10 An elementary account of Amari's expected geometry

Frank Critchley, Paul Marriott and Mark Salmon

Differential geometry has found fruitful application in statistical inference. In particular, Amari's (1990) expected geometry is used in higher-order asymptotic analysis and in the study of sufficiency and ancillarity. However, we can see three drawbacks to the use of a differential geometric approach in econometrics and statistics more generally. First, the mathematics is unfamiliar and the terms involved can be difficult for the econometrician to appreciate fully. Secondly, their statistical meaning can be less than completely clear. Finally, the fact that, at its core, geometry is a visual subject can be obscured by the mathematical formalism required for a rigorous analysis, thereby hindering intuition. All three drawbacks apply particularly to the differential geometric concept of a non-metric affine connection.

The primary objective of this chapter is to attempt to mitigate these drawbacks in the case of Amari's expected geometric structure on a full exponential family. We aim to do this by providing an elementary account of this structure that is clearly based statistically, accessible geometrically and visually presented.

Statistically, we use three natural tools: the score function and its first two moments with respect to the true distribution. Geometrically, we are largely able to restrict attention to tensors; in particular, we are able to avoid the need formally to define an affine connection. To emphasise the visual foundation of geometric analysis we parallel the mathematical development with graphical illustrations using important examples of full exponential families. Although the analysis is not restricted to this case, we emphasise one-dimensional examples so that simple pictures can

This work has been partially supported by ESRC grant 'Geodesic Inference, Encompassing and Preferred Point Geometry in Econometrics' (Grant Number R000232270).

be used to illustrate the underlying geometrical ideas and aid intuition. It turns out that this account also sheds some new light on the choice of parameterisation as discussed by Amari (1990), extending earlier work by Bates and Watts (1980, 1981), Hougaard (1982) and Kass (1984). There are also a number of points of contact between our presentation and Firth (1993).

A key feature of our account is that all expectations and induced distributions are taken with respect to one fixed distribution, namely, that assumed to give rise to the data. This is the so-called preferred point geometrical approach developed in Critchley, Marriott and Salmon (1993, 1994), on whose results we draw where appropriate.

Our hope is that the following development will serve to broaden interest in an important and developing area. For a more formal but still readable treatment of differential geometry, see Dodson and Poston (1977). For broader accounts of the application of differential geometry to statistics, see the review chapters or monographs by Barndorff-Nielsen, Cox and Reid (1986), Kass (1987, 1989), Amari (1990) and Murray and Rice (1993).

The chapter is organised as follows. The elementary prerequisites are established in section 1. The key elements of Amari's expected geometry of general families of distributions are briefly and intuitively reviewed in section 2. In particular, his α-connections are discussed in terms of the characteristic statistical properties of their associated affine parameterisations. Section 3 contains our account of this geometry in the full exponential family case, as outlined above, and section 4 considers the effect of changing the sample size.

1 Preliminaries

1.1 The general framework

Let

$$M = \{p(x, \theta) : \theta \in \Theta\}$$

be a p-dimensional parametric family of probability (density) functions. The available data $\mathbf{x} = (x_1, \ldots, x_n)^T$ is modelled as a random sample from some unknown true distribution $p(x, \phi) \in M$. Let the parameter space Θ be an open connected subset of $\mathbf{R}^p$. The family M is regarded as a manifold, with the parameter θ playing the role of a coordinate system on it. Formally, certain regularity conditions are entailed. These are detailed in Amari (1990, p. 16).

1.2 The score function

The score function

$$s(\theta, \mathbf{x}) = \left(\frac{\partial}{\partial \theta^1} \ln p(\mathbf{x}, \theta), \ldots, \frac{\partial}{\partial \theta^p} \ln p(\mathbf{x}, \theta) \right)^T$$

is very natural to work with statistically as it contains precisely all the relevant information in the likelihood function. Integrating over Θ recovers the log-likelihood function, l, up to an additive constant which is independent of θ. This is equivalent to the likelihood up to a multiplicative positive factor which may depend on $\mathbf{x}$ but not on θ. As discussed by Cox and Hinkley (1974, p. 12), two different choices of the constant do not affect the essential likelihood information, which we refer to as the shape of the likelihood. Visually, the graph of the score function displays the shape of the likelihood in a natural and direct way. We use this to advantage later.

The score function is also a very natural tool to work with geometrically. An important concept of differential geometry is that of the tangent space. We can avoid the general abstract definition here as we have a concrete representation of this space in terms of the score function. Regarding $\mathbf{x}$ now as a random vector and following Amari (1990), we identify the tangent space TM_θ at each fixed $p(\mathbf{x}, \theta) \in M$ with the vector space of random variables spanned by

$$\{s_i(\theta, \mathbf{x}) = \frac{\partial}{\partial \theta^i} \ln p(\mathbf{x}, \theta) : i = 1, \ldots, p\}.$$

Under the regularity conditions referenced in section 2.3 of chapter 1, this vector space has dimension p, the dimension of M.

1.3 Distribution of the score vector

Naturally associated with each fixed tangent space TM_θ is the joint distribution ρ_θ^ϕ of the components of the score vector $s(\theta, \mathbf{x})$. This may be known analytically but can always, by the central limit theorem, be approximated asymptotically by the multivariate normal distribution $N_p(\mu^\phi(\theta), g^\phi(\theta))$, where

$$\mu^\phi(\theta) = \mathbf{E}_{p(x,\phi)}[s(\theta, \mathbf{x})] = n\mathbf{E}_{p(x,\phi)}[s(\theta, x)]$$

and

$$g^\phi(\theta) = \mathrm{Cov}_{p(x,\phi)}[s(\theta, \mathbf{x})] = n\mathrm{Cov}_{p(x,\phi)}[s(\theta, x)].$$

These last two quantities are statistically natural tools that we shall employ in our account of Amari's geometry. The matrix $g^{\phi}(\theta)$ is assumed to be always positive definite.

Note that, for all ϕ,

$$\mu^{\phi}(\phi) = 0 \quad \text{and} \quad g^{\phi}(\phi) = I(\phi) = ni(\phi),$$

where I and i denote the Fisher information for the sample and for a single observation, respectively.

For later use we define the random vector $\epsilon^{\phi}(\theta, \mathbf{x})$ by the decomposition

$$s(\theta, \mathbf{x}) = \mu^{\phi}(\theta) + \epsilon^{\phi}(\theta, \mathbf{x})$$

so that $\mathbf{E}_{p(x,\phi)}[\epsilon^{\phi}(\theta, \mathbf{x})]$ vanishes identically in θ and ϕ.

In the one-dimensional case there is a particularly useful graphical representation of the three tools on which our account is based. For a particular realisation of the data $\mathbf{x}$, the plot of the graph of $s(\theta, \mathbf{x})$ against θ can give great insight into the shape of the observed likelihood function. We call this graph the observed plot. Together with this we use the expected plot. This is a graph of the true mean score together with an indication of variability. We make extensive use of this graphical method for several important examples below.

1.4 Reparameterisation

So far, we have worked in a single parameterisation θ. It is important to consider what happens under a reparameterisation.

