An Introduction to Differential Geometry in Econometrics

Paul Marriott
National University of Singapore

Mark Salmon
Department of Banking and Finance
City University Business School

January 22, 2000

1 Introduction

In this introductory chapter we seek to cover sufficient differential geometry in order to understand its application to Econometrics. It is not intended to be a comprehensive review of either differential geometric theory, nor of all the applications which geometry has found in statistics. Rather it is aimed as a rapid tutorial covering the material needed in the rest of this volume and the general literature. The full abstract power of a modern geometric treatment is not always necessary and such a development can often hide in its abstract constructions as much as it illuminates.

In Section 2 we show how econometric models can take the form of geometrical objects known as manifolds, in particular concentrating on classes of models which are full or curved exponential families.

This development of the underlying mathematical structure leads into Section 3 where the tangent space is introduced. It is very helpful, to be able view the tangent space in a number of different, but mathematically equivalent ways and we exploit this throughout the chapter.

Section 4 introduces the idea of a metric and more general tensors illustrated with statistically based examples. Section 5 considers the most important tool that a differential geometric approach offers, the affine connection. We look at applications of this idea to asymptotic analysis, the relationship between geometry and information theory and the problem of the choice of parameterisation. The last two sections look at direct applications of this geometric framework. In particular at the problem of inference in curved families and at the issue of information loss and recovery.

Note that while this chapter aims to give a reasonably precise mathematical development of the required theory an alternative and perhaps more intuitive approach can be found in the chapter by Critchley, Marriott and Salmon later in this volume. For a more exhaustive and detailed review of current geometrical statistical theory see Kass and Vos (1997) or from a more purely mathematical background, see Murray and Rice (1993).

2 Parametric Families and Geometry

In this section we look at the most basic relationship between parametric families of distribution functions and geometry. We begin by first introducing the statistical examples to which the geometric theory most naturally applies; the class of *full* and *curved exponential families*.

Examples are given to show how these families include a broad range of econometric models. Families outside this class are considered in Section 2.3.

Section 2.4 then provides the necessary geometrical theory which defines a *manifold* and shows how one manifold can be defined as a curved sub-family of another. It is shown how this construction gives a very natural framework in which we can describe clearly the geometrical relationship between full and curved exponential families. It further gives the foundations on which a fully geometrical theory of statistical inference can be built.

It is important at the outset to make clear one notational issue in that we shall follow throughout the standard geometric practice of denoting components of a set of parameters by an upper index in contrast to standard econometric notation. In other words if $\theta \in \mathbf{R}^r$ is a r-dimensional parameter vector, then we write it in component terms as

$$\theta = \left(\theta^1, \theta^2, \ldots, \theta^r\right)'.$$

This allows us to use the *Einstein summation convention* where a repeated index in both superscript and subscript is implicitly summed over. For example if $x = (x_1, \ldots, x_r)'$ then the convention states that

$$\theta^i x_i = \sum_{i=1}^{r} \theta^i x_i.$$

2.1 Exponential families

We start with the formal definition. Let $\theta \in \Theta \subseteq \mathbf{R}^r$ be a parameter vector, X a random variable, continuous or discrete, and $s(X) = (s_1(X), \ldots, s_r(X))'$ an r-dimensional statistic. Consider a family of continuous or discrete probability densities, for this random variable, of the form

$$p(x|\theta) = \exp\{\theta^i s_i - \psi(\theta)\} m(x). \tag{1}$$

Remember we are using the Einstein summation convention in this definition. The densities are defined with respect to some fixed dominating measure, ν. The function $m(x)$ is non-negative and independent of the parameter vector θ. We shall further assume that the components of s are not linearly dependent. We call Θ the *natural parameter space* and we shall assume it contains all θ such that

$$\int \exp\{\theta^i s_i\} m(x)\, d\nu < \infty.$$

A parametric set of densities of this form is called a *full exponential family.* If Θ is open in $\mathbf{R}^r$ then the family is said to be *regular*, and the statistics $(s_1, \ldots, s_r)'$ are called the *canonical statistics.*

The function $\psi(\theta)$ will play an important role in the development of the theory below. It is defined by the property that the integral of the density is one, hence

$$\psi(\theta) = \log\left(\int \exp\{\theta^i s_i\} m(x) d\nu\right).$$

It can also be interpreted in terms of the moment generating function of the canonical statistic S. This is given by $M(S; t, \theta)$ where

$$M(S; t, \theta) = \exp\{\psi(\theta + t) - \psi(\theta)\}, \tag{2}$$

see for example Barndorff-Nielsen and Cox (1994, pp4).

The geometric properties of full exponential families will be explored later. However it may be helpful to remark that in Section 5 it is shown that they have a natural geometrical characterisation as the *affine subspaces* in the space of all density functions. They therefore play the role that lines and planes do in three dimensional Euclidean geometry.

2.1.1 Examples

Consider what are perhaps the simplest examples of full exponential families in econometrics; the standard regression model and the linear Simultaneous Equation Model. Most of the standard building blocks of univariate statistical theory are in fact full exponential families including the Poisson, Normal, Exponential, Gamma, Bernoulli, Binomial and Multinomial families. These are studied in more detail in Critchley *et al* later in this volume.

