partial/\partial\eta^*_r$. We also have that $\eta^r = \partial^r\phi(\eta^*)$ and $\psi(\eta) + \phi(\eta^*) - \eta^r\eta^*_r = 0$. From the properties of ψ and ϕ, it is easily checked that

$$D(\eta_1, \eta_2) = \psi(\eta_1) + \phi(\eta_2^*) - \eta_1^r\eta^*_{2r}. \tag{2.8}$$

From (2.8) we have

$$D(\eta_1, \eta_2) + D(\eta_2, \eta_3) = D(\eta_1, \eta_3) + (\eta_1^r - \eta_2^r)(\eta^*_{3r} - \eta^*_{2r}). \tag{2.9}$$

Rewriting (2.9) in parameter-free form, we have

$$D(p_1, p_3) = D(p_1, p_2) + D(p_2, p_3) - \langle \vec{v}(p_2, p_1), \vec{v}^*(p_2, p_3)\rangle. \tag{2.10}$$

The divergence is a distance-like quantity that measures how near two points are. Equation (2.10) is related to the following identity for squared distances

$$\begin{aligned} &||\vec{v}(p_2, p_1) - \vec{v}^*(p_2, p_3)||^2 \\ &\quad = ||\vec{v}(p_2, p_1)||^2 + ||\vec{v}^*(p_2, p_3)||^2 - 2\langle \vec{v}(p_2, p_1), \vec{v}^*(p_2, p_3)\rangle \end{aligned} \tag{2.11}$$

where $||\vec{v}||^2 = \langle \vec{v}, \vec{v}\rangle$. Comparing (2.10) and (2.11) we see that the divergence behaves like one half times a squared distance. Certainly, a Taylor's series expansion of $D(\eta_1, \eta_2)$ makes it clear that

$$D(p_1, p_2) \approx \frac{1}{2}||\vec{v}(p_2, p_1)||^2 \tag{2.12}$$

for p_1 near p_2. What (2.11) shows is that in some ways the divergence behaves like one half times a squared distance globally. When $\vec{v}(p_2, p_1)$ and $\vec{v}^*(p_2, p_3)$ are orthogonal, (2.10) reduces to Amari's ((1985), p. 86) Pythagorean relationship for divergences.

The following proposition shows that a divergence is uniquely defined by specifying the metric matrix and divergence parameterization.

PROPOSITION 2.1. *If $D_A(p_1, p_2)$ and $D_B(p_1, p_2)$ are both compatible with $\langle \cdot, \cdot \rangle$ and each has divergence parameter η, then $D_A(p_1, p_2) = D_B(p_1, p_2)$ for all $p_1, p_2 \in S$.*

PROOF. From (2.8) we have

$$D_A(\eta_1, \eta_2) = \psi_A(\eta_1) + \phi_A(\eta_2^A) - \eta_1^r\eta^A_{2r}, \tag{2.13a}$$

$$D_B(\eta_1, \eta_2) = \psi_B(\eta_1) + \phi_B(\eta_2^B) - \eta_1^r\eta^B_{2r} \tag{2.13b}$$

where η^A is dual to η for D_A and η^B is dual to η for D_B. Since D_A and D_B are compatible with the same metric, $\partial_r\partial_s\psi_A = \partial_r\partial_s\psi_B$ and so

$$\psi_A(\eta) - \psi_B(\eta) = K_r\eta^r + C_1, \qquad \eta^A_r - \eta^B_r = K_r$$

where K_r and C_1 are constants. Since ϕ_A and ϕ_B are potential functions, $g^{rs}\partial_s\phi_A = \eta^r = g^{rs}\partial_s\phi_B$ so that

$$\phi_A(\eta^A) - \phi_B(\eta^B) = C_2$$

where C_2 is a constant. Subtracting (2.13b) from (2.13a) and making substitutions from the above equations, we obtain

$$D_A(\eta_1, \eta_2) - D_B(\eta_1, \eta_2) = C_1 + C_2.$$

Setting $\eta_1 = \eta_2$, we see that $C_1 + C_2 = 0$. □

Next, we consider the relationship between a manifold with divergence $D(\cdot, \cdot)$ and a statistical manifold. Lauritzen (1987) defines a statistical manifold as a Riemannian manifold $(S, \langle \cdot, \cdot \rangle)$ together with a symmetric tri-linear map $T(\cdot, \cdot, \cdot)$ called the skewness tensor. Although we also consider statistical models in which the moments beyond the variance have been left unspecified, we still call $T(\cdot, \cdot, \cdot)$ a skewness tensor. We have seen that a divergence defines a pair of dual connections; Lauritzen (1987) shows that there is a skewness tensor defined by these dual connections. The relationship between the dual connections and $T(\cdot, \cdot, \cdot)$ is

$$T(A, B, C) = \langle \nabla_A B - \nabla_A^* B, C \rangle \tag{2.14}$$

for vector fields A, B, C. The definition of $T(\cdot, \cdot, \cdot)$ in terms of the divergence is given in Proposition 2.2.

PROPOSITION 2.2. *Let η be the divergence parameter for $D(\cdot, \cdot)$ defined on S. The functions $\langle \cdot, \cdot \rangle$ and $T(\cdot, \cdot, \cdot)$ defined by*

$$\langle \partial_r, \partial_s \rangle = \partial_{1r}\partial_{1s} D(\eta_1, \eta_2), \tag{2.15a}$$

$$T(\partial_r, \partial_s, \partial_t) = -\partial_{1r}\partial_{1s}\partial_{1t} D(\eta_1, \eta_2) \tag{2.15b}$$

are symmetric tensors that make S a statistical manifold $(S, \langle \cdot, \cdot \rangle, T(\cdot, \cdot, \cdot))$.

Eguchi (1983) and Lauritzen (1987) prove this for the case where S is defined from a linear exponential family and η is the natural parameter. Their argument also holds for our case.

3. f-divergence

From (2.8) it is easily seen that the *dual divergence* $D^*(\eta_1^*, \eta_2^*) = D(\eta_2, \eta_1)$ where η_1 (resp., η_2) is a function of η_1^* (resp., η_2^*), is indeed a divergence with divergence parameter η^*. Both D and D^* are compatible with the metric $\langle \cdot, \cdot \rangle$. It is natural to ask what other divergences are compatible with $\langle \cdot, \cdot \rangle$. A divergence compatible with a given metric will be denoted by D_f and called an f-divergences for the following reason.

For many statistical inference problems, the family of probability models determine a Riemannian manifold with metric defined in terms of the Fisher information matrix. The estimates obtained by minimizing the likelihood divergence have optimal higher order asymptotic properties (Amari (1985)). Even when full distributional assumptions are weakened to assumptions about the first two moments, it is often possible to define a Riemannian manifold with metric given by the covariance structure. In these situations, estimates are typically obtained by minimizing the deviance divergence; although other divergences provide equally good estimates in some cases (Vos (1991)). The divergence parameter for the likelihood divergence and the deviance divergence is the mean parameter. So in a statistical setting, we are given a manifold S with divergence $D(\mu_1, \mu_2)$ that is compatible with the metric $\langle \cdot, \cdot \rangle$ on S and divergence parameter $\mu \in \mathcal{M}$. Since $(\eta, \mathcal{N})$ is another parameterization (coordinate chart) on S, there exists a diffeomorphism f such that $f(\mu) = \eta$. By Proposition 2.1, we can label any divergence compatible with $\langle \cdot, \cdot \rangle$ by the diffeomorphism $f : \mathcal{M} \mapsto \mathcal{N}$. Clearly, not all diffeomorphisms will provide an f-divergence. The image of f must be the divergence parameter for a divergence compatible with the metric; that is,

$$\partial_{1r}\partial_{1s} D_f(\eta_1, \eta_2) = \langle \partial_r, \partial_s \rangle = \frac{\partial \mu^i}{\partial \eta^r} g_{ij} \frac{\partial \mu^j}{\partial \eta^s}.$$

Each f-divergence on S determines a connection, called the f-connection $\overset{f}{\nabla}$, and a dual f-connection $\overset{f}{\nabla}{}^*$. The corresponding exponential and vector maps will be denoted by $\overset{f}{\exp}$, $\overset{f}{\exp}{}^*$, $\overset{f}{\vec{v}}$ and $\overset{f}{\vec{v}}{}^*$. The curvature and skewness tensors for the f-connections will be denoted by R^f and T^f, respectively.