We consider reparameterisations $\theta \to \xi(\theta)$ that are smooth and invertible. Define

$$B_i^{\alpha}(\theta) = \frac{\partial \xi^{\alpha}}{\partial \theta^i} \quad \text{and} \quad \bar{B}_{\alpha}^{i}(\xi) = \frac{\partial \theta^i}{\partial \xi^{\alpha}},$$

for $1 \leq i, \alpha \leq p$. By the chain rule, the components of the score vector transform as 1-tensors. That is:

$$s_{\alpha}(\xi(\theta), \mathbf{x}) := \frac{\partial l}{\partial \xi^{\alpha}} = \sum_{i=1}^{p} \bar{B}_{\alpha}^{i}(\xi(\theta)) \frac{\partial l}{\partial \theta^i} := \sum_{i=1}^{p} \bar{B}_{\alpha}^{i}(\theta) s_i(\theta, \mathbf{x}) \tag{1}$$

for each fixed θ. This amounts to a change of basis for the vector space TM_{θ}. By linearity of expectation, the components of $\mu^{\phi}(\theta)$ are also 1-tensors. That is:

$$\mu_{\alpha}^{\xi(\phi)}(\xi(\theta)) = \sum_{i=1}^{p} \bar{B}_{\alpha}^{i}(\theta) \mu_i^{\phi}(\theta). \tag{2}$$

As covariance is a bilinear form, we see that $g^{\phi}(\theta)$ is a 2-tensor. That is, its components transform according to:

$$g^{\xi(\phi)}_{\alpha\beta}(\xi(\theta)) = \sum_{i=1}^{p}\sum_{j=1}^{p} \bar{B}^{i}_{\alpha}(\theta)\bar{B}^{j}_{\beta}(\theta)g^{\phi}_{ij}(\theta). \tag{3}$$

By symmetry, the assumption of positive definiteness, and since $g^{\phi}(\theta)$ varies smoothly with θ, $g^{\phi}(\theta)$ fulfils the requirements of a metric tensor (see Amari (1990), p. 25). It follows at once, putting $\theta = \phi$, that the Fisher information also enjoys this property.

In parallel with this tensor analysis, plotting the observed and expected plots for different parameterisations of the model can be extremely useful in conveying the effects of reparameterisation on the shape of the likelihood and the statistical properties of important statistics such as the maximum likelihood estimate (MLE). The question of parameterisation is therefore an important choice that has to be taken in statistical analysis.

2 Some elements of Amari's expected geometry

2.1 Connections

Formally, Amari's expected geometry is a triple (M, I, ∇^{+1}) in which M is a family of probability (density) functions and I the Fisher information metric tensor, as described above. The major difficulty in understanding revolves around the third component, ∇^{+1}, which is a particular non-metric affine connection. In section 3 we obtain a simple, statistical interpretation of it in the full exponential family case. Here we note certain facts concerning connections and Amari's geometry, offering intuitive explanations and descriptions where possible. For a formal treatment, see Amari (1990). We emphasise that such a treatment is not required here, as our later argument proceeds in terms of the elementary material already presented.

A connection allows us to (covariantly) differentiate tangent vectors and, more generally, tensors (see Dodson and Poston (1977), chapter 7). A connection therefore determines which curves in a manifold shall be called 'geodesic' or 'straight'. Generalising familiar Euclidean ideas, these are defined to be those curves along which the tangent vector does not change.

A metric tensor induces in a natural way an associated connection called the Levi–Civita or metric connection. In Amari's structure the Fisher information I induces the affine connection denoted by ∇^{0}. The Levi–Civita connection has the property that its geodesics are curves of

minimum length joining their endpoints. No concept of length is associated with the geodesics corresponding to non-metric connections.

Amari shows that the two connections ∇^0 and ∇^{+1} can be combined to produce an entire one-parameter family $\{\nabla^\alpha : \alpha \in \mathbf{R}\}$ of connections, called the α-connections. The most important connections statistically correspond to $\alpha = 0, \pm\frac{1}{3}, \pm 1$, as we now explain.

2.2 Choice of parameterisation

For each of Amari's connections it can happen that a parameterisation θ of M exists such that the geodesic joining the points labelled θ_1 and θ_2 simply consists of the points labelled $\{(1-\lambda)\theta_1 + \lambda\theta_2 : 0 \leq \lambda \leq 1\}$. For example, Cartesian coordinates define such a parameterisation in the Euclidean case. When this happens, M is said to be flat, such a parameterisation is called affine, and the parameters are unique up to affine equivalence. That is, any two affine parameterisations are related by a non-singular affine transformation. In the important special case of a metric connection, M is flat if and only if there exists a parameterisation θ in which the metric tensor is independent of θ.

For a connection to admit an affine parameterisation is a rather special circumstance. When it does, we may expect the affine parameterisation to have correspondingly special properties. This is indeed the case with Amari's expected geometry. When an α-connection has this property, the manifold is called α-flat and the associated parameterisations are called α-affine. Amari (1990, Theorem 5.12, p. 152), established the following characteristic features of certain α-affine parameterisations:

1. $\alpha = 1$ corresponds to the natural parameter, θ.
2. $\alpha = \frac{1}{3}$ corresponds to the normal likelihood parameter.
3. $\alpha = 0$ gives a constant asymptotic covariance of the MLE.
4. $\alpha = -\frac{1}{3}$ gives zero asymptotic skewness of the MLE.
5. $\alpha = -1$ gives zero asymptotic bias of the MLE.

These correspond to the $\delta = 0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1$ parameterisations, respectively, of Hougaard (1982), who studied the one-dimensional curved exponential family case. In any one-dimensional family an α-affine parameter exists for every α. A full exponential family, of any dimension, is always $+1$-flat and -1-flat, with the natural and mean value parameters, respectively, being affine. Amari (1990) also established the duality result that M is α-flat if and only if it is $-\alpha$-flat. This duality between ∇^α and $\nabla^{-\alpha}$ has nice mathematical properties but has not been well understood statistically.

3 The expected geometry of the full exponential family

3.1 Introduction

We restrict attention now to the full exponential family. In the natural parameterisation, θ, we have

$$p(x, \theta) = \exp\left\{\sum_{i=1}^{p} t_i(x)\theta^i - \psi(\theta)\right\}.$$

The mean value parameterisation is given by $\eta = (\eta^1, \ldots, \eta^p)$, where

$$\eta^i(\theta) = \mathbf{E}_{p(x,\theta)}[t_i(x)] = \frac{\partial \psi}{\partial \theta^i}(\theta).$$

These two parameterisations are therefore affinely equivalent if and only if ψ is a quadratic function of θ, as with the case of normal distributions with constant covariance. As we shall see, this is a very special circumstance.