Example 1. The Standard Linear Model. Consider a linear model of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of the single endogenous variable, $\mathbf{X}$ is an $n \times (k+1)$ matrix of the k weakly exogenous variables and the intercept term and $\boldsymbol{\epsilon}$ is the $n \times 1$ matrix of disturbance terms which we assume satisfies the Gauss-Markov conditions. In particular for all i in $1, \ldots, n$

$$\epsilon_i \sim N(0, \sigma^2).$$

The density function of Y *conditionally* on the values of the exogenous variables can then be written as

$$\exp\left\{\left(\frac{\beta}{\sigma^2}\right)'(\mathbf{X}'\mathbf{Y}) + \left(\frac{1}{-2\sigma^2}\right)(\mathbf{Y}'\mathbf{Y}) - \left(\frac{\beta'\mathbf{X}'\mathbf{X}\beta}{2\sigma^2} + (n/2)\log(2\pi\sigma^2)\right)\right\}.$$

This is in precisely the form for a full exponential family with the parameter vector

$$\theta' = \begin{pmatrix} \frac{\beta'}{\sigma^2} & \frac{1}{-2\sigma^2} \end{pmatrix}$$

and canonical statistics

$$(s(\mathbf{Y}))' = \begin{pmatrix} \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}.$$

Example 2. The Simultaneous Equation Model. Consider the set of simultaneous linear equations

$$\mathbf{B}\mathbf{Y_t} + \mathbf{\Gamma}\mathbf{X_t} = \mathbf{U_t},$$

where $\mathbf{Y}$ are endogenous variables, $\mathbf{X}$ weakly exogenous, $\mathbf{U}$ the random component and t indexes the observations. Moving to the Reduced Form we have

$$\mathbf{Y_t} = -\mathbf{B^{-1}}\mathbf{\Gamma}\mathbf{X_t} + \mathbf{B^{-1}}\mathbf{U_t},$$

which gives a full exponential family in a similar way to Example 1. However an important point to notice is that the natural parameters θ in the standard full exponential form are now highly non-linear functions of the parameters in the structural equations. We shall see how the geometric analysis allows us to understand the effect of such non-linear reparametrisations below.

Example 3. Poisson Regression. Moving away from linear models, consider the following Poisson Regression Model. Let μ_i denote the expected value for independent Poisson variables Y_i, $i = 1, \ldots, n$. We shall initially assume that the μ_i parameters are unrestricted. The density for $(y_1, \ldots, y_n)$ can be written as,

$$\exp\{\sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \mu_i\} \prod_{i=1}^{n} \frac{1}{y_i!}.$$

Again this is in full exponential family form, with the natural parameters and canonical statistics being

$$\theta^i = \log(\mu_i), s_i((y_1, \ldots, y_n)) = y_i,$$

respectively. For a true Poisson regression model, the μ_i parameters will be predicted using covariates. This imposes a set of restrictions on the full exponential family which we consider in Section 2.2.

2.1.2 Parametrisations

There is a very strong relationship between geometry and parameterisation. In particular it is important in a geometrically based theory to distinguish between those properties of the model which are dependent on a particular choice of parametrisation and those which are independent of this choice. Indeed one can define the geometry of a space to be those properties which are invariant to changes in parameterisation, see Dodson and Poston (1991).

In Example 2 we noted that the parameters in the structural equations need not be simply related to the natural parameters, θ. Structural parameters will often have a direct econometric interpretation, which will be context dependent. However there are also sets of parameters for full exponential families which always play an important role. The natural parameters, θ are one such set. A second form are the *expected* parameters η. These are defined by

$$\eta^i(\theta) = E_{p(x,\theta)}\left(s_i(x)\right).$$

From Equation 2 it follows that these parameters can be expressed as

$$\eta^i(\theta) = \frac{\partial \psi}{\partial \theta^i}(\theta). \tag{3}$$

In a regular full exponential family the change of parameters from θ to η is a diffeomorphism. This follows since the Jacobian of this transformation is given from Equation 3 as

$$\frac{\partial \eta^i}{\partial \theta^j} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\theta).$$

This will be non-zero since for a regular family ψ is a strictly convex function, see Kass and Vos (1997, page 16, Theorem 2.2.1).

2.1.3 Repeated sampling and sufficient statistics

One important aspect of full exponential families concerns the properties of their sufficient statistics. Let us assume that we have a random sample $(x_1, \ldots, x_n)$ where each observation is drawn from a density

$$p(x, \mid \theta) = \exp\{\theta^i s_i(x) - \psi(\theta)\} m(x).$$

The log-likelihood function for the full sample will be

$$\ell(\theta; (x_1, \ldots, x_n)) = \theta^i \sum_{j=1}^{n} s_i(x_j) - n\psi(\theta).$$

Thus if the parameter space is r-dimensional then there is always an r-dimensional sufficient statistic, namely

$$\left(\sum_{j=1}^{n} s_1(x_j), \ldots, \sum_{j=1}^{n} s_r(x_j) \right).$$

Note that the dimension of this sufficient statistic will be independent of the sample size n. This is an important property which we shall see in Section 2.3 has important implications for the geometric theory.

2.2 Curved exponential families

In the previous section we mentioned that full exponential families will be shown to play the role of affine subspaces in the space of all density functions. Intuitively they can be thought of as lines, planes and higher dimensional Euclidean spaces. We can then ask what would be the properties of *curved* subfamilies of full exponential families?

In general there are two distinct ways in which subfamilies can be defined. Firstly by imposing restrictions on the parameters of the full family and secondly as parametric families in their own right. We use this second approach as the basis of a formal definition.