We have already mentioned that the term divergence has been defined differently by other authors. The same is true for f-divergence. Csiszar's f-divergence (1967) is a divergence in the broader sense of a contrast functional.

Next, we consider some properties of the f-divergences and their corresponding geometric structure. It should be noted that the f_1- and f_2-divergences may be the same even when $f_1 \neq f_2$. Proposition 3.1 characterizes when two f-divergences are equal. Before giving this proposition we make a definition. Two parameterizations $\eta = (\eta^r)$ and $\xi = (\xi^\rho)$ on S are called *affine equivalent* or simply *equivalent* if η is a nonsingular affine transformation of ξ, i.e., $\eta^r = L^r_\rho \xi^\rho + K^r$ where $L = (L^r_\rho)$ is a nonsingular $n \times n$ matrix and $K = (K^r) \in \mathbb{R}^n$. The collection of parameterizations that are equivalent to η will be written $[\eta]$. When η and ξ are divergence parameterizations for $D_{f_1}(\eta_1, \eta_2)$ and $D_{f_2}(\xi_1, \xi_2)$, respectively, and $[\eta] = [\xi]$, then

$$f_1^r(\cdot) = L^r_\rho f_2^\rho(\cdot) + K^r. \tag{3.1}$$

Functions f_1 and f_2 are called equivalent if they satisfy equation (3.1) and we write $[f_1] = [f_2]$. Diffeomorphisms are called divergence equivalent if they define the same divergence. The following proposition shows that divergence equivalence and affine equivalence are the same.

PROPOSITION 3.1. *If $D_{f_1}(p_1,p_2)$ and $D_{f_2}(p_1,p_2)$ are compatible with the metric then*

$$D_{f_1}(p_1,p_2) = D_{f_2}(p_1,p_2) \Longleftrightarrow [f_1] = [f_2].$$

PROOF. Let $\eta = (\eta^r)$ be the f_1-divergence parameter and let $\xi = (\xi^\rho)$ be the f_2-divergence parameter. Furthermore, let $\partial_r = \partial/\partial\eta^r$ and $\partial_\rho = \partial/\partial\xi^\rho$ so that

$$\partial_r = \partial_r\xi^\rho\partial_\rho \quad \text{and} \quad \partial_\rho = \partial_\rho\eta^r\partial_r. \tag{3.2}$$

If $D_{f_1}(\eta_1,\eta_2) = D_{f_2}(\xi_1,\xi_2)$, then the metric defined by each must be the same, so

$$\langle\partial_\rho,\partial_\sigma\rangle = \partial_\rho\eta^r\partial_\sigma\eta^s\langle\partial_r,\partial_s\rangle. \tag{3.3}$$

Since $\langle\partial_r,\partial_s\rangle = \partial_{1r}\partial_{1s}D_{f_1}(\eta_1,\eta_2)$ and $\langle\partial_\rho,\partial_\sigma\rangle = \partial_{1\rho}\partial_{1\sigma}D_{f_2}(\xi_1,\xi_2)$, (3.3) can be rewritten as

$$\partial_{1\rho}\partial_{1\sigma}D_{f_2}(\xi_1,\xi_2) = \partial_\rho\eta^r\partial_\sigma\eta^s\partial_{1r}\partial_{1s}D_{f_1}(\eta_1,\eta_2). \tag{3.4}$$

Using (3.2) and our hypothesis, we find

$$\partial_{1\rho}\partial_{1\sigma}D_{f_2}(\xi_1,\xi_2) = \partial_\rho\eta^r\partial_\sigma\eta^s\partial_{1r}\partial_{1s}D_{f_1}(\eta_1,\eta_2) + \partial_\sigma\partial_\rho\eta^r\partial_{1r}D_{f_1}(\eta_1,\eta_2). \tag{3.5}$$

Comparing (3.4) and (3.5), we see that $\partial_\sigma\partial_\rho\eta^r = 0$ so $\partial_\rho\eta^r = L^r_\rho$ and, therefore, $\eta^r = L^r_\rho\xi^\rho + K^r$. The matrix L^r_ρ must be nonsingular to ensure that $\eta = (\eta^r)$ is a diffeomorphism. Conversely, suppose $[f_1] = [f_2]$ so that $\eta = L\xi + K$ where $L = (L^r_\rho)$ is a nonsingular $n\times n$ matrix and $K = (K^r) \in \mathbb{R}^n$. It is easily checked that the function defined by $D(\xi_1,\xi_2) = D_{f_1}(L\xi_1 + K, L\xi_2 + K)$ is a divergence compatible with the metric and that ξ is a divergence parameterization. By Proposition 2.1 we have $D_{f_2}(\xi_1,\xi_2) = D(\xi_1,\xi_2)$ and so $D_{f_2}(\xi_1,\xi_2) = D_{f_1}(\eta_1,\eta_2)$. □

For a given f-divergence, Proposition 3.2 shows that there exists an f^*-divergence whose f^*-connection is dual to the f-connection.

PROPOSITION 3.2. *For every f-divergence $D_f(p_1,p_2)$, there exists an f^*-divergence $D_{f^*}(p_1,p_2)$ such that*

$$D_{f^*}(p_1,p_2) = D_f(p_2,p_1) = D^*_f(p_1,p_2).$$

Furthermore, $\overset{f}{\nabla}{}^* = \overset{f^*}{\nabla}$.

PROOF. The equality between the second two divergences follows from the definition of the dual divergence D^*. For the first equality we simply define $f^* : \mathcal{M} \mapsto \mathcal{N}^*$ by $f^* = \varphi^*_\eta \circ \varphi^{-1}_\eta \circ f$ where φ_η is the divergence parameterization on S (for $D_f(p_1,p_2)$) and φ^*_η is the dual parameterization. From (2.7) we have $\overset{f}{\nabla}{}^*_{\partial^r}\partial^s = 0$ since η^* is the dual parameter for D_f and since η^* is the primal divergence parameter for D_{f^*}, $\overset{f^*}{\nabla}_{\partial^r}\partial^s = 0$. Hence, $\overset{f}{\nabla}{}^* = \overset{f^*}{\nabla}$. □

Equation (2.12) shows that locally an f-divergence is a squared distance and the defining properties of a divergence and equation (2.10) show that even globally an f-divergence retains some of the properties of a squared distance. The following proposition explores the conditions when D_f is a squared distance.

PROPOSITION 3.3. *Let $D_f(\cdot,\cdot)$ be compatible with the metric $\langle\cdot,\cdot\rangle$ and have divergence parameter $\eta = (\eta^r)$, dual divergence parameter $\eta^* = (\eta^*_r)$, and metric matrix $g_{rs} = \langle\partial_r, \partial_s\rangle$. Then the following are equivalent:*

(1) $D_f(\eta_1,\eta_2) = D_f(\eta_2,\eta_1)$,

(2) $[f^*] = [f]$,

(3) g_{rs} *is constant on* S,

(4) $\overset{f}{T}(\cdot,\cdot,\cdot) = 0$,

(5) $D_f(\eta_1,\eta_2) = \frac{1}{2}(\eta_1^r - \eta_2^r)g_{rs}(\eta_1^s - \eta_2^s)$.