In natural parameters, the score function is

$$s_i(\theta, \mathbf{x}) = n\left\{\bar{t}_i(\mathbf{x}) - \frac{\partial \psi}{\partial \theta^i}(\theta)\right\} = n\{\bar{t}_i(\mathbf{x}) - \eta^i(\theta)\}, \tag{4}$$

where $n\bar{t}_i(\mathbf{x}) = \sum_{r=1}^{n} t_i(x_r)$. From (4) we have the useful fact that the maximum likelihood estimator $\hat{\eta}^i := \eta^i(\hat{\theta}) = \bar{t}_i$. Further, the first two moments of the score function under $p(x, \phi)$ are given by

$$\mu^{\phi}{}_i(\theta) = n\left\{\frac{\partial \psi}{\partial \theta^i}(\phi) - \frac{\partial \psi}{\partial \theta^i}(\theta)\right\} = n\{\eta^i(\phi) - \eta^i(\theta)\} \tag{5}$$

$$g^{\phi}{}_{ij}(\theta) = n\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\phi) = I_{ij}(\phi). \tag{6}$$

3.2 Examples

The following one-dimensional examples are used for illustrative purposes: Poisson, normal with constant (unit) variance, exponential and Bernoulli.

Although, of course, the sample size affects the ϕ-distribution of $\bar{t}$, it enters the above equations for the score and its first two moments only as a multiplicative constant. Therefore our analysis, which is based solely on these quantities, is essentially invariant under independent repeated samples. Our third and fourth examples implicitly cover the gamma and binomial families and together, then, these examples embrace most

Table 10.1. *One-dimensional examples: Poisson, normal, exponential and Bernoulli*

	Poisson (θ) (Figure 10.1)	Normal ($\theta, 1$) (Figure 10.2)	Exponential (θ) (Figure 10.3)	Bernoulli (θ) (Figure 10.4)
$t(x)$	x	x	$-x$	x
$\psi(\theta)$	e^{θ}	$\frac{1}{2}\theta^2$	$-\ln\theta$	$\ln(1+e^{\theta})$
$s(\theta, \mathbf{x})$	$n(\bar{x}-e^{\theta})$	$n(\bar{x}-\theta)$	$n(-\bar{x}+\theta^{-1})$	$n[\bar{x}-e^{\theta}(1+e^{\theta})^{-1}]$
$\mu^{\phi}(\theta)$	$n(e^{\phi}-e^{\theta})$	$n(\phi-\theta)$	$n(-\phi^{-1}+\theta^{-1})$	$n\frac{e^{\phi}}{1+e^{\phi}}-n\frac{e^{\theta}}{1+e^{\theta}}$
$g^{\phi}(\theta)$	ne^{ϕ}	n	$n\phi^{-2}$	$ne^{\phi}(1+e^{\phi})^{-2}$
$\xi(\theta)$	$\eta(\theta)=e^{\theta}$	$\theta^{1/3}$	$\eta(\theta)=-\theta^{-1}$	$\eta(\theta)=e^{\theta}(1+e^{\theta})^{-1}$
$\bar{B}(\theta)$	ξ^{-1}	$3\xi^2$	ξ^{-2}	$[\xi(1-\xi)]^{-1}$
$s(\xi, \mathbf{x})$	$n(\bar{x}-\xi)\xi^{-1}$	$3n(\bar{x}-\xi^3)\xi^2$	$-n(\bar{x}+\xi)\xi^{-2}$	$n(\bar{x}-\xi)[\xi(1-\xi)]^{-1}$
$\mu^{\xi(\phi)}(\xi)$	$n[\xi(\phi)-\xi]\xi^{-1}$	$3n[\xi^3(\phi)-\xi^3]\xi^2$	$n[\xi(\phi)-\xi]\xi^{-2}$	$n\frac{[\xi(\phi)-\xi]}{[\xi(1-\xi)]}$
$g^{\xi(\phi)}(\xi)$	$n\xi(\phi)\xi^{-2}$	$9n\xi^4$	$n\xi(\phi)^2\xi^{-4}$	$n\frac{\xi(\phi)[1-\xi(\phi)]}{[\xi(1-\xi)]^2}$
ϕ	0	0	1	0
n	10	10	10	10

of the distributions widely used in generalised linear models (McCullagh and Nelder, 1989).

The examples are summarised algebraically in table 10.1, and are displayed visually in figures 10.1 to 10.4, respectively. For each example, for a chosen ϕ and n shown in table 10.1, we give observed and expected plots, both in the natural parameterisation θ and in a non-affinely equivalent parameterisation $\xi(\theta)$.

We take $\xi(\theta)$ to be the mean value parameter $\eta(\theta)$ except in the normal case, where we take $\xi(\theta)=\theta^{\frac{1}{3}}$. We use this last parameterisation for illustration only, even though it is not invertible at $\theta=0$. In each case, ξ is an increasing function of θ. In the expected plots, we illustrate the first two moments of the score function under the true distribution (that is, under $p(x, \phi)$) by plotting the mean ± 2 standard deviations. In the observed plots, to give some idea of sampling variability, we plot five observed score functions corresponding to the 5%, 25%, 50%, 75% and 95% points of the true distribution of $\bar{t}$ for the continuous families and the closest observable points to these in the discrete cases. Recall that these