Let Θ be the r-dimensional natural parameter space for the full exponential family given by

$$p(x|\ \theta) = \exp\{\theta^i s_i - \psi(\theta)\}m(x).$$

Assume that there is a mapping from Ξ, an open subset of $\mathbf{R}^p$ to Θ,

$$\begin{aligned} A : \Xi &\rightarrow \Theta \\ \xi &\mapsto \theta(\xi). \end{aligned}$$

Which obeys the following conditions;

1. the dimension of Ξ is less than that of Θ,
2. the mapping is one-to-one, smooth and its derivative has full rank everywhere,
3. if the sequence of points $\{\theta_i, i = 1, \ldots, r\} \subseteq A(\Xi)$ converges to $\theta_0 \in A(\Xi)$, then $A^{-1}(\theta_i)$ converges to $A^{-1}(\theta_0)$ in Ξ.

Under these conditions the parametric family defined by

$$p(x|\ \xi) = \exp\{\theta^i(\xi)s_i - \psi(\theta(\xi))\}m(x)$$

is called a *curved exponential* family. In particular noting the dimensions of the relevant spaces, it is an (r, p)-curved exponential family.

2.2.1 Examples

We now look at a set of examples to see how this class of curved exponential families is relevant to econometrics. For further examples see Kass and Vos (1997) or Barndorff-Nielsen and Cox (1994), where many forms of Generalised Linear Models, including logistic, binomial and exponential regressions, non-linear regression models, time series models and stochastic processes are treated. Another important source of curved exponential families is the imposition and testing of parametric restrictions, see Example 5 . Finally we mention some general approximation results which state that *any* parametric family can be approximated using a curved exponential family, see for example Barndorff-Nielsen and Jupp (1989).

Example 3. Poisson regression (continued). Let us now assume that the parameters in the Poisson regression model treated above are assumed to be determined by a set of covariates. As a simple example we could assume the means follow the equation

$$\log(\mu_i) = \alpha + \beta X_i,$$

where X is an exogenous variable. Hence in terms of the natural parameters we have

$$\theta^i = \alpha + \beta X_i.$$

Thus the map defining the curved exponential family is

$$(\alpha, \beta) \rightarrow \left(\theta^1(\alpha, \beta), \ldots, \theta^n(\alpha, \beta)\right),$$

and we have a $(n, 2)$-curved exponential family.

Example 4. $AR(1)$**-model.** Consider the simple $AR(1)$ model

$$x_t = \alpha x_{t-1} + \epsilon_t$$

where the disturbance terms are independent $N(0, \sigma^2)$ variables, and we assume $x_0 = 0$. The density function will then be of the form

$$\exp\left\{\left(\frac{-1}{2\sigma^2}\right)\sum_{i=1}^{n} x_i^2 + \left(\frac{\alpha}{\sigma^2}\right)\sum_{i=1}^{n} x_t x_{t-1} + \left(\frac{-\alpha^2}{2\sigma^2}\right)\sum_{i=1}^{n} x_{t-1}^2 - \frac{n}{2}\log(2\pi\sigma^2)\right\}.$$

This is a curved exponential family since the parameters can be written in the form

$$\theta^1(\alpha, \sigma) = \frac{-1}{2\sigma^2}, \theta^2(\alpha, \sigma) = \frac{\alpha}{\sigma^2}, \theta^3(\alpha, \sigma) = \frac{-\alpha^2}{2\sigma^2}.$$

The geometry of this and more general $ARMA$-families has been studied in Ravishanker (1994).

Example 5. COMFAC model. Curved exponential families can also be defined by imposing restrictions on the parameters of a larger, full or curved, exponential family. As we will see, if these restrictions are non-linear in the natural parameters the restricted model will, in general, be a curved exponential family. As an example consider the COMFAC model,

$$y_t = \gamma x_t + u_t,$$

where x is weakly exogenous and the disturbance terms follow a normal $AR(1)$ process

$$u_t = \rho u_{t-1} + \epsilon_t.$$

Combining these gives a model

$$y_t = \rho y_{t-1} + \gamma x_t - \rho\gamma x_{t-1} + \epsilon_t$$

which we can think of as a restricted model in an unrestricted autoregressive model.

$$y_t = \alpha_0 y_{t-1} + \alpha_1 x_t + \alpha_2 x_{t-1} + \omega_t.$$

We have already seen that the autoregressive model gives rise to a curved exponential structure. The COMFAC restriction in this simple case is given by a polynomial in the parameters

$$\alpha_2 + \alpha_0\alpha_1 = 0.$$

The family defined by this nonlinear restriction will also be a curved exponential family. Its curvature is defined by a nonlinear restriction in a family which is itself curved. Thus the COMFAC model is curved exponential and testing the validity of the model is equivalent to testing the validity of one curved exponential family in another. We shall see later how the geometry of the embedding of a curved exponential family affects the properties of such tests as discussed by Van Garderen in this volume and Critchley, Marriott and Salmon(1996) among many others.

2.3 Non exponential families

Of course not all parametric families are full or curved exponential and we therefore need to consider families which lie outside this class and how this affects the geometric theory. We only have space to highlight the issues here but it is clear that families which have been excluded include the Weibull, Generalised Extreme Value and Pareto distributions and these are of practical relevance in a number of areas of econometric application. An important feature which these families have is that the dimension of their sufficient statistics grows with the sample size. While this does not make an exact geometrical theory impossible it does considerably complicate matters.

Another property which the non-exponential families can exhibit is that the support of the densities can be parameter dependent. Thus members of the same family need not be mutually absolutely continuous. Again while this need not exclude a geometrical theory it does make the development more detailed and we will not consider this case.