We note that (1) $\Rightarrow$ (4) *holds for general yokes.*

PROOF. We prove these equivalences in the following manner: (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3) $\Leftrightarrow$ (4), (5) $\Rightarrow$ (3) and (1) $\Rightarrow$ (5). Proposition 3.2 shows that (1) is equivalent to $D_f(\eta_1,\eta_2) = D_{f^*}(\eta_1,\eta_2)$, and so by Proposition 3.1, (1) $\Leftrightarrow$ (2). By Proposition 3.1 and the definition of f^*, it follows that (2) is equivalent to

$$\eta^*_s = L_{rs}\eta^r + K_s \tag{3.6}$$

where (L_{rs}) is a nonsingular $n \times n$ matrix and $(K_s) \in \mathbb{R}^n$. Since $\partial_r\eta^*_s = g_{rs}$, (2) $\Leftrightarrow$ (3) and, in fact, $L_{rs} = g_{rs}$. For the divergence parameter η we have $T_{rst} = \partial_r g_{st}$; this shows (3) $\Leftrightarrow$ (4). Since $\partial_{1r}\partial_{1s}D_f(\eta_1,\eta_2) = g_{rs}$ is a function of η_1 alone, it is easily checked that g_{rs} in the right-hand side of (5) must be constant. Hence, (5) $\Rightarrow$ (3). Now assume (1) holds. Then (3.6) holds with $L_{rs} = g_{rs}$ and we must have

$$(\eta^*_{1r} - \eta^*_{2r}) = g_{rs}(\eta_1^s - \eta_2^s). \tag{3.7}$$

Using (2.8) to write the f-divergence in terms of its potential functions, (1) implies

$$2D_f(\eta_1,\eta_2) = (\psi(\eta_1) + \phi(\eta^*_2) - \eta_1^r\eta^*_{2r}) + (\psi(\eta_2) + \phi(\eta^*_1) - \eta_2^r\eta^*_{1r}). \tag{3.8}$$

Using the identity $\psi(\eta) + \phi(\eta^*) - \eta^r\eta^*_r = 0$, equation (3.8) can be rewritten as

$$2D_f(\eta_1,\eta_2) = (\eta_1^r - \eta_2^r)(\eta^*_{1r} - \eta^*_{2r}). \tag{3.9}$$

Equations (3.7) and (3.9) give (5) and the proposition is proved. □

4. Submanifolds and auxiliary manifolds

We have considered the geometry induced by a divergence on an n-dimensional manifold S. Next, we consider a smooth submanifold $M \subset S$ and let $u : M \mapsto \mathcal{U} \subset \mathbb{R}^m$ be a parameterization for M. The tangent space T_pM of M at p is spanned by $\{\partial_1, \ldots, \partial_m\}$ where $\partial_a = \partial/\partial u^a$. This tangent space is a subspace of T_pS since $\partial_a = \partial\mu^i/\partial u^a \partial_i$. Notice that we have dropped the notational distinction between vectors and vector fields. A geometric structure on M can be induced from S in a natural manner so that M is not only a Riemannian manifold but a statistical manifold as well. Since $T_pM \subset T_pS$, $\langle \cdot, \cdot \rangle$ and $T(\cdot, \cdot, \cdot)$ are defined on T_pM and the corresponding quantities on M are simply defined by restriction.

We shall assume that there exists a tubular neighborhood U_M around M. This tubular neighborhood can be thought of as a collection of $(n-m)$-dimensional auxiliary manifolds $A(p)$ (called, ancillary manifolds in Amari (1985)) at each point $p \in M$, $U_M = \cup_{p \in M} A(p)$. In order for U_M to be a tubular neighborhood, these auxiliary manifolds must be combined in a smooth manner such that there exists a parameterization $u = (u^1, \ldots, u^m)'$ on M and parameterization $w = (w^1, \ldots, w^n)'$ on U_M such that $w^\alpha = u^a$ for $a = \alpha$ where $\alpha = 1, \ldots, n$ and $a = 1, \ldots, m$. The other $n - m$ components of w will be denoted v^κ for $\kappa = m+1, \ldots, n$. In brief, we write this relationship between w, u and v as $(w^\alpha) = (u^a, v^\kappa)$. The parameter v is defined so that $(u^a, 0)$ names a point on the manifold M and if p_0 is a fixed point in M and $u_0 = u(p_0)$, then $w = (u_0^a, v^\kappa)$ names a point in $A(p_0)$. The parameter v is a coordinate chart for each $A(p)$.

For this paper we shall only consider auxiliary manifolds A that are orthogonal to M; that is, T_pM is the orthogonal complement of $T_pA(p)$ in T_pS for each $p \in M$. We let $\partial_a = \partial/\partial u^a$, $a = 1, \ldots, m$ be the natural basis associated with the parameter u and $\partial_\kappa = \partial/\partial v^\kappa$, $\kappa = m, \ldots, n$ be the natural basis for the parameter v so that $\langle \partial_a, \partial_\kappa \rangle = 0$ at all points on M. An important tensor defined on the auxiliary submanifolds is the imbedding curvature which has components defined by

$$H_{\kappa\lambda a} = \langle \nabla_{\partial_\kappa} \partial_\lambda, \partial_a \rangle. \tag{4.1}$$

We shall be most interested in the curvature of $A(p)$ evaluated on the submanifold M. It will be useful to express this tensor using another parameterization μ. We write $\partial_a = U_a^i \partial_i$ and $\partial_\lambda = V_\lambda^i \partial_i$ where $\partial_i = \partial/\partial\mu^i$, $U_a^i = \partial\mu^i/\partial u^a$ and $V_\lambda^i = \partial\mu^i/\partial v^\lambda$. Notice that U_a^i depends on how the submanifold M is imbedded in S while V_λ^i depends on the auxiliary submanifolds. Making the appropriate substitutions in (4.1) we have

$$H_{\kappa\lambda a} = V_\kappa^i (\partial_i V_\lambda^j) U_a^k g_{jk} + V_\kappa^i V_\lambda^j U_a^k \Gamma_{ijk}. \tag{4.2}$$

The following contraction of $H_{\kappa\lambda a}$, called the square of the auxiliary imbedding curvature, will be used

$$(H^2)_{ab} = H_{\kappa\lambda a} H_{\kappa'\lambda' b} g^{\kappa\kappa'} g^{\lambda\lambda'} \tag{4.3}$$

where $g^{\lambda\lambda'}$ is the $(n-m)\times(n-m)$ matrix formed from the lower right corner of the $n\times n$ matrix $g^{\alpha\alpha'}$, the matrix inverse of $g_{\alpha\alpha'}$. The simplest contraction of $H_{\kappa\lambda a}$ is the trace

$$\operatorname{tr}(H)_a = H_{\kappa\lambda a}g^{\kappa\lambda}. \tag{4.4}$$

This tensor will also be considered in the sequel.

Before calculating $A^f(p)$ explicitly, we discuss estimation and auxiliary submanifolds. We shall assume that the vector of observations y can be mapped into the manifold S; this is typically done using the expectation parameter so that we choose an element $p(y)$ in S having expectation parameter y. We also assume that the true distribution of Y belongs to the submanifold M. Recall that the points of S and M are equivalence classes of probability distributions, so by the assumption that y has distribution in M, we mean that the distribution of Y belongs to one of the equivalence classes in M. Let $p(y)$ be the point in S corresponding to y and let $p_0\in M$ contain the true distribution of Y. One way to estimate p_0 is to find the point in M that is nearest, in some sense, to $p(y)$. How we estimate p_0 will depend, then, on how we measure the "distance" between points in S. Hence, to each f-divergence there will be an estimator called the minimum f-divergence estimator or f-estimator for short. When S is defined from a linear exponential family, then the maximum likelihood estimator belongs to the family of f-estimators.

Let $\overset{f}{\hat{p}}$ be the minimum f-divergence estimate for $p(y)$; i.e.,

$$D_f\bigl(p(y),\overset{f}{\hat{p}}\bigr) = \min_{p\in M} D_f(p(y),p). \tag{4.5}$$

Notice that we can also define a minimum divergence estimator by interchanging p and $p(y)$ in (4.5). By Proposition 3.2, we see that this is just the f^*-estimator. To calculate $A^f(p)$, we need the following property for divergence functions.