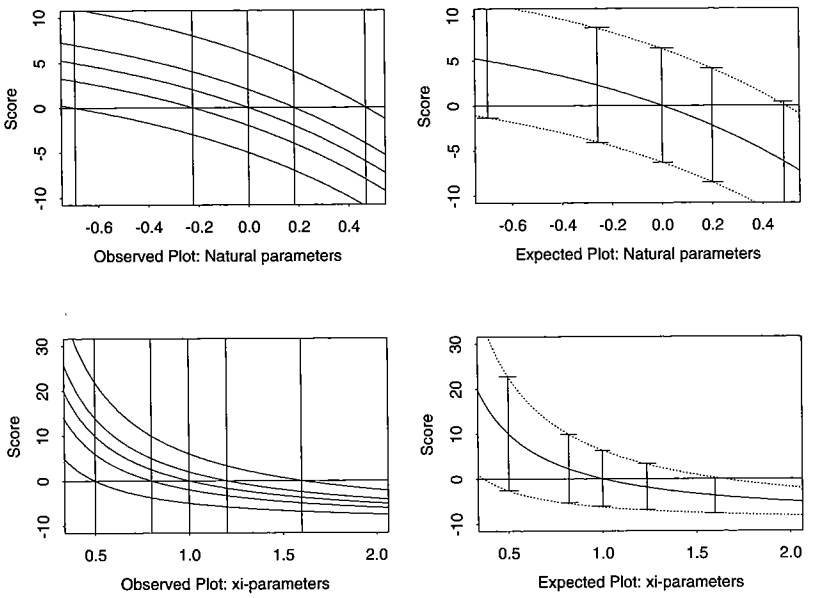

Figure 10.1 One-dimensional Poisson example

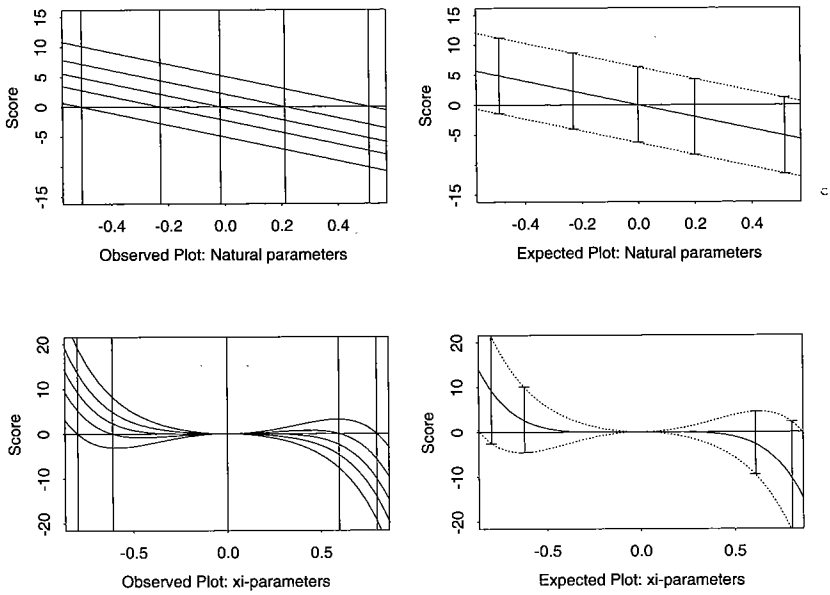

Figure 10.2 One-dimensional normal example

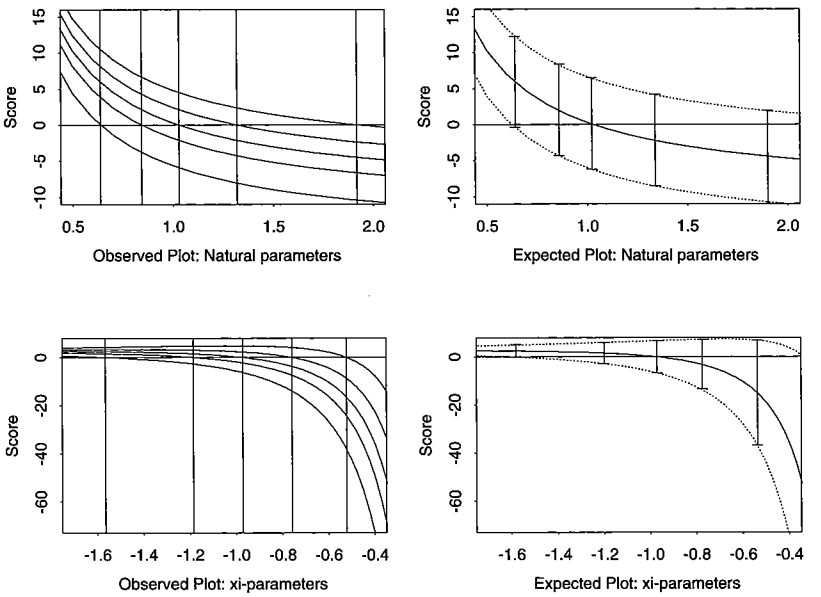

Figure 10.3 One-dimensional exponential example

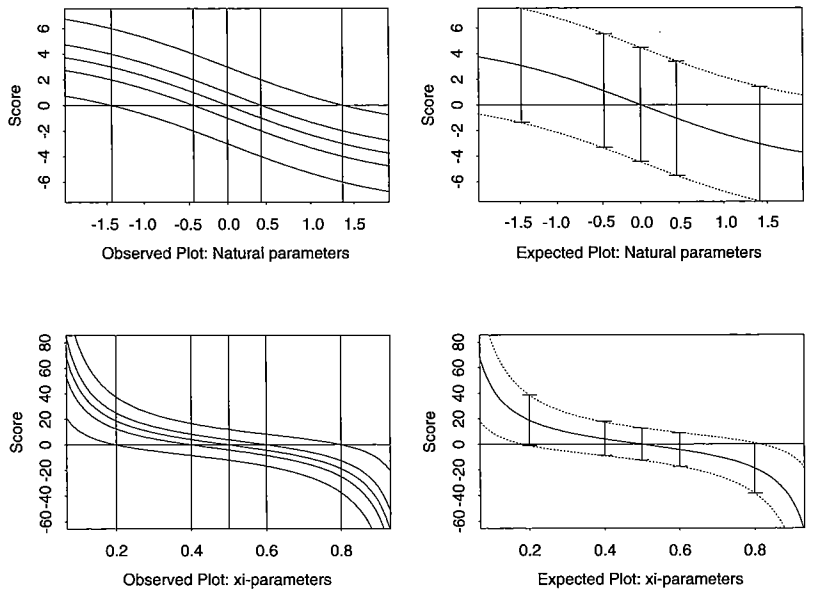

Figure 10.4 One-dimensional Bernoulli example

plots precisely contain the shape of the observed and expected likelihood functions and thus are a direct and visual representation of important statistical information.

The observed score graphs do not cross since, for each fixed parameter value, the observed score function is a non-decreasing affine function of $\bar{t}$. This holds in all parameterisations, using (1). From (1), (2), (4) and (5) it is clear that, in any parameterisation, the graph of the true mean score function coincides with that of the observed score for data where $\bar{t}(\mathbf{x})$ equals its true mean $\eta(\phi)$. In the examples, the true distribution of $n\bar{t}$ is given by Poisson$(\phi + \ln n)$, normal$(n\phi, n)$, gamma(ϕ, n) and binomial(n, ϕ), respectively.

The most striking feature of the plots is the constancy of the variance of the score across the natural parameterisation, and the fact that this property is lost in the alternative parameterisation. Also remarkable is the linearity of the normal plots in the natural parameterisation. A close inspection reveals that for each example, in the natural parameterisation, each of the observed plots differs by only a vertical translation. Again this property will not hold in a general parameterisation. We use these and other features of the plots to better understand Amari's expected geometry.

Certain information is evident from the plots straight away. Under standard regularity conditions, the unique maximum likelihood estimate of a parameter for given data occurs when the graph of the corresponding observed score function crosses the horizontal axis from above. Thus, as $\bar{t} = \hat{\eta}$ in our examples (even in the degenerate Bernoulli case), these five crossing points are the 5%, 25%, 50%, 75% and 95% points of the true distribution of the maximum likelihood estimate. The position of these five crossing points gives visual information about this distribution, in particular about its location, variance and skewness.

Of more direct relevance to our present concern is the fact that, in these one-dimensional cases, there is a straightforward visual representation of the tangent space at each point. TM_θ can be identified with the vertical line through θ, and ρ_θ^ϕ with the distribution of the intersection of this line with the graph of the observed score function. Identical remarks apply in any parameterisation. These tangent spaces are shown in both parameterisations, at the above five percentage points of the maximum likelihood estimator, as lines in the observed plots and as vertical bars in the expected plots.

In the observed plot, the five intersection points with any given tangent space TM_θ are the five corresponding percentage points of ρ_θ^ϕ. The same is true in any increasing reparameterisation ξ. Thus, comparing the position of these five intersection points at corresponding parameter values in

the two observed plots gives direct visual information on the difference between ρ_θ^ϕ and $\rho_{\xi(\theta)}^{\xi(\phi)}$; in particular, on changes in skewness. The observed plots also show very clearly that, as the natural parameter varies, the true distribution of the score changes only in its location, whereas this is not so in a general parameterisation.