In general the development below covers families that satisfy standard regularity conditions found, for instance, in Amari (1990, page 16). In detail these conditions for a parametric family $p(x|\theta)$ are:

1. all members of the family have common support,
2. let $\ell(\theta\ ;\ x) = \log \mathrm{Lik}(\theta\ ;\ x)$, then the set of functions

$$\{\frac{\partial \ell}{\partial \theta^i}(\theta\ ;\ x)|i = 1, \ldots, n\}$$

   are linearly independent,
3. moments of $\frac{\partial \ell}{\partial \theta^i}(\theta\ ;\ x)$ exist up to sufficiently high order,
4. for all relevant functions integration and taking partial derivatives with respect to θ are commutative.

These conditions exclude a number of interesting models but will not, in general, be relevant for many standard econometric applications. All full exponential families satisfy these conditions, as do a large number of other classes of families.

2.4 Geometry

We now look at the general relationship between parametric statistical families and geometric objects known as *manifolds*. These can be thought of intuitively as multidimensional generalisations of surfaces. The theory of manifolds is fundamental to the development of differential geometry although we do not need the full abstract theory which would be found in any modern treatment such as Spivak (1979) or Dodson and Poston (1991). We develop a simplified theory suitable to explain the geometry of standard econometric models. Fundamental to this approach is the idea of an *embedded manifold*. Here the manifold is defined as a subset of a much simpler geometrical object called an *affine space*. This affine space construction avoids complications created by the need to fully specify and manipulate the infinite dimensional space of all proper density functions. Nothing is lost by just considering this affine space approach when the affine space we consider is essentially defined as the space of all log likelihood functions. An advantage is that with this construction we can trust our standard Euclidean intuition based on surfaces inside 3-dimensional spaces regarding the geometry of the econometric models we want to consider.

The most familiar geometry is 3-dimensional Euclidean space, consisting of points, lines and planes. This is closely related to the geometry of a real vector space except for the issue of the choice of origin. In Euclidean Space, unlike a vector space, there is no natural choice of origin. It is an example of an affine geometry, for which we have the following abstract definition.

An affine space (X, V) consists of a set X, a vector space V, together with a translation operation $+$. This is defined for each $v \in V$, as a function

$$\begin{aligned} X &\to X \\ x &\mapsto x + v \end{aligned}$$

which satisfies

$$(x + v_1) + v_2 = x + (v_1 + v_2)$$

and is such that, for any pair of points in X there is a *unique* translation between them.

Most intuitive notions about Euclidean space will carry over to general affine spaces, although care has to be taken in infinite dimensional examples. We shall therefore begin our definition of a manifold by first considering curved subspaces of Euclidean space.

2.4.1 Embedded manifolds

As with our curved exponential family examples, curved subspaces can be defined either using parametric functions, or as solutions to a set of restrictions. The following simple but abstract example can most easily get the ideas across.

Example 6. The Sphere model. Consider in $\mathbf{R}^3$, with some fixed origin and axes, the set of points which are the solutions of

$$x^2 + y^2 + z^2 = 1.$$

This is of course the unit sphere, centred at the origin. It is an example of an embedded manifold in $\mathbf{R}^3$ and is a curved 2-dimensional surface. At least part of the sphere can also be defined more directly, using parameters, as the set of points

$$\left\{ \left(\cos(\theta^1) \sin(\theta^2), \sin(\theta^1) \sin(\theta^2), \cos(\theta^2) \right) \mid \theta^1 \in (-\pi, \pi), \theta^2 \in (0, \pi) \right\}.$$

Note that both the north and south poles have been excluded in this definition, as well as the curve

$$\left(-\sin(\theta^2), 0, \cos(\theta^2) \right).$$

The poles are omitted in order to ensure that the map from the parameter space to $\mathbf{R}^3$ is invertible. The line is omitted to give a geometric regularity known as an *immersion*. Essentially we want to keep the *topology* of the parameter space consistent with that of its image in $\mathbf{R}^3$.

The key idea here is we want the parameter space, which is an open set in Euclidean space, to represent the model as faithfully as possible. Thus it should have the same topology and the same smoothness structure.

We shall now give a formal definition of a manifold which will be sufficient for our purposes. Our manifolds will always be subsets of some fixed affine space, so more properly we are defining a submanifold.

Consider a smooth map from Φ, an open subset of $\mathbf{R}^r$, to the affine space (X, V) defined by

$$i : \Phi \to X.$$

The set $i(\Phi)$ will be an embedded manifold if the following conditions apply:

(A) the derivative of i has full rank r for all points in Φ,

(B) i is a *proper* map, that is the inverse image of any compact set is itself compact, see Bröcker and Jänich (1982, page 71).

In the sphere example it is Condition (A) which makes us exclude the poles and Condition (B) which excludes the line. This is necessary for the map to be a diffeomorphism and this in turn is required to ensure the parameters represent unique points in the manifold and hence the econometric model is well defined and identified.

Another way of defining a (sub-)manifold of an affine space it to use a set of restriction functions. Here the formal definition is; Consider a smooth map ρ from an n-dimensional affine space (X, V) to $\mathbf{R}^r$. Consider the set of solutions of the restriction

$$\{x \mid \rho(x) = 0\},$$

and suppose that for all points in this set the Jacobian of ρ has rank r, then the set will be an $(n-r)$-dimensional manifold.

There are two points to notice in this alternative definition. Firstly we have only applied it to restrictions of finite dimensional affine spaces. The generalisation to the infinite dimensional case is somewhat more technical. Secondly the two alternatives will be locally equivalent due to the Inverse Function Theorem, see Rudin (1976).