PROPOSITION 4.1. *If $\hat{p}$ is a minimum divergence estimator not on the boundary of M and $m<n$, then we must have $\vec{v}(\hat{p},p(y))\perp T_{\hat{p}}M$.*

The proof of Proposition 4.1 follows from Theorem 3.8 in Amari (1985) and noting that Amari's result holds when S is defined using a quasi-likelihood function.

We can now define the f-auxiliary manifold $A^f(p)$ at $p\in M$,

$$A^f(p) = \overset{f}{\exp}_p(\mathcal{D}\cap(T_pM)^{\perp})$$

where $\mathcal{D}$ is the domain of $\overset{f}{\exp}_p$ and $(T_pM)^{\perp}$ is the orthogonal complement of T_pM in T_pS. In light of Proposition 4.1, $A^f(p)$ is almost the set of all points in S whose image under the f-estimator is p. The auxiliary submanifolds may intersect, so that some auxiliary submanifolds $A^f(p)$ may contain points that do not provide p as the f-estimate. This is because the restriction that the residual vector $\overset{f}{\vec{v}}(p(y),p)$

be orthogonal to T_pM is necessary but not sufficient (once again, provided p is not on the boundary of M and $m < n$). Locally, that is in the tubular neighborhood U_M, the auxiliary submanifolds do not intersect and this condition is sufficient. In practice, we generally cannot assume that $p(y) \in U_M$ and extra care must be exercised in finding the f-estimate. The f-auxiliary submanifold is particularly easy to express in the f-divergence parameter η

$$\eta\left(A^f(p)\right) = \{\eta : \eta = \eta(p) + t_\kappa v^\kappa\} \tag{4.6}$$

where $t_\kappa = (t^1_\kappa, \ldots, t^n_\kappa)'$, $\vec{t}_\kappa = t^r_\kappa \partial_r$ and $v^\kappa \vec{t}_\kappa \in \mathcal{D} \cap (T_pM)^\perp$. For a tubular neighborhood, we can define vector fields $\vec{t}_\kappa$ on a neighborhood around p such that $\vec{t}_\kappa(q)$ span $(T_qM)^\perp$ for all q in this neighborhood. If $u = (u^1, \ldots, u^m)'$ is parameterization for M, then $(w^\alpha) = (u^a, v^\kappa)$ where v is defined in (4.6) parameterizes a tubular neighborhood U_M of M. Notice that the f-auxiliary submanifolds are orthogonal to M; i.e., $\langle \partial_a, \partial_\kappa \rangle = 0$ on M.

In practice, there will often be a large selection of f-divergences compatible with the metric. We will assume there exists a unique f_0 so that the f_0-estimator is third order efficient. Further details appear in the next section. Consider the statistical manifold $\left(S, \langle\cdot,\cdot\rangle, \overset{f_0}{T}(\cdot,\cdot,\cdot)\right)$, where $\overset{f_0}{T}(\cdot,\cdot,\cdot)$ is the skewness tensor defined by the divergence $D_{f_0}(\cdot,\cdot)$. From the definition of $A^{f_0}(p)$ it is clear that the f_0-imbedding curvature of $A^{f_0}(p)$, call it $\overset{f_0}{H}{}^{f_0}_{\kappa\lambda a}$, is zero on M for all κ, λ and a. For the α-connections, Lauritzen (1987) calls such a manifold α-geodesic. Following this terminology, we say that $A^{f_0}(p)$ is f_0-geodesic. Any f-geodesic submanifold imbedded in an f-flat manifold is f-flat. The converse, however, is not true. The f-geodesic submanifold $A^f(p)$ is f-flat and by (2.7) f^*-flat, but $A^f(p)$ is not f^*-geodesic. The preceding two statements follow from the relationship between the imbedding curvature and Riemannian curvature expressed by the Gauss equation for dual connections (Vos (1989)).

Typically, f_0 is unknown, so we cannot define $A^{f_0}(p)$. One example where this occurs is when we model some transformation of the data and it is assumed that the "correct" transformation is a member of some family, but we do not know which one. Rather than use $D_{f_0}(\cdot,\cdot)$, we use $D_f(\cdot,\cdot)$ and define auxiliary submanifolds $A^f(p)$ at each $p \in M$. Since the optimal auxiliary submanifolds $A^{f_0}(p)$ have zero imbedding curvature in the f_0-connection, the curvature of $A^f(p)$ in the f_0-connection will give us a measure of how much the f-estimator differs from the f_0-estimator. We shall use $\overset{f_0}{H}{}^f_{\kappa\lambda a}$ to indicate the f_0-imbedding curvature of A^f, $(\overset{f_0}{H}{}^f)^2_{ab}$ to indicate its square, and $\mathrm{tr}(\overset{f_0}{H}{}^f)_a$ to indicate its trace. The f_0-imbedding curvature measures the curvature of A^f only at the point of intersection with M. Since the f-estimator will depend on how A^f behaves through out S (or at least U_M), one might think that $\overset{f_0}{H}{}^f_{\kappa\lambda a}$ would be inadequate to compare auxiliary manifolds. Recall, however, that $A^f(p)$ is f-geodesic and so it has zero imbedding curvature everywhere, not just at its point of intersection with M. It is reasonable to expect the f_0-imbedding curvature of $A^f(p)$ to be approximately constant, especially if f is near f_0. This interpretation of $\overset{f_0}{H}{}^f_{\kappa\lambda a}$ has relied solely on the

geometry of the f- and f_0-estimators. In the next section, we show that $(\overset{f_0}{H}{}^f)^2_{ab}$ measures loss of efficiency and $\mathrm{tr}(\overset{f_0}{H}{}^f)_a$ measures bias under repeated sampling.

We are now ready to obtain explicit formulae for $\overset{f_0}{H}{}^f_{\kappa\lambda a}$. From (4.2) and the fact that $\overset{f}{H}{}^f_{\kappa\lambda a} = 0$, we see that

$$V^i_\kappa(\partial_i V^j_\lambda)U^k_a g_{jk} = -V^i_\kappa V^j_\lambda U^k_a \overset{f}{\Gamma}_{ijk}. \tag{4.7}$$

Substituting (4.7) into (4.2) we have

$$\begin{aligned}\overset{f_0}{H}{}^f_{\kappa\lambda a} &= V^i_\kappa(\partial_i V^j_\lambda)U^k_a g_{jk} + V^i_\kappa V^j_\lambda U^k_a \overset{f_0}{\Gamma}_{ijk} \\ &= V^i_\kappa V^j_\lambda U^k_a \big(\overset{f_0}{\Gamma} - \overset{f}{\Gamma}\big)_{ijk}.\end{aligned} \tag{4.8}$$

Substituting (4.8) into (4.3) and (4.4) we find

$$\big(\overset{f_0}{H}{}^f\big)^2_{ab} = V^i_\kappa V^j_\lambda \big(\overset{f_0}{\Gamma} - \overset{f}{\Gamma}\big)_{ijk}\big(\overset{f_0}{\Gamma} - \overset{f}{\Gamma}\big)_{i'j'k'} V^{i'}_{\kappa'} V^{j'}_{\lambda'} g^{\kappa\kappa'} g^{\lambda\lambda'} U^k_a U^{k'}_b, \tag{4.9}$$

$$\mathrm{tr}\big(\overset{f_0}{H}{}^f\big)_a = \big(\overset{f_0}{\Gamma} - \overset{f}{\Gamma}\big)_{ijk} V^i_\kappa V^j_\lambda g^{\kappa\lambda} U^k_a. \tag{4.10}$$