This brings to light a certain natural duality between the maximum likelihood estimator and the score function. Consider the observed plots in the natural and mean value parameterisations. For any given point consider its corresponding tangent space TM_θ and $TM_{\eta(\theta)}$ in the two plots. In each plot we have five horizontal and five vertical crossing points, as above, giving information about the distribution of the maximum likelihood estimator and the score function respectively in the same parameterisation. Now, these two plots are far from independent. As $\hat{\eta}(\mathbf{x}) = \eta(\theta) + n^{-1}s(\theta, \mathbf{x})$, the horizontal crossing points in the mean parameter plot are just an affine transformation of the vertical crossing points in the natural parameter plot. The converse is true asymptotically. As we discuss below, this simple and natural duality between the maximum likelihood estimator and the score function corresponds with the duality present in Amari's expected geometry.

3.3 Amari's +1-geometry

The above one-dimensional plots have already indicated two senses in which the natural parameterisation is very special. We note here that this is so generally. Our analysis then provides a simple statistical interpretation of Amari's +1-connection.

From (4) we see that in the natural parameterisation the score function has the form of a stochastic part, independent of θ, plus a deterministic part, independent of the data. Recalling (1) and (4) we see that this property is lost in a non-affine reparameterisation ξ, since $\bar{B}(\theta)$ $(:= \bar{B}_1^1(\theta))$ is independent of θ if and only if ξ is an affine transformation of θ. An equivalent way to describe this property is that the 'error term' $\epsilon^\phi(\theta, \mathbf{x})$ in the mean value decomposition of $s(\theta, \mathbf{x})$ defined at the end of section 1.3 is independent of θ. Or again, as $\mu^\phi(\phi)$ vanishes, that this decomposition has the form

$$s(\theta, \mathbf{x}) = \mu^\phi(\theta) + s(\phi, \mathbf{x}). \tag{7}$$

Note that ρ_θ^ϕ differs from $\rho_{\theta'}^\phi$ only by the translation $\mu^\phi(\theta) - \mu^\phi(\theta')$. In this parameterisation, from one sample to the next, the whole graph of the observed score function just shifts vertically about its ϕ-expectation by the same amount $s(\phi, \mathbf{x})$.

As a consequence of (7), the ϕ-covariance of the score function is independent of θ (and therefore coincides with $g^{\phi}(\phi) = I(\phi)$). But $g^{\phi}(\theta)$ is a metric tensor (section 1.4) and, in this parameterisation, the metric is constant across all tangent spaces. Recalling section 2.2, we note that if a metric is constant in a parameterisation then the parameterisation is affine for the metric connection. All tangent spaces thus have the same geometric structure and differ only by their choice of origin. For more details on this geometric idea of flatness, see Dodson and Poston (1977).

The metric connection is the natural geometric tool for measuring the variation of a metric tensor in any parameterisation. But Critchley, Marriott and Salmon (1994) prove that, in the full exponential family, the metric connection induced by $g^{\phi}(\theta)$ coincides with Amari's $+1$-connection. Thus we have the simple statistical interpretation that ∇^{+1} is the natural geometric measure of the non-constancy of the covariance of the score function in an arbitrary parameterisation. In the one-dimensional case, the $+1$-connection measures the variability of variance of the observed score across different points of M. Looking again at figures 10.1 to 10.4 we see a visual representation of this fact in that the ± 2 standard deviation bars on the expected plot are of a constant length for the θ-parameterisation, and this does not hold in the non-affine ξ-parameterisation.

3.4 Amari's 0-geometry

The fact that in the natural parameterisation all the observed score functions have the same shape invites interpretation. From (7) we see that the common information conveyed in all of them is that conveyed by their ϕ-mean. What is it?

The answer is precisely the Fisher information for the family. This is clear since μ^{ϕ} determines I via

$$I_{ij}(\theta) = -\frac{\partial \mu_j^{\phi}}{\partial \theta^i}(\theta),$$

while the converse is true by integration, noting that $\mu^{\phi}(\phi) = 0$. Thus, in natural parameters, knowing the Fisher information at all points is equivalent to knowing the true mean of the score function (and hence all the observed score functions up to their stochastic shift term). In particular, in the one-dimensional case, the Fisher information is conveyed visually by minus the slope of the graph of $\mu^{\phi}(\theta)$ as, for example, in the natural parameter expected plots of figures 10.1 to 10.4.

Amari uses the Fisher information as his metric tensor. It is important to note that, when endowed with the corresponding metric connection, an exponential family is not in general flat. That is, there does not, in general, exist any parameterisation in which the Fisher information is constant. The multivariate normal distributions with constant covariance matrix and any one-dimensional family are notable exceptions. In the former case, the natural parameters are affine. In the latter case, using (3), the affine parameters are obtained as solutions to the equation

$$\left(\frac{\partial\theta}{\partial\xi}(\theta)\right)^2 \psi''(\theta) = \text{constant}.$$

For example, in the Poisson family where $\psi(\theta) = \exp(\theta)$ one finds $\xi(\theta) = \exp(\theta/2)$, as in Hougaard (1982).

Thus far we have seen that, in the case of the full exponential family, the fundamental components of Amari's geometry (M, I, ∇^{+1}) can be simply and naturally understood in terms of the first two moments of the score function under the distribution assumed to give rise to the data. I is defined by the true mean, and ∇^{+1} by I and the true covariance. Further, they can be understood visually in terms of the expected plots in our one-dimensional examples. We now go on to comment on duality and choice of parameterisation.

3.5 Amari's -1-geometry and duality

The one-dimensional plots above have already indicated a natural duality between the score vector and the maximum likelihood estimator, and that there is a natural statistical curvature, even in the one-dimensional case, unless the manifold is *totally flat*; that is, unless the graph of the true mean score function is linear in the natural parameterisation. We develop these remarks here.

Amari (1990) shows that the mean value parameters

$$\eta(\theta) = \mathbf{E}_{p(x,\theta)}[t(x)] = \psi'(\theta)$$

are -1-affine and therefore, by his general theory, duality related to the natural $+1$-affine parameters θ. We offer the following simple and direct statistical interpretation of this duality. We have,

$$\hat{\eta} = \eta(\theta) + n^{-1}s(\theta, \mathbf{x}).$$

Expanding $\theta(\hat{\eta})$ to first order about η gives an asymptotic converse

$$\hat{\theta} \doteq \theta + n^{-1}\bar{B}(\theta)s(\theta, \mathbf{x}) = \theta + n^{-1}s(\eta, \mathbf{x}),$$

the right-hand equality following from (1) and where we use $\doteq$ to denote first-order asymptotic equivalence. Note that $\bar{B}(\theta) = i^{-1}(\theta)$. Thus the duality between the $+1$ and -1 connections can be seen as the above strong and natural asymptotic correspondence between the maximum likelihood estimator in one parameterisation and the score function in another. In fact this simple statistical interpretation of Amari's duality is not restricted to the full exponential family (see Critchley, Marriott and Salmon (1994)). It is established formally in a more general case than $+1$ duality here in section 3.7.

3.6 Total flatness and choice of parameterisation

The above approximation to $\hat{\theta}$ is exact when θ and η are affinely equivalent. In this case, $\hat{\theta}$ and $\hat{\eta}$ are in the same affine relationship and so their distributions have the same shape. In particular, as normality is preserved under affine transformations, these distributions are as close to normality as each other whatever the definition of closeness that is used. In the case where M is a constant covariance normal family, $\hat{\theta}$ and $\hat{\eta}$ are both exactly normally distributed.