We note again that many standard differential geometric textbooks do not assume that a manifold need be a subset of an affine space, and therefore they require a good deal more machinery in their definitions. Loosely the general definition states that a manifold is *locally* diffeomorphic to an open subset of Euclidean space. At each point of the manifold there will be a small local region in which the manifold looks like a curved piece of Euclidean space. The structure is arranged in such a way that these local subsets can be combined in a smooth way. There are a number of technical issues which are required to make such an approach rigorous in the current setting. Again we emphasise that we will always have an embedded structure for econometric applications, thus we can sidestep a lengthy theoretical development.

Also it is common, in standard geometric works, to regard parameters, not as labels to distinguish points, but rather functions of these points. Thus if M is an r-dimensional manifold then a set of parameters $(\theta^1, \ldots, \theta^r)$ is a set of smooth functions

$$\theta^i : M \to \mathbf{R}.$$

In fact this is very much in line with an econometric view of parameters in which the structural parameters of the model are functions of the probability structure of the model. For example we could parameterise a family of distributions using a finite set of moments. Moments are clearly most naturally thought of as functions of the points, when points of the manifolds are actually distribution functions.

2.4.2 Statistical manifolds

In this section we show how parametric families of densities can be seen as manifolds. First we need to define the affine space that embeds all our families and we follow the approach of Murray and Rice (1993) in this development. Rather than working with densities directly we work with log-likelihoods since this enables the natural affine structure to become more apparent. However because of the nature of the likelihood function some care is needed with this definition.

Consider the set of all (smooth) positive densities with a fixed common support S, each of which is defined relative to some fixed measure ν. Let this family be denoted by $\mathcal{P}$. Further let

us denote by $\mathcal{M}$ the set of all positive measures which are absolutely continuous with respect to ν. It will be convenient to consider this set up to scaling by a positive constant. That is we will consider two such measures equivalent if and only if they differ by multiplication of a constant. We denote this space by $\mathcal{M}^*$. Define X by

$$X = \{\log(m) \mid m \in \mathcal{M}^*\}.$$

Because $m \in \mathcal{M}^*$ is defined only up to a scaling constant we must have the identification in X that

$$\log(m) = \log(Cm) = \log(m) + \log(C), \quad \forall\, C \in \mathbf{R}^+.$$

Note that the space of log-likelihood functions is a natural subset of X. A log-likelihood is defined *only* up to the addition of a constant, Cox and Hinkley (1974). Thus any log-likelihood $\log(p(x))$ will be equivalent to $\log(p(x)) + \log(C)$ for all $C \in \mathbf{R}^+$. Finally define the vector space V by $V = \{f(x) | f \in C^\infty(S, \mathbf{R})\}$.

The pair (X, V) is given an affine space structure by defining translations as,

$$\log(m) \mapsto \log(m) + f(x) = \log(\exp(f(x))m)$$

Since $\exp(f(x))m$ is a positive measure, the image of this map does lie in X. It is then immediate that $(\log(m) + f_1) + f_2 = \log(m) + (f_1 + f_2)$ and the translation from $\log(m_1)$ to $\log(m_2)$ is uniquely defined by $\log(m_2) - \log(m_1) \in C^\infty(S, \mathbf{R})$, hence the conditions for an affine space apply.

Using this natural affine structure, consider a parametric family of densities which satisfies the regularity conditions from Section 2.3. Condition 1 implies that the set of log-likelihoods defined by this family will lie in X. From Condition 2 it follows that Condition (A) holds immediately. Condition (B) will hold for almost all Econometric models, in particular it will always hold if the parameter space is compact and in practice this will not be serious restriction. Hence the family will be an (embedded) manifold.

We note further that the set $\mathcal{P}$ is defined by a simple restriction function as a subset of $\mathcal{M}$. This is because all elements of $\mathcal{P}$ must integrate to one. There is some intuitive value in therefore thinking of $\mathcal{P}$ as a submanifold of $\mathcal{M}$. However as pointed out in Section 2.4.1 the definition of a manifold by a restriction function works most simply when the embedding space is finite dimensional. There are technical issues involved in formalising the above intuition, which we do not discuss here. However, this intuitive idea is useful for understanding the geometric nature of full exponential families. Their log-likelihood representation will be

$$\theta^i s_i(x) - \psi(\theta).$$

This can be viewed in two parts. Firstly, an affine function of the parameters θ, which fits naturally into the affine structure of X. Secondly there is a normalising term $\psi(\theta)$ which ensures that the integral of the density is constrained to be one. Very loosely think of $\mathcal{M}$ as an affine space in which $\mathcal{P}$ is a curved subspace, the role of the function ψ is to project an affine function of the natural parameters back into $\mathcal{P}$.

Example 4. $AR(1)$**-model (continued).** We illustrate the previous theory with an explicit calculation for the $AR(1)$ model. We can consider this family as a sub-family of the n-dimensional multivariate Normal model, where n is the sample size. This is the model which determines the innovation process. Thus it is a sub-model of a n-dimensional full exponential family. In fact it lies in a 3-dimensional sub-family of this full exponential family. This is the smallest full family which contains the $AR(1)$ family and its dimension is determined by the

dimension of the minimum sufficient statistic. The dimension of the family itself is determined by its parameter space, given in our case by α and σ. It is a $(3, 2)$ curved exponential family.

Its log likelihood representation is

$$\ell(\alpha, \sigma : x) = \left(\frac{-1}{2\sigma^2}\right) \sum_{i=1}^{n} x_i^2 + \left(\frac{\alpha}{\sigma^2}\right) \sum_{i=1}^{n} x_t x_{t-1} + \left(\frac{-\alpha^2}{2\sigma^2}\right) \sum_{i=1}^{n} x_{t-1}^2 - \frac{n}{2} \log(2\pi\sigma^2)$$

2.4.3 Repeated samples

The previous section has demonstrated that a parametric family $p(x \mid \theta)$ has the structure of a geometric manifold. However in statistical application we need to deal with repeated samples; independent or dependent. So we need to consider a set of related manifolds which are indexed by the sample size n. The exponential family has a particularly simple structure to this sequence of manifolds.