One further simplification can be made in (4.9). By the orthogonality of ∂_a and ∂_κ, we have that $V^i_\kappa V^{i'}_{\kappa'} g^{\kappa\kappa'} = \bar{g}^{ii'}$ where

$$\bar{g}^{ii'} = g^{ii'} - U^i_a U^{i'}_{a'} g^{aa'}. \tag{4.11}$$

In many applications m is quite small while $n-m$ is large and V^i_κ is computationally non-trivial. Equation (4.11) shows that we need not calculate V^i_κ to find $V^i_\kappa V^{i'}_{\kappa'} g^{\kappa\kappa'}$, $(\overset{f_0}{H}{}^f)^2_{ab}$, or $\mathrm{tr}(\overset{f_0}{H}{}^f)_a$. Equations (4.9) and (4.10) now become

$$\big(\overset{f_0}{H}{}^f\big)^2_{ab} = \big(\overset{f_0}{\Gamma} - \overset{f}{\Gamma}\big)_{ijk}\big(\overset{f_0}{\Gamma} - \overset{f}{\Gamma}\big)_{i'j'k'} \bar{g}^{ii'}\bar{g}^{jj'} U^k_a U^{k'}_b, \tag{4.12}$$

$$\mathrm{tr}\big(\overset{f_0}{H}{}^f\big)_a = \big(\overset{f_0}{\Gamma} - \overset{f}{\Gamma}\big)_{ijk} \bar{g}^{ij} U^k_a. \tag{4.13}$$

Equation (4.12) ((4.13)) shows that the square (trace) of the imbedding curvature for the auxiliary submanifold of any divergence estimator is a quadratic (linear) form in the difference between the components of two connections.

5. Asymptotic considerations

In the previous section we showed how the imbedding curvature $\overset{f_0}{H}{}^f_{\kappa\lambda a}$ can be used to compare estimators with auxiliary submanifolds orthogonal to M. In this section, we interpret $\overset{f_0}{H}{}^f_{\kappa\lambda a}$ when the data y is obtained under repeated sampling.

That is, we assume $y = (y_1 + \cdots + y_N)/N$ so that $\tilde{Y} = \sqrt{N}(Y - \mu) = O_p(1)$ where $\mu = E(Y_I)$, $I = 1, \ldots, N$. We have assumed that $E(\tilde{Y}^i) = 0$ and $\operatorname{cov}(\tilde{Y}^i, \tilde{Y}^j) = g^{ij}$. For first order asymptotic calculations it is enough to assume

$$E(\tilde{Y}^i) = O(N^{-1/2}) \quad \text{and} \quad \operatorname{cov}(\tilde{Y}^i, \tilde{Y}^j) = g^{ij} + O(N^{-1/2}).$$

Vos (1991) shows that all f-estimators are first order efficient. This also follows for exponential families from Theorem 5.2 in Amari ((1985), p. 130) which says that all estimators having auxiliary families orthogonal to M are first order efficient. In order to interpret $\overset{f_0}{H}{}^{f}_{\kappa\lambda a}$, which measures the difference between the f_0- and f-estimators, we will need to consider higher order asymptotic calculations and make further assumptions. We make these assumptions in terms of $\tilde{\eta}^r = \sqrt{N}(f_0^r(Y) - f_0^r(\mu))$:

$$\begin{aligned} &E(\tilde{\eta}^r) = O(N^{-3/2}), \qquad \operatorname{cov}(\tilde{\eta}^r, \tilde{\eta}^s) = g^{rs} + O(N^{-3/2}), \\ &\operatorname{cum}(\tilde{\eta}^r, \tilde{\eta}^s, \tilde{\eta}^t) = N^{-1/2}\overset{f_0}{T}{}^{rst} + O(N^{-3/2}), \\ &\operatorname{cum}(\tilde{\eta}^r, \tilde{\eta}^s, \tilde{\eta}^t, \tilde{\eta}^u) = N^{-1}\overset{f_0}{S}{}^{rstu} + O(N^{-3/2}) \end{aligned} \tag{5.1}$$

where $\overset{f_0}{T}{}^{rst} = \partial^r\partial^s\partial^t\overset{f_0}{\phi}(\eta^*)$ and $\overset{f_0}{S}{}^{rstu} = \partial^r\partial^s\partial^t\partial^u\phi(\eta^*)$. An easy calculation shows that $\overset{f_0}{T}{}^{rst}$ is the contravariant version of $\overset{f_0}{T}_{rst}$ defined in (2.15b) for the f_0-divergence. The assumptions have been strengthened in (5.1) by specifying two further cumulants and requiring a closer approximation to all four cumulants. In practical terms, (5.1) says that the functional relationship between the mean and variance is better described using the transformed data $f_0(Y)$ and the parameter η. The calculations that we do here are identical to those in Amari ((1985), Section 4.4) for exponential families, so we shall not repeat all the details. By following the steps in Amari (1985), one finds that the exponential family assumptions can be replaced by those in (5.1). For our applications, these assumptions can be weakened slightly; the $O(N^{-3/2})$ term for the covariance and third order cumulant can be replaced with $O(N^{-1})$. For the main result, Proposition 5.1, $\overset{f_0}{S}{}^{rstu}$ cancels and, therefore, need not be specified.

Let $w = (u, v)$ be a parameterization for the tubular neighborhood U_M defined by the f-estimator and let $w_0 = (u_0, 0)$ be the w coordinates of the "true" point in M. We take $\hat{\eta} = \eta(Y)$ and let $\hat{w} = (\hat{u}, \hat{v})$ be the corresponding w coordinate for $\hat{\eta}$. From Section 5.2 of Amari (1985), we see that $E(\overset{f}{\hat{u}}{}^a - u_0^a) = b^a(u_0) + O(N^{-3/2})$ where $b^a(u_0) = -\big(\overset{f_0}{\Gamma}{}^a_{cd}g^{cd} + \operatorname{tr}(\overset{f_0}{H}{}^f)_{a'}g^{a'a}\big)/2N$ is the second order bias term. If we had used the f_0-estimator, then the second order bias terms would simply be $-\overset{f_0}{\Gamma}{}^a_{cd}g^{cd}/2N$ since $\operatorname{tr}(\overset{f_0}{H}{}^{f_0})_a = 0$. Thus, $\operatorname{tr}(\overset{f_0}{H}{}^f_a)$ measures the change in the second order bias term resulting from using the f-estimator rather than the f_0-estimator. Notice that the f-estimator may, in fact, reduce the second order bias term.

Consider next the bias corrected f-divergence estimator $\overset{f}{\hat{u}}{}^\star = \overset{f}{\hat{u}} - b(\overset{f}{\hat{u}})$. Following the steps found in Amari ((1985), pp. 132–133), we can find the Edgeworth

expansion for $\overset{f}{\tilde{u}}{}^{\star} = \sqrt{N}(\overset{f}{\hat{u}}{}^{\star} - u)$ and thereby obtain the following proposition analogous to Theorem 5.4 in Amari ((1985), p. 133).

PROPOSITION 5.1. *Under the assumptions found in* (5.1), *the mean square error of the bias corrected* f*-estimator* $\overset{f}{\tilde{u}}{}^{\star}$ *is*

$$E(\overset{f}{\tilde{u}}{}^{\star a}\overset{f}{\tilde{u}}{}^{\star b}) = g^{ab} + \frac{1}{2N}\left\{(\overset{f_0}{\Gamma})^{2ab} + 2(\overset{f_0}{H}{}^{*}_{M})^{2ab} + (\overset{f_0}{H}{}^{f})^{2ab}\right\} + O(N^{-3/2})$$

where $(\overset{f_0}{\Gamma})^{2ab}$, $(\overset{f_0}{H}{}^{*}_{M})^{2ab}$ *and* $(\overset{f_0}{H}{}^{f})^{2ab}$ *are contravariant versions of*

$$(5.2)\qquad \begin{aligned} (\overset{f_0}{\Gamma})^{2}_{ab} &= \overset{f_0}{\Gamma}_{cda}\overset{f_0}{\Gamma}_{efb}g^{ce}g^{df}, \qquad (\overset{f_0}{H}{}^{*}_{M})^{2}_{ab} = \overset{f_0^*}{H}_{ac\kappa}\overset{f_0^*}{H}_{bd\lambda}g^{cd}g^{\kappa\lambda}, \\ (\overset{f_0}{H}{}^{f})^{2}_{ab} &= \overset{f_0}{H}{}^{f}_{\kappa\lambda}\overset{f_0}{H}{}^{f}_{\kappa'\lambda'}g^{\kappa\kappa'}g^{\lambda\lambda'}. \end{aligned}$$

As Amari points out, the $O(N^{-1})$ terms given in (5.2) allow the following interpretations. The first term is one half the square of the f-connection on M. This is not a tensor and depends on the parameterization of M, but it is the same for all estimators. The second term is the square of the f^*-curvature of M; it is a tensor and so depends only on the geometric properties of M and is the same for all estimators. The third term is one half the square of the f_0-curvature of A^f discussed in Section 4. We see then that $(\overset{f_0}{H}{}^{f})^{2}_{ab}$ measures the third order efficiency not recovered by the bias corrected f-estimator.