Affine equivalence of θ and η is a very strong property. When it holds, much more is true. It is the equivalent in the full exponential family case of the general geometric notion of total flatness defined and studied in Critchley, Marriott and Salmon (1993). Recall that the natural parameterisation θ has already been characterised by the fact that the true covariance of the score function is constant in it. Total flatness entails this same parameterisation simultaneously has other nice properties. It is easy to show the following equivalences:

$$
\begin{aligned}
& \theta \text{ and } \eta \text{ are affinely equivalent} \\
\iff \quad & \psi \text{ is a quadratic function of } \theta \\
\iff \quad & I(\theta) \text{ is constant in the natural parameters} \\
\iff \quad & \mu^{\phi}(\theta) \text{ is an affine function of } \theta \\
\iff \quad & \exists\, \alpha \neq \beta \text{ with } \nabla^{\alpha} = \nabla^{\beta} \\
\iff \quad & \forall \alpha, \forall \beta, \quad \nabla^{\alpha} = \nabla^{\beta} \\
\iff \quad & \text{the } \theta \text{ parameterisation is } \alpha\text{-affine for all } \alpha
\end{aligned}
$$

(see Critchley, Marriott and Salmon (1993)). In particular, the maximum likelihood estimators of any α-affine parameters are all equally close (in any sense) to normality.

It is exceptional for a family M to be totally flat. Constant covariance multivariate normal families are a rare example. In totally flat manifolds the graph of $\mu^{\phi}(\theta)$ is linear in the natural parameterisation, as remarked upon in the one-dimensional normal example of figure 10.2. More usually, even in the one-dimensional case, a family M of probability (density) functions will exhibit a form of curvature evidenced by the non-linearity of the graph of $\mu^{\phi}(\theta)$.

Recall that the graph of $\mu^{\phi}(\theta)$ enables us to connect the distribution of $\hat{\theta}$ and $\hat{\eta}$. In the natural parameterisation θ, each observed graph is a vertical shift of the expected graph. This shift is an affine function of $\bar{t} = \hat{\eta}$. The intersection of the observed plot with the θ axis determines $\hat{\theta}$. When the expected plot is linear (the totally flat case), then $\hat{\theta}$ and $\hat{\eta}$ are affinely related and so their distributions have the same shape. When it is non-linear they will not be affinely related. This opens up the possibility that, in a particular sense of 'closeness', one of them will be closer to normality.

In all cases, the 0-geometry plays a pivotal role between the ± 1-geometries. That is, the graph of $\mu^{\phi}(\theta)$ determines the relationship between the distributions of the maximum likelihood estimators $\hat{\theta}$ and $\hat{\eta}$ of the ± 1-affine parameters. We illustrate this for our examples in figure 10.5. Both distributions are of course exactly normal when the parent distribution is. In the Poisson case, the concavity of $\mu^{\phi}(\theta)$ means that the positive skewness of $\hat{\eta}$ is reduced. Indeed, $\hat{\theta}$ has negative skew, as figure 10.5a illustrates. The opposite relationship holds in the exponential case, where $\mu^{\phi}(\theta)$ is convex (figure 10.5c). In our Bernoulli example, the form of $\mu^{\phi}(\theta)$ preserves symmetry while increasing kurtosis so that, in this sense, the distribution of $\hat{\theta}$ is closer to normality than that of $\hat{\eta}$ (figure 10.5d).

3.7 Amari's $\pm\frac{1}{3}$-geometry and duality

Amari's $\frac{1}{3}$-connection can be simply interpreted in terms of linearity of the graph of the true mean score function, at least in the one-dimensional situation where the $\frac{1}{3}$-affine parameters are known to exist. If M is totally flat, this graph is linear in the natural parameterisation, as in the normal constant covariance family. It is therefore natural to pose the question: Can a parameterisation be found for a general M in which this graph is linear?

This question can be viewed in two ways. First, for some given $p(x, \phi)$, is such a parameterisation possible? However, in this case, any parameterisation found could be a function of the true distribution. In general, there will not be a single parameterisation that works for all ϕ. The

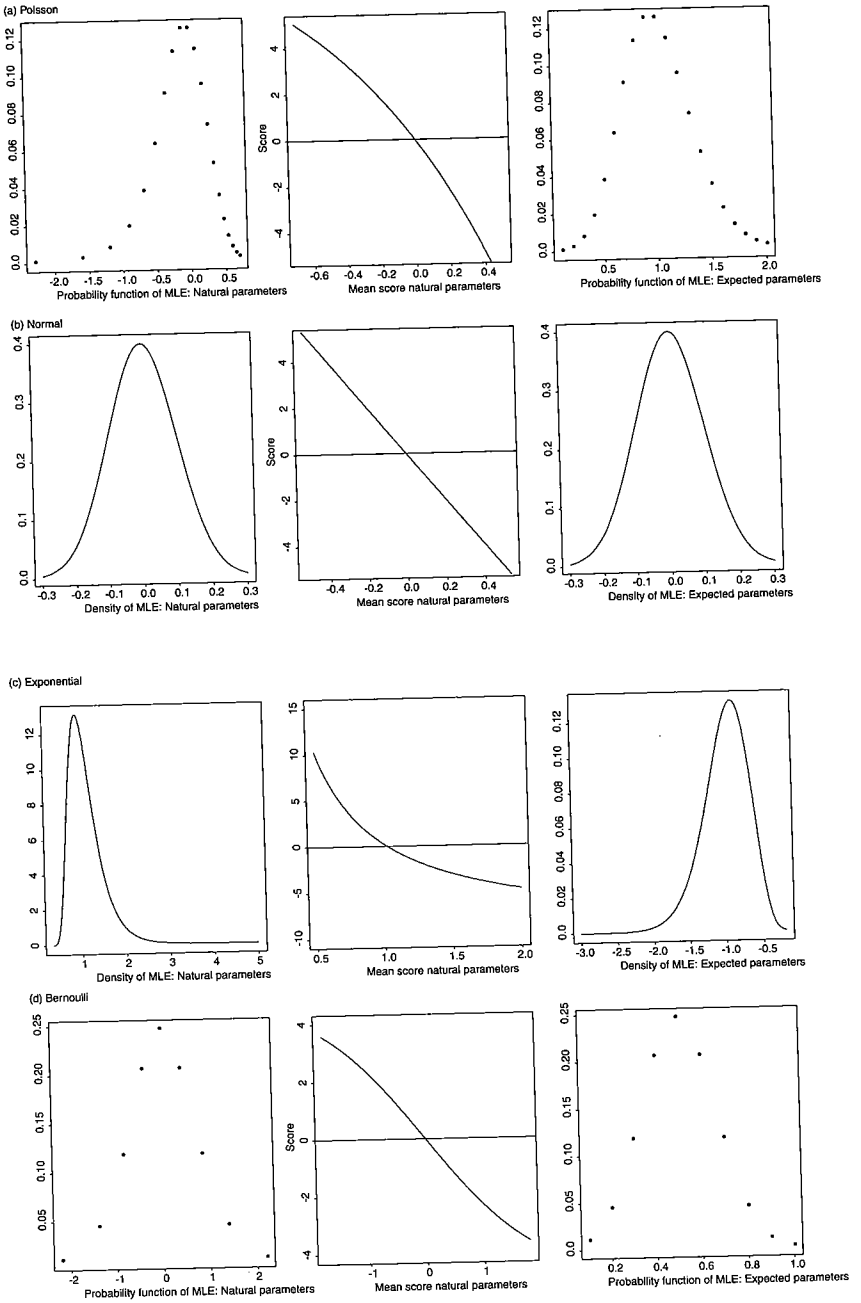

Figure 10.5 The distributions of the natural and expected parameter estimates

second way is to look locally to ϕ. This is the more fruitful approach statistically. The question then becomes: Can a single parameterisation $\theta \to \xi$ be found such that, for all ϕ, the graph of the true mean score is linear locally to $\xi = \xi(\phi)$? In the one-dimensional case, we seek ξ such that

$$\forall \phi, \quad \frac{\partial^2 \mu^{\xi(\phi)}(\xi)}{\partial \xi^2}\Big|_{\xi=\xi(\phi)} = 0.$$