One reason for the simple structure is the fact that the dimension of the sufficient statistic does not depend on the sample size. If X has density function given by (1), then an i.i.d. sample $(x_1, \ldots, x_n)$ has density

$$p((x_1, \ldots, x_n) | \ \theta) = \exp\left\{\theta^i \sum_{j=1}^{n} s_i(x_j) - n\psi(\theta)\right\} \prod_{j=1}^{n} m(x_j)$$

This is also therefore a full exponential family, hence an embedded manifold. Much of the application of geometric theory is concerned with asymptotic results. Hence we would be interested in the limiting form of this sequence of manifolds as $n \to \infty$. The simple relationship between the geometry and the sample size in full exponential families is then used to our advantage.

In the case of linear models, or dependent data the story will of course be more complex. There will still be a sequence of embedded manifolds but care needs to be taken with, for example, the limit distribution of exogenous variables. As long as the likelihood function for a given model can be defined the geometric construction we have set up will apply. In general econometric models with dependent data and issues of exogeneity the correct conditional distributions have to be used to define the appropriate likelihood for our geometric analysis as was implicitly done in the $AR(1)$ example above with the prediction error decomposition.

2.4.4 Bibliographical remarks

The term *curved exponential family* is due to Efron (1975, 1978 & 1982) in a series of seminal papers which revived interest in the geometric aspects of statistical theory. This work was followed by a series of paper by Amari *et al*, most of the material from which can be found in Amari (1990). The particular geometry treatment in this section owes a lot to Murray and Rice's (1993) more mathematical based approach, also to the excellent reference work by Kass and Vos (1997).Since the exponential family class includes all the standard building blocks of statistical theory the relevant references go back to the beginnings of probability and statistics. Good general references are however Brown (1986), Barndorff-Nielsen (1978, 1988).

3 The Tangent Space

We have seen that parametric families of density functions can take the mathematical form of manifolds. However this in itself has not defined the geometric structure of the family.

It is only the foundation stone on which a geometric development stands. In this section we concentrate on the key idea of a differential geometric approach. This is the notion of a *tangent space*. We first look at this idea from a statistical point of view, defining familiar statistical objects such as the score vector. We then show that these are precisely what the differential geometric development requires. Again we shall depart from the form of the development that a standard abstract geometric text might follow as we can exploit the embedding structure which was carefully set up in Section 2.4.2. This structure provides the simplest accurate description of the geometric relationship between the score vectors, the maximum likelihood estimator and likelihood based inference more generally.

3.1 Statistical representations

We have used the log-likelihood representation above as an important geometric tool and closely related is the score vector defined as

$$\left(\frac{\partial \ell}{\partial \theta^1}, \ldots, \frac{\partial \ell}{\partial \theta^r}\right)'.$$

One of the fundamental properties of the score comes from the following familiar argument. Since

$$\int p(x \mid \theta) d\nu = 1$$

it follows that

$$\frac{\partial}{\partial \theta^i} \int p(x \mid \theta) d\nu = \int \frac{\partial}{\partial \theta^i} p(x \mid \theta) d\nu = 0$$

using Property 4 in Section 2.3, then

$$E_{p(x,\theta)}\left(\frac{\partial \ell}{\partial \theta^i}\right) = \int \frac{1}{p(x \mid \theta)} \frac{\partial}{\partial \theta^i} p(x \mid \theta) p(x \mid \theta) d\nu = 0. \tag{4}$$

We present this argument in detail as it has important implications for the development of the geometric theory.

Equation 4 is the basis of many standard asymptotic results when combined with a Taylor expansion around the maximum likelihood estimate (MLE), $\hat{\theta}$,

$$\hat{\theta}^i - \theta^i = \mathcal{I}^{ij} \frac{\partial \ell}{\partial \theta^j} + O(\frac{1}{n}) \tag{5}$$

where $\left(-\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\hat{\theta})\right)^{-1} = \mathcal{I}^{ij}$, see Cox and Hinkley (1974). This shows that in an asymptotically shrinking neighbourhood of the data generating process the score statistic will be directly related to the MLE. The geometric significance of this local approximation will be shown in Section 3.2.

The efficiency of the maximum likelihood estimates is usually measured by the covariance of the score vector or the expected Fisher information

$$I_{ij} = E_{p(x,\theta)}\left(-\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}\right) = \mathrm{Cov}_{p(x,\theta)}\left(\frac{\partial \ell}{\partial \theta^i}, \frac{\partial \ell}{\partial \theta^j}\right).$$

Efron and Hinkley (1978) however argue that a more relevant and hence accurate measure of this precision is given by the observed Fisher information

$$\mathcal{I}_{ij} = -\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\hat{\theta}),$$

since this is the appropriate measure to be used after conditioning on suitable ancillary statistics.

The final property of the score vector we need is its behaviour under conditioning by an ancillary statistic. Suppose that the statistic a is exactly ancillary for θ, and we wish to undertake inference conditionally on a. We then should look at the conditional log-likelihood function

$$\ell(\theta \mid a) = \log(p(x \mid \theta, a)).$$

However when a is exactly ancillary

$$\frac{\partial \ell}{\partial \theta^i}(\theta : |a) = \frac{\partial \ell}{\partial \theta^i}(\theta),$$

in other words, the conditional score will equal the unconditional. Thus the score is unaffected by conditioning on *any* exact ancillary. Because of this the statistical manifold we need to work with is the same whether we work conditionally or unconditionally because the affine space differs only by an translation which is invariant.