6. The power family of divergences

In applications, it may be more practical to restrict our attention from the set of all diffeomorphisms compatible with the metric to an indexed family. An analogous situation occurs when searching for an appropriate data transformation. In this case, the family of power transformations is often useful. Correspondingly, we restrict f to the power family of diffeomorphisms

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

Notice that the transformations $f(\cdot\,;\lambda)$ act componentwise; that is, $f^r(\mu;\lambda)$ is a function of μ^i for $i = r$ alone. In order for the transformations $f(\cdot\,;\lambda)$ to be diffeomorphisms, we require that $\mu^i \neq 0$ for any i. Typically we take the domain of $f(\cdot\,;\lambda)$ to be $\mathcal{M} = (0, B)^n$ where $0 < B \leq \infty$. Sometimes the power transformation family is defined so that it is continuous in λ,

$$\underline{f}(\mu;\lambda) = \begin{cases} \dfrac{\mu^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log(\mu) & \lambda = 0. \end{cases}$$

Notice that for fixed λ, $[f(\mu;\lambda)] = [\underline{f}(\mu;\lambda)]$ and so, by Proposition 3.1, the divergence and geometric structures are the same for either version of the power family of transformations. To distinguish these divergences from general f-divergences we shall use the terminology λ-divergence and notation such as $\overset{\lambda}{\nabla}$, $\overset{\lambda}{H}$, etc. We shall see that the geometry defined by the power divergence is closely related to Amari's α-geometry (Amari (1985)).

We shall restrict our attention to covariance matrices $V(\mu)$ belonging to $\mathcal{V} = \{V(\mu) : V(\mu) = [\text{diag}(\mu)]^d, d \in \mathbb{R}\}$ where $\text{diag}(\mu)$ is an $n \times n$ diagonal matrix with diagonal $\mu \in \mathcal{M}$. Although there are more general covariance matrices, many important covariance structures are contained in $\mathcal{V}$. When $d = 0, 1, 2$ and 3, $V(\mu)$ corresponds to the second order moment structure for the normal, Poisson, gamma and inverse Gaussian distributions, respectively. If $\lambda \neq 0$, then $\eta = \mu^\lambda$ so that $\partial_i \eta^r = \lambda(\mu^i)^{\lambda-1}$ if $i = r$, zero otherwise, and $\partial_i \partial_j \eta^r = \lambda(\lambda-1)(\mu^i)^{\lambda-2}$ if $i = j = r$, zero otherwise. We also have

$$g_{rs} = \begin{cases} \lambda^{-2}(\mu^i)^{2-2\lambda-d} & r = s = i \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

Equation (6.1) follows from $g_{rs} = (\partial\mu^i/\partial\eta^r)g_{ij}(\partial\mu^j/\partial\eta^s)$ and $g_{ij} = (\mu^i)^{-d}$ if $i = j$. Making the above substitutions into (2.3), we obtain

$$\overset{\lambda}{\Gamma}_{ijk} = \begin{cases} (\lambda-1)(\mu^i)^{-(d+1)} & i = j = k \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

Since $\partial_i g_{jk} = -d(\mu^i)^{-(d+1)}$, we have, by definition of the dual connection, that

$$\overset{\lambda}{\Gamma}^*_{ijk} = \begin{cases} (1-\lambda-d)(\mu^i)^{-(d+1)} & i = j = k \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

It is easily checked that (6.2) and (6.3) hold for $\lambda = 0$ as well. From (6.2), (6.3), and $\overset{\lambda}{T}_{ijk} = \overset{\lambda}{\Gamma}_{ijk} - \overset{\lambda}{\Gamma}^*_{ijk}$ we see that $\overset{\lambda}{T}_{ijk} = (d + 2(\lambda-1))(\mu^i)^{-(d+1)}$ for $i = j = k$ and $\overset{\lambda}{T}_{ijk} = 0$ otherwise. The covariant version of this tensor is

$$\overset{\lambda}{T}{}^{ijk} = \begin{cases} (d + 2(\lambda-1))(\mu^i)^{2d-1} & i = j = k \\ 0 & \text{otherwise.} \end{cases} \tag{6.4}$$

The λ-divergence can be easily expressed when $V \in \mathcal{V}$. For each real d, we can define the λ-divergence $\bar{D}_{\lambda,d}(\mu_1, \mu_2)$ by

$$\bar{D}_{\lambda,d}(\mu_1, \mu_2) = \sum \frac{\mu_1^{2-d} - \mu_2^{2-d} + \dfrac{2-d}{\lambda}\mu_2^{2-d}\{1 - (\mu_1/\mu_2)^\lambda\}}{(2-d-\lambda)(2-d)} \tag{6.5}$$

provided $\lambda \neq 0$, $d \neq 2$ and $\lambda + d \neq 2$. We use the notation $\bar{D}_{\lambda,d}$ rather than $D_{\lambda,d}$ in (6.5) because μ is the mean parameter which is for $\lambda \neq 1$ not the divergence parameter. We reserve D for the divergence defined directly on the manifold or in

terms of the divergence parameter so that $\bar{D}_{\lambda,d}(\mu_1,\mu_2) = D_{1,d'}(f(\mu_1;\lambda), f(\mu_2;\lambda))$ for suitably chosen d'. In (6.5), raising a vector to a power and division 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

$$\bar{D}_{\lambda,2-\lambda}(\mu_1,\mu_2) = \sum \frac{\mu_2^\lambda - \mu_1^\lambda + \lambda\mu_1^\lambda \log(\mu_1/\mu_2)}{\lambda^2} \tag{6.6}$$

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

$$\bar{D}_{\lambda,2}(\mu_1,\mu_2) = \sum \frac{\lambda \log(\mu_2/\mu_1) + (\mu_1/\mu_2)^\lambda - 1}{\lambda^2}. \tag{6.7}$$

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

$$\bar{D}_{0,d}(\mu_1,\mu_2) = \sum \frac{\mu_1^{2-d} - \mu_2^{2-d} + (2-d)\mu_2^{2-d}\log(\mu_2/\mu_1)}{(2-d)^2}. \tag{6.8}$$

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

$$\bar{D}_{0,2}(\mu_1,\mu_2) = \sum \frac{(\log\mu_1 - \log\mu_2)^2}{2}. \tag{6.9}$$

For the special case when $d = 1$, $D_{\lambda,d}$ is closely related to the power divergence of Read and Cressie (1988). We note that the dual connections given in (6.2) and (6.3) can be obtained directly from the above divergences using (2.4) and (2.5) for general contrast functionals.

Using Proposition 3.1, equations (6.2) and (6.3), and the fact that $D_{\lambda,d}$ can be defined for all λ and d (using (6.5)–(6.9)), we have the following result.