Such a local approach is sufficient asymptotically when the observed score function will be close to its expected value and the maximum likelihood estimate will be close to the true parameter. Thus in such a parameterisation, whatever the true value, the observed log-likelihood will asymptotically be close to quadratic near the MLE. Hence the name, normal likelihood parameter. Amari (1990) shows that such parameters always exist for a one-dimensional full exponential family, and that they are the $\frac{1}{3}$-affine parameters.

The vanishing of the second derivative of the true expected score function in one parameterisation ξ finds a dual echo in the vanishing of the asymptotic skewness of the true distribution of the maximum likelihood estimator in another parameterisation λ. This is called the $-\frac{1}{3}$-affine parameterisation, because it is induced by Amari's $-\frac{1}{3}$-connection. Note again that the duality is between the score function and the maximum likelihood estimator, as in section 3.5. This can be formalised as follows.

Consider any one-dimensional full exponential family,

$$p(x, \theta) = \exp\{t(x)\theta - \psi(\theta)\}.$$

Let ξ and λ be any two reparameterisations. Extending the approach in section 3.5, it is easy to show the following equivalences:

$$\hat{\xi} \doteq \xi + n^{-1} s(\lambda, \mathbf{x}) \iff \hat{\lambda} \doteq \lambda + n^{-1} s(\xi, \mathbf{x}) \iff \frac{\partial \lambda}{\partial \theta} \frac{\partial \xi}{\partial \theta} = \psi''(\theta).$$

In this case, we say that ξ and λ are ψ-dual. Clearly, the natural ($+1$-affine) and mean value (-1-affine) parameters are ψ-dual. A parameter ξ is called self ψ-dual if it is ψ-dual to itself. In this case we find again the differential equation for the 0-affine parameters given in section 3.4. More generally, it can be shown that for any $\alpha \in \mathbf{R}$

$$\xi \text{ and } \lambda \text{ are } \psi\text{-dual} \Rightarrow [\xi \text{ is } \alpha\text{-affine} \iff \lambda \text{ is } -\alpha\text{-affine}].$$

For a proof see the appendix to this chapter. Thus the duality between the score function and the maximum likelihood estimator coincides quite generally with the duality in Amari's expected geometry.

Note that the simple notion of ψ-duality gives an easy way to find $-\alpha$-affine parameters once $+\alpha$-affine parameters are known. For example,

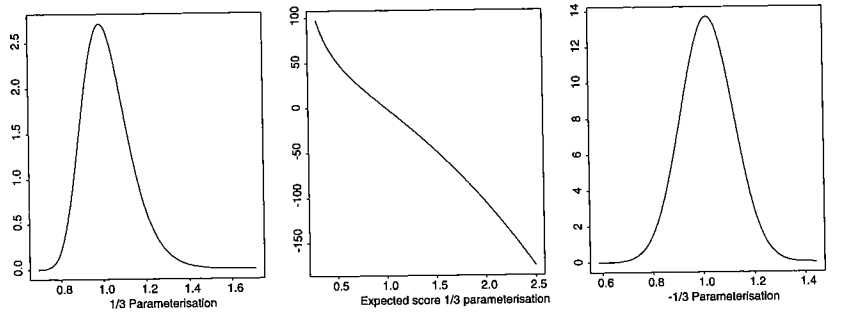

Figure 10.6 The distributions of the 1/3 affine parameter estimates: the exponential case

given that $\xi = \theta^{\frac{1}{3}}$ is $\frac{1}{3}$-affine in the exponential family (Hougaard, 1982) where $\psi(\theta) = -\ln(\theta)$, one immediately has

$$\frac{\partial \lambda}{\partial \theta} = 3\theta^{-4/3},$$

whence $\theta^{-1/3}$ is $-\frac{1}{3}$-affine. Again, in the Poisson family, $\xi = \exp(\theta/3)$ is $\frac{1}{3}$-affine gives at once that $\exp(2\theta/3)$ is $-\frac{1}{3}$-affine.

The local linearity of the true score in $+\frac{1}{3}$-parameters suggests that asymptotically the distributions of the maximum likelihood estimator of the $\pm\frac{1}{3}$-affine parameters will be relatively close compared, for example, with those of the ± 1-affine parameters. In particular, it suggests that both will show little skewness. Figure 10.6, which may be compared with figure 10.5(c), conveys this information for our exponential family example.

4 Sample size effects

In this section we look at the effect of different sample sizes on our plots of the graph of the score vector. For brevity we concentrate on the exponential model. In figure 10.7 we plot the observed scores, taken as before at the 5%, 25%, 50%, 75% and 95% points of the distribution of the score vector. We do this in the natural θ-parameters and the -1-affine mean value η-parameters, for sample sizes 5, 10, 20 and 50.

In the natural parameters we can see that the distribution of $\hat{\theta}$ approaches its asymptotic normal limit. Its positive skewness visibly decreases as the sample size increases. More strikingly, the non-linearity in each of the graphs of the observed scores reduces quickly as n increases. For the sample size 50 case, we see that each graph is, to a close degree of approximation, linear. This implies that at this sample size

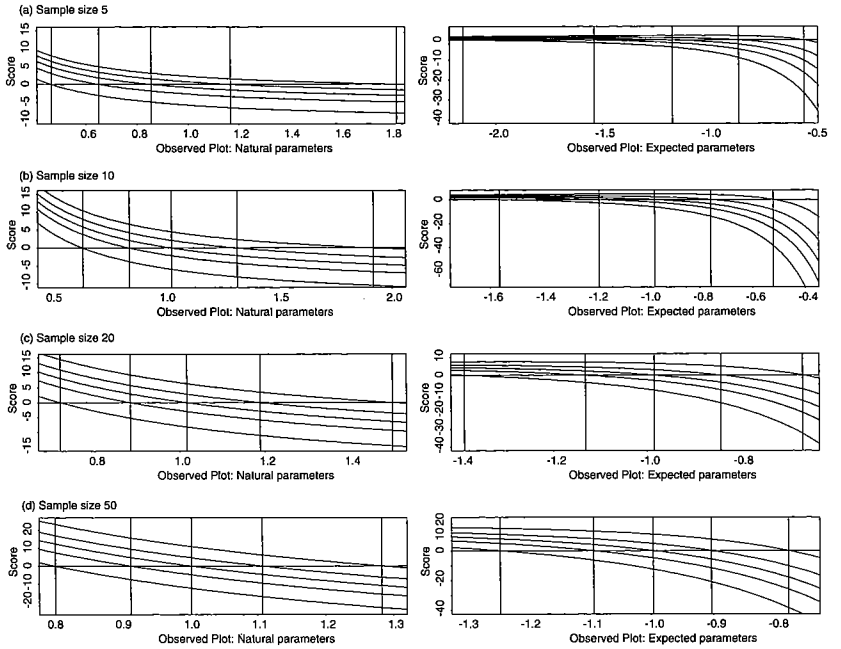

Figure 10.7 The effect of sample size on the relationship between the score vector and the MLE: the exponential case