3.2 Geometrical theory

Having reviewed the properties of the score vector we now look at the abstract notion of a *tangent space* to a manifold. It turns out that the space of score vectors defined above will be a statistical representation of this general and important geometrical construction. We shall look at two different, but mathematically equivalent, characterisations of a tangent space.

Firstly we note again that we only study manifolds which are embedded in affine spaces. These manifolds will in general be non-linear, or curved objects. It is natural to try and understand a non-linear object by linearising. Therefore we could study a curved manifold by finding the best affine approximation at a point. The properties of the curved manifold, in a small neighbourhood of this point, will be approximated by those in the vector space.

The second approach to the tangent space at a point is to view it as the set of all *directions* in the manifold at that point. If the manifold were r-dimensional then we would expect that this space will have the same dimension, in fact to be an r-dimensional affine space.

3.2.1 Local affine approximation

We first consider the local approximation approach. Let M be an r-dimensional manifold, embedded in an affine space N, and let p be a point in M. We first define a tangent vector to a curve in a manifold M. A curve is defined to be a smooth map

$$\begin{aligned} \gamma : (-\epsilon, \epsilon) \subset \mathbf{R} &\rightarrow M \\ t &\mapsto \gamma(t), \end{aligned}$$

such that $\gamma(0) = p$. The tangent vector at p will be defined by

$$\gamma'(0) = \lim_{h \to 0} \frac{\gamma(h) - \gamma(0)}{h}.$$

We note that since we are embedded in N, $\gamma'(0)$ will be an element of this affine space, see Dodson and Poston (1991). It will be a vector whose origin is p. The tangent vector will be the best linear approximation to the curve γ, at p. It is the unique line which approximates the curve to first order, see Willmore (1959, page 8).

We can then define TM_p, the tangent space at p to be the set of all tangent vectors to curves through p. Let us put a parameterisation θ of an open neighbourhood which includes $p$on M. We define this as a map ρ

$$\rho : \Phi(\subseteq \mathbf{R}^r) \to N$$

where $\rho(\Phi)$ is an open subset of M which contains p, and the derivative is assumed to have full rank for all points in Φ. Any curve can then be written as a composite function in terms of the parameterisation,

$$\rho \circ \gamma : (-\epsilon, \epsilon) \to \Phi \to N$$

Thus any tangent vector will be of the form

$$\frac{\partial \rho}{\partial \theta^i} \frac{d\theta^i}{dt}.$$

Hence TM_p will be spanned by the vectors of N given by

$$\left\{ \frac{\partial \rho}{\partial \theta^i} \ , \ i = 1, \ldots, r \right\}.$$

Thus TM_p will be a p-dimensional affine subspace of N. For completeness we need to show that the construction of TM_p is in fact independent of the choice of the parameterisation. We shall see this later but for details see Wilmore (1959).

3.2.2 Space of directions

The second approach to defining the tangent space is to think of a tangent vector as defining a direction in the manifold M. We define a direction in terms of a directional derivative. Thus a tangent vector will be viewed as a differential operator which corresponds to the directional derivative in a given direction.

The following notation is used for a tangent vector which makes clear its role as a directional derivative

$$\frac{\partial}{\partial \theta^i} = \partial_i.$$

It is convenient in this viewpoint to use a axiomatic approach. Suppose M is a smooth manifold. A tangent vector at $p \in M$ is a mapping

$$X_p : C^\infty(M) \to \mathbf{R}$$

such that for all $f, g \in C^\infty(M)$, and $a \in \mathbf{R}$:

1. $X_p(a.f + g) = aX_p(f) + X_p(g)$,
2. $X_p(f.g) = g.X_p(f) + f.X_p(g)$.

It can be shown that the set of such tangent vectors will form an r-dimensional vector space, spanned by the set

$$\{\partial_i \ , \ i = 1, \ldots, r\}.$$

Further that this vector space will be isomorphic to that defined in Section 3.2.1. For details see Dodson and Poston (1991).

It is useful to have both viewpoints of the nature of a tangent vector. The clearest intuition follows from the development in Section 3.2.1, whereas for mathematical tractability the axiomatic view, in this section is superior.

3.2.3 The dual space

We have seen that the tangent space TM_p is a vector space whose origin is at p. We can think of it as a subspace of the affine embedding space. Since it is a vector space it is natural to consider its dual space TM_p^*. This is defined as the space of all linear maps

$$TM_p \to \mathbf{R}.$$

Given a parameterisation, we have seen we have a basis for TM_p given by

$$\{\partial_1, \ldots, \partial_r\}.$$

The standard theory for dual spaces, see Dodson and Poston (1991), shows that we can define a basis for TM_p^* to be

$$\left\{d\theta^1, \ldots, d\theta^r\right\}$$

where each $d\theta^i$ is defined by the relationship

$$\partial_i(d\theta^j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

We can interpret $d\theta^i$, called a 1-*form* or *differential*, as a real valued function defined on the manifold M which is constant in all tangent directions apart from the ∂_i direction. The level sets of the 1-forms defines the coordinate grid on the manifold given by the parameterisation θ.

3.2.4 The change of parametrisation formulae

So far we have defined the tangent space explicitly by using a set of basis vectors in the embedding space. This set was chosen through the choice of parameterisation. Since this parameterisation is not intrinsic to the geometry of the manifold we must be able to convert from the basis defined in terms of one parameterisation to that defined in terms of another. Indeed we must show that the tangent space and its dual exist independently of the choice of parameters used.