PROPOSITION 6.1. *Let S be $(0,B)^n$ with metric given by the covariance structure $V \in \mathcal{V}$. For every $f = f(\cdot,\lambda)$ in the power family of transformations, the f-divergence $D_{\lambda,d}(\cdot,\cdot)$ exists and has dual $D_{\lambda^*,d}(\cdot,\cdot)$ where $\lambda^* = (2-d)-\lambda$. There exists a unique power transformation that makes the divergence symmetric; namely, the transformation with $\lambda = 1 - d/2$.*

Next, we shall compare the λ-connections and the α-connections. Suppose we have an n-dimensional exponential family with covariance structure $V \in \mathcal{V}$ and let $\mu = (\mu^i)$ be the expectation parameter so that $\overset{(1)}{\Gamma}_{ijk} = \partial_i g_{jk}$ and $\overset{(-1)}{\Gamma}_{ijk} = 0$. From (2.29) in Amari ((1985), p. 40) we see that

$$\overset{\alpha}{\Gamma}_{ijk} = \frac{(1+\alpha)}{2}\partial_i g_{jk}.$$

(We note that equation (2.29) holds for any parameterization; Amari (1985) uses Γ_{ijk} for the components of a connection using the natural parameter while we

use the same symbol for the components of a connection using the expectation parameter. Hence, the -1-connection is represented by $\overset{(-1)}{\Gamma}_{ijk} = 0$ in this paper and by $\overset{(-1)}{\Gamma}_{ijk} = \partial_i g_{jk}$ in Amari (1985).)

Since g_{jk} are the components of the inverse of the covariance structure and $V \in \mathcal{V}$, we have

$$\overset{\alpha}{\Gamma}_{ijk} = \begin{cases} -d\dfrac{(1+\alpha)}{2}(\mu^i)^{-(d+1)} & i = j = k \\ 0 & \text{otherwise.} \end{cases} \tag{6.10}$$

Comparing (6.2) and (6.10), we see that $\overset{\lambda}{\nabla} = \overset{\alpha}{\nabla}$ provided $(1-\lambda) = d(1+\alpha)/2$ and $V \in \mathcal{V}$. Although the collection of all λ-connections is the same as the collection of all α-connections when $V \in \mathcal{V}$ and $d \neq 0$, in general, they can be quite different. The α-connections are defined by a linear combination of the dual connections on a Riemannian manifold S. The f-connections, and the λ-connections in particular, are defined from an f-divergence whose metric matrix is compatible with the metric on S. Hence, for each λ-connection there is a λ-divergence, but the same need not be true for the α-connections. When $V \in \mathcal{V}$, the relationship between the λ-connections and α-connections means that the α-divergence exists for all real α. This means the manifold S is α-flat for all real α. Notice that when $d = 0$, then $\overset{\lambda}{T}_{iii} = 2(\lambda - 1)(\mu^i)^{-1}$ while $\overset{\alpha}{T}_{iii} = 0$ for all α.

The imbedding curvature of A^λ in the λ-connection, $\overset{\lambda_0(\lambda)}{H}_{\kappa\lambda a}$, can be easily calculated when $V \in \mathcal{V}$. We place the superscript λ in parentheses because it denotes the λ-connection while the subscript λ is an index for the v parameter. Using (4.8) and (6.2) we see that

$$\overset{\lambda_0(\lambda)}{H}_{\kappa\lambda a} = (\lambda_0 - \lambda)(\mu^i)^{-(d+1)}\delta_{ij}V^j_\kappa V^j_\lambda U^j_a. \tag{6.11}$$

From (6.2), (4.12) and (4.13) we see that the square and trace of this curvature are

$$\left(\overset{\lambda_0(\lambda)}{H}\right)^2_{ab} = (\lambda_0 - \lambda)^2 \sum_{i,j}^{n} (\mu^i)^{-(d+1)}(\mu^j)^{-(d+1)}\bar{g}^{ij}\bar{g}^{ij}U^i_a U^j_b, \tag{6.12}$$

$$\operatorname{tr}\left(\overset{\lambda_0(\lambda)}{H}\right)_a = (\lambda_0 - \lambda)\sum_{i}^{n}(\mu^i)^{-(d+1)}\bar{g}^{ii}U^i_a. \tag{6.13}$$

Equations (6.12) and (6.13) show how the "distance" between power transformations can be measured by the difference in their lambda parameters. Hoaglin *et al.* (1983) obtain a similar result for the strength of data transformations.

7. An example

Bates and Watts ((1988), pp. 110–121) analyzed the data from an experiment that measured the nitrate utilization (nmol/g hr) of portions of three bean plant leaves subjected to eight levels of light (μE/m^{2}s). This same experiment was repeated on a different day. The data for both days is plotted in Fig. 1. It was expected that the nitrate utilization would be zero at zero light intensity and to approach an asymptote for increased light intensity, but no explicit functional relationship is known.

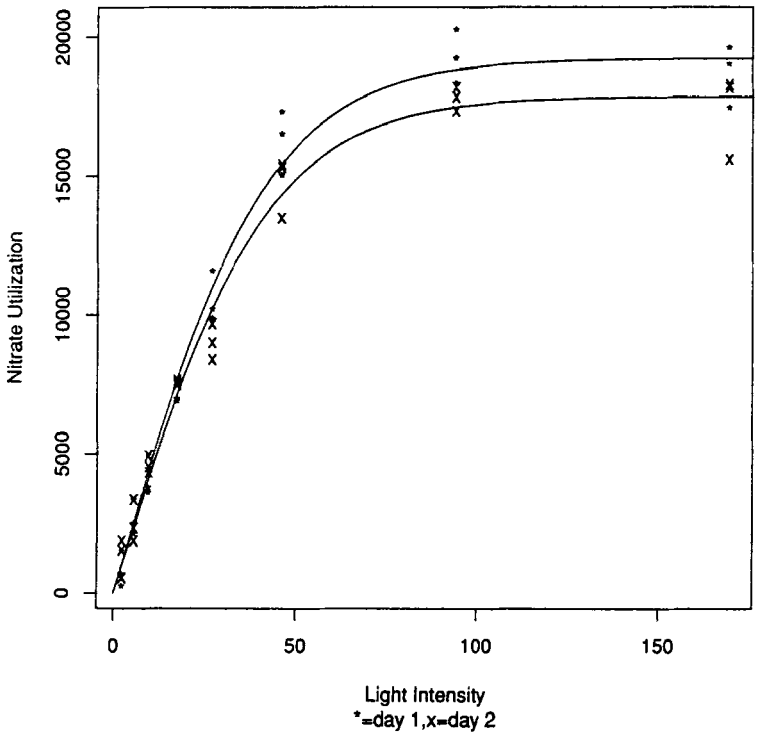

Fig. 1. Nitrate utilization for bean plants.

Bates and Watts (1988) show that the data is consistent with the assumption of homoscedastic errors and use nonlinear regression techniques to analyze the data. They failed to find a satisfactory model for the mean structure that has an asymptote. Our analysis differs both in the mean structure as a function of light intensity and the assumptions on the error distribution.