there will be almost an affine relationship between the score in θ coordinates and the maximum likelihood estimator $\hat{\theta}$, thus demonstrating their well-known asymptotic affine equivalence. It also throws light on the familiar asymptotic equivalence of the score test, the Wald test and (given the asymptotic normality of the maximum likelihood estimate) the likelihood ratio test.

For any model in any smooth invertible reparameterisation of the natural parameters asymptotically the graphs of the observed score will tend to the natural parameterisation plot of the normal distribution shown in figure 10.2. In this limit the graphs become straight and parallel. We can see both these processes in the η-parameterisation of figure 10.7. In this example, a higher sample size than for the natural parameter case is needed to reach the same degree of asymptotic approximation. The highly non-linear and non-parallel graphs of sample size 5 and 10 have been reduced to a much more moderate degree of non-linearity for sample size 50. However, this sample size is not quite sufficient to produce the parallel, linear graphs of the θ-parameterisation, thus there will still not quite be an affine relationship between the score and the maximum likelihood estimator.

Appendix

We give the proof of the equivalence claimed in section 3.7. We assume here familiarity with the use of Christoffel symbols (see Amari (1990), p. 42).

Theorem *Let M be a one-dimensional full exponential family, and assume the parameterisations ξ and λ are ψ-dual. Then ξ is $+\alpha$-affine if and only if λ is $-\alpha$-affine.*

Proof From Amari (1990) we have in the natural θ-parameterisation

$$\Gamma^{\alpha}(\theta) = \left(\frac{1-\alpha}{2}\right)\psi'''(\theta).$$

Thus in ξ-parameters, by the usual transformation rule, the Christoffel symbols are

$$\Gamma^{\alpha}(\xi) = \left(\frac{\partial\theta}{\partial\xi}\right)^{3}\Gamma^{\alpha}(\theta) + i(\theta)\frac{\partial\theta}{\partial\xi}\frac{\partial^{2}\theta}{\partial\xi^{2}}$$

$$= \left(\frac{1-\alpha}{2}\right)\psi'''(\theta)\left(\frac{\partial\theta}{\partial\xi}\right)^{3} + \psi''(\theta)\frac{\partial\theta}{\partial\xi}\frac{\partial^{2}\theta}{\partial\xi^{2}}.$$

Thus ξ is α-flat if and only if

$$\left(\frac{1-\alpha}{2}\right)\psi'''(\theta) + \psi''(\theta)\left(\frac{\partial^{2}\theta}{\partial\xi^{2}}\right)\left(\frac{\partial\xi}{\partial\theta}\right)^{2} = 0. \tag{A.1}$$

Similarly in λ parameters we have λ is $-\alpha$-flat if and only if

$$\left(\frac{1+\alpha}{2}\right)\psi'''(\theta) + \psi''(\theta)\left(\frac{\partial^{2}\theta}{\partial\lambda^{2}}\right)\left(\frac{\partial\lambda}{\partial\theta}\right)^{2} = 0. \tag{A.2}$$

Since ξ and λ are ψ-dual we have

$$\frac{\partial\theta}{\partial\lambda}\frac{\partial\theta}{\partial\xi} = (\psi'')^{-1}(\theta).$$

Differentiating both sides with respect to θ using the chain rule gives

$$\frac{\partial^{2}\theta}{\partial\lambda^{2}}\frac{\partial\lambda}{\partial\theta}\frac{\partial\theta}{\partial\xi} + \frac{\partial^{2}\theta}{\partial\xi^{2}}\frac{\partial\xi}{\partial\theta}\frac{\partial\theta}{\partial\lambda} = -\left(\frac{1}{\psi''}(\theta)\right)^{2}\psi'''(\theta),$$

and multiplying through by $(\psi'')^{2}$ and using the ψ-duality gives

$$\frac{\partial^2\theta}{\partial\lambda^2}\left(\frac{\partial\lambda}{\partial\theta}\right)^2\psi''(\theta)+\frac{\partial^2\theta}{\partial\xi^2}\left(\frac{\partial\xi}{\partial\theta}\right)^2\psi''(\theta)=-\psi'''(\theta). \tag{A.3}$$

Substituting (A.3) into (A.2) gives (A.1), and (A.3) into (A.1) gives (A.2) as required.

References

Amari, S. (1990), *Differential-Geometrical Methods in Statistics*, 2nd edn, Lecture Notes in Statistics No. 28, Berlin: Springer-Verlag.

Barndorff-Nielsen, O.E., D.R. Cox and N. Reid (1986), 'The Role of Differential Geometry in Statistical Theory', *International Statistical Review*, 54: 83–86.

Bates, D.M. and D.G. Watts (1980), 'Relative Curvature Measures of Non-linearity', *Journal of the Royal Statistical Society*, Series B, 40: 1–25.

(1981), 'Parametric Transforms for Improving Approximate Confidence Regions in Non-linear Least Squares', *Annals of Statistics*, 9: 1152–1167.

Cox, D.R. and D.V. Hinkley (1974), *Theoretical Statistics*, London: Chapman & Hall.

Critchley, F., P.K. Marriott and M. Salmon (1993), 'Preferred Point Geometry and Statistical Manifolds', *Annals of Statistics*, 21: 1197–1224.

(1994), 'On the Local Differential Geometry of the Kullback–Leibler Divergence', *Annals of Statistics*, 22: 1587–1602.

Dodson, C.T.J. and T. Poston (1977), *Tensor Geometry*, London: Pitman.

Firth, D. (1993), 'Bias Reduction of Maximum Likelihood Estimates', *Biometrika*, 80: 27–38.

Hougaard, P. (1982), 'Parametrisations of Nonlinear Models', *Journal of the Royal Statistical Society*, Series B, 44: 244–252.

Kass, R.E. (1984), 'Canonical Parametrisation and Zero Parameter Effects Curvature', *Journal of the Royal Statistical Society*, Series B, 46: 86–92.

(1987), 'Introduction', in S.I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen and C.R. Rao (eds.), *Differential Geometry in Statistical Inference*, Hayward, Calif.: Institute of Mathematical Statistics.

(1989), 'The Geometry of Asymptotic Inference', *Statistical Sciences*, 4: 188–234.

McCullagh, P. and J.A. Nelder (1989), *Generalised Linear Models*, 2nd edn, London: Chapman & Hall.

Murray, M.K. and J.W. Rice (1993), *Differential Geometry and Statistics*, London: Chapman & Hall.