First, we consider the functional relationship between mean nitrate utilization and light intensity. Bates and Watts (1988) try the models

$$\mu = h(x,\beta) = \frac{\beta^1 x}{\beta^2 + x}, \tag{7.1}$$

$$\mu = h(x,\beta) = \beta^1(1 - e^{-\beta^2 x}) \tag{7.2}$$

where x is light intensity. These models are inadequate until the day effect and a quadratic term are added. Inspection of Fig. 1 shows that if the mean approaches an asymptote, it does so quite quickly; that is, the "bend" in the curve is a sharp one. The problem with models (7.1) and (7.2) is that they contain only "45 mph" curves while the data exhibits a sharper "30 mph" curve. A better model for this data is a function whose slope is large for small values of x and changes rapidly to small positive slope for larger values of x. One such function is the hyperbolic tangent. Since there appears to be an effect for the day the experiment was conducted, we fit the model

$$\mu = h(x, \beta) = (\beta^1 + \beta^3 x_2) \tanh((\beta^2 + \beta^4 x_2) x_1) \tag{7.3}$$

where x_1 is light intensity and x_2 is 0 for day 1 and 1 for day 2. Notice that (7.3) is not a generalized linear model since there is not transformation of μ that is linear in $\beta = (\beta^1, \beta^2, \beta^3, \beta^4)'$. Following Bates and Watts (1988), we use ordinary least squares to estimate β, i.e., we find β that minimizes $D_{1,0}(y, h(x, \beta))$. The parameter β^4 is not found significant and the analysis for the three parameter model is summarized in Table 1. The lack-of-fit analysis produces an F-statistic with value 1.44, so that the fit is reasonable. A plot of the residuals shows no irregularities. Adding a quadratic term to obtain

$$h_1(x; \beta) = (\beta^1 + \beta^3 x_2) \tanh(\beta^2 x_1 + \beta^5 (x_1)^2)$$

and estimating $\beta = (\beta^1, \beta^2, \beta^3, \beta^5)'$ show β^5 is not significant ($t = 1.27$). There appears to be little evidence suggesting that nitrate utilization does not approach an asymptote when the hyperbolic tangent models are used.

Table 1. Constant variance and zero skewness model ($d = 0, \lambda = 1$).

Parameter	Estimate	Standard error	t ratio	Correlation matrix		
β^1	19300	343	56.3	1.00		
β^2	240×10^{-4}	8.44×10^{-4}	28.4	−0.55	1.00	
β^3	−1484	407	−3.6	−0.61	0.04	1.00

Next, we consider the error distribution for these data. Estimating β by minimizing $D_{1,0}(y, h(x, \beta))$ assumes homoscedasticity and zero skewness for the data. Although the homoscedasticity assumption appears reasonable, the zero skewness assumption is unwarranted. Since the nitrate utilization cannot be negative and we have assumed constant variance, the data must be positively skewed, at least for μ near zero. For large μ the skewness may be negligible, but there are six observations at $x_1 = 2.2$ that are near zero. The average for the three observations taken on day 1 is about 1 standard deviation from zero and the average for the three observations taken on day 2 is about 1.6 standard deviations from zero. (The variance estimate comes from the replications mean square error (32 d.f.).)

Whether the skewness of these six observations can safely be ignored is unclear without further investigation.

We model the skewness by using the minimum λ-divergence estimate for constant variance, i.e., we minimize $D_{\lambda,0}(y, h(x,\beta))$. By (6.4), we see that the skewness of the transformed random variable is proportional to $2(\lambda - 1)$. Values of λ far from 1 should be used to model large skewness. Rather than attempt to estimate the skewness from the data, we can investigate the parameter estimates for several values of λ. Even for λ as small as 0, the estimates change little. The estimates for this model are summarized in Table 2. The parameter σ^2 is estimated by the minimum divergence estimate $s^2 = \min_\beta D_{0,0}(y, h(x,\beta))/45$. One normally would not choose $\lambda > 1$ since the skewness of the resulting transformed data would be larger than the original data; even so, for $\lambda \leq 1.5$, the estimates are not too different from those in Table 1.

Table 2. Constant variance and positive skewness model ($d = 0, \lambda = 0$).

Parameter	Estimate	Standard error	t ratio	Correlation matrix		
β^1	19045	353	54.0	1.00		
β^2	244×10^{-4}	9.00×10^{-4}	27.1	−0.55	1.00	
β^3	−1753	418	−4.2	−0.61	0.04	1.00

Even though there is not sufficient evidence to reject the constant variance assumption, a plot of the replication standard deviation versus the replication average shows that variance may also be modeled as an increasing function of the mean. If we take $V(Y) = \mu^d$ for $d > 0$ and use the maximum quasi-likelihood estimate, then (6.4) shows that this estimator is optimal for distributions with skewness $d\mu^{2d-1}$. In other words, changing the constant variance assumption has the added advantage of simultaneously introducing a positive skewness structure. The maximum quasi-likelihood estimates for the case $d = .5$ are given in Table 3. A plot of the replication standard deviation divided by $\mu^{.5}$ versus the replication average supports this choice of d. Again, the dispersion parameter is estimated by the minimum divergence estimator $s^2 = \min_\beta D_{1,.5}(Y, h(x,\beta))/45$.

Table 3. Non-constant variance model ($d = .5, \lambda = 1$).

Parameter	Estimate	Standard error	t ratio	Correlation matrix		
β^1	19244	415	46.4	1.00		
β^2	240×10^{-4}	9.07×10^{-4}	26.5	−0.57	1.00	
β^3	−1376	475	−2.9	−0.60	0.04	1.00

Another approach to this problem is to transform the data to homoscedasticity (see, e.g., Ruppert and Aldershof (1989)). However, after any transformation

from the power family with $\lambda > 0$, the data will still be constrained (at the transformation of the origin) so that the homoscedasticity assumption near this point is wrong. Although transformations with $\lambda \leq 0$ avoid this problem, they can often be too strong to transform to homoscedasticity. By modeling the heteroscedasticity we avoid this problem.

In conclusion, the asymptote for day 1 is around 19000 (nmol/g hr) for all three models, while the asymptote for day 2 is about 18000 for the models summarized in Tables 1 and 3, but closer to 17000 for the model in Table 2. Correspondingly, the day effect is more significant for the constant variance, positive skewness model (Table 2). On the other hand, for the nonconstant variance model (Table 3), the day effect is lessened, but still produces a significant t-value. The nominal standard errors for the estimates of the nonconstant variance model are also increased. The major conclusions, the existence of an asymptote and difference between day 1 and day 2, are the same for each model. Hence, these conclusions are relatively insensitive to perturbations in the variance and skewness structures. In picking a final model, however, there is a slight preference for the nonconstant variance model. The constant variance, zero skewness model is clearly inappropriate. The constant variance, positive skewness model may be appropriate, but the skewness structure is difficult to estimate. If we had seriously followed this approach, a model different from the one given in Table 2 would likely have been found. The model in Table 2, was chosen to check the harmful effects of wrongly assuming zero skewness. For the nonconstant variance model with $d = .5$, the variance model is easy to check and appears reasonable. This model for mean nitrate utilization is plotted along with the data in Fig. 1.

Acknowledgement

The author wishes to thank a referee for pointing out the relationship between divergence, yoke, and contrast functional and for his (her) other valuable comments.

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*, 19–94, IMS Lecture Notes-Monograph Series, Vol. 10, Hayward, California.

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

Bates, D. and Watts, D. (1988). *Nonlinear Regression Analysis and Its Applications*, Wiley, New York.

Blæsild, P. (1987). Yokes: elemental properties with statistical applications, *Geometrization of Statistical Theory, Proceedings of the GST Workshop, University of Lancaster* (ed. C. T. J. Dodson), 193–198, ULDM Publications, University of Lancaster.

Csiszar, I. (1967). On topological properties of f-divergence, *Studia Sci. Math. Hungar.*, **2**, 329–339.

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.

Hoaglin, D., Mosteller, F. and Tukey, J. (1983). *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

Lauritzen, S. L. (1987). Statistical manifolds, *Differential Geometry in Statistical Inference*, 163–216, ISM Lecture Notes-Monograph Series, Vol. 10, Hayward, California.

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

Pfanzagl, J. (1973). Asymptotic expansions related to minimum contrast estimators, *Ann. Statist.*, **1**, 993–1026.

Rao, C. R. (1987). Differential metrics in probability spaces, *Differential Geometry in Statistical Inference*, 217–240, ISM Lecture Notes-Monograph Series, Vol. 10, 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. (1989). Fundamental equations for statistical submanifolds with applications to the Bartlett correction, *Ann. Inst. Statist. Math.*, **41**, 429–450.

Vos, P. W. (1991). Minimum f-divergence estimators and quasi-likelihood functions, to appear in *Ann. Inst. Statist. Math.*