The change of parameterisation formulae in fact turn out to be repeated applications of the standard chain rule from multivariate calculus. Suppose we have two parameterisations θ and η, for an open subset of the manifold M. Due to the full rank of the derivative of parameterisations there is a diffeomorphism connecting the parameters, i.e., $\eta = \eta(\theta)$ is a smooth invertible map, whose inverse is smooth. This follows from the inverse function theorem and the fact that we are only working locally in an open neighbourhood of some point in the manifold, see for example Kass and Vos (1997, page 300). We write $\partial_a \eta^i$ to denote the ith partial derivative of η with respect to the ath component of θ. So we have

$$\partial_a \eta^i = \frac{\partial \eta^i}{\partial \theta^a}(\theta).$$

Application of the chain rule then allows us to connect the basis in the θ-parameters, which we denote by $\{\partial_a \mid a = 1, \ldots, r\}$, indexed by a, with the basis relative to the η-parameterisation, which is denoted by $\{\partial_i \mid i = 1, \ldots, r\}$. Thus, recalling from Section 3.2.2, that the basis elements are defined as differential operators, we can apply the chain rule. This gives the formula connecting the two basis as being

$$\partial_a = \left(\partial_a \eta^i\right) \partial_i.$$

Note here, as throughout, the use of the Einstein summation notation. Since this is just an invertible linear transformation this result immediately implies that the tangent space spanned by the two basis sets will be the same. Hence we have shown that the previous definitions are indeed well defined.

Further, to complete the change of basis formulae suppose we have a tangent vector X. We could write it in component terms relative to the θ-parameterisation as

$$X = X^a \partial_a.$$

Thus changing basis gives

$$\begin{aligned} X &= X^a \left(\partial_a \eta^i\right) \partial_i \\ &= \left(X^a \partial_a \eta^i\right) \partial_i. \end{aligned}$$

Hence the *components* will have the following change of basis formula

$$(X^1, \ldots, X^r) \mapsto \left(X^a \partial_a \eta^1, \ldots, X^a \partial_a \eta^1\right). \tag{6}$$

The change of basis formulae for the dual space can be derived in a similar way. The relationship connecting the basis relative to the θ-parameterisation, $\{d\theta^a \mid : a = 1, \ldots, r\}$, with that of the η-parameterisation, $\{d\eta^i \mid : i = 1, \ldots, r\}$is given by

$$d\theta^a = \partial_i \theta^a d\eta^i,$$

where $\partial_i \theta^a = \frac{\partial \eta^i}{\partial \theta^a}$. Note that there is also the important relationship between the two changes of basis matrices, which states

$$\partial_i \theta^a \partial_a \eta^j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{7}$$

That is, viewed as matrices, $\partial_i \theta^a$ and $\partial_a \eta^j$ are mutual inverses.

3.2.5 Statistical interpretation

We can now relate this previous theory to our statistical manifolds. Recall that we are representing the manifolds by their log-likelihood representations. Let us now consider the representation of the tangent space in terms of the local affine approximation to the manifold.

Let the parameteric family be given by $p(x \mid \theta)$, then its log-likelihood representation is given by $\log p(\theta \mid x) = \ell(\theta)$. A path in this manifold, through the point represented by θ_0, is given by

$$\begin{aligned} \gamma : (-\epsilon, \epsilon) &\rightarrow \Phi \rightarrow X \\ t &\mapsto \ell(\theta(t)). \end{aligned}$$

Thus the tangent space will be spanned by the set of random variables

$$\left\{ \frac{\partial \ell}{\partial \theta^1}, \ldots, \frac{\partial \ell}{\partial \theta^r} \right\}.$$

Hence the set of score vectors spans the tangent space.

It will be convenient to swop freely between the two intepretations of tangents vectors, firstly random variables which are elements of the above vector space, and secondly directional derivatives. We shall use the following notation throughout,

$$\partial_i \ell(\theta) = \frac{\partial \ell}{\partial \theta^i}$$

when we wish to emphasis the random variable nature of the tangent vector.

3.2.6 Taylor's theorem

Suppose we have a real valued function h defined on our statistical manifold, M. Most asymptotic theory is defined using a form of expansion of such functions. Using the machinery we have already developed, we can look at a first order Taylor expansion which is relevant for curved manifolds. Suppose we have a parameterisation θ on the manifold then Taylor's expansion will be

$$h(\theta) = h(\theta_0) + \partial_i h(\theta_0) d\theta^i + \text{ higher order terms.}$$

We note however due the change of basis formulae in Section 3.2.4, in particular Equation 7, that this expansion will be invariant. That is to say the expansion will be the same for all parameterisations. The question of invariance to reparameterisation is an important one in Statistics, and has motivated much of our geometric development. We shall see that the extension of this formula, which is needed for higher order asymptotics, requires more knowledge of the underlying geometry of the manifold, and is considered in detail in Blæsild (1987).

3.2.7 Further structures

Having defined a single tangent space for a manifold M at the point $p \in M$, we can define the *tangent bundle*, denoted by TM. This is the collection of all tangent spaces, i.e.

$$TM = \{TM_p \mid p \in M\}.$$

The tangent bundle itself will have the structure of a manifold, in fact if M is r-dimensional then TM will be a $2r$-dimensional manifold, see Dodson and Poston (1991).

We will also use the notion of a tangent field on the manifold. Intuitively we can think of this associating a tangent vector at each point p which lies in the tangent space TM_p. Formally, a tangent field X is a (smooth) map between manifolds

$$\begin{aligned} X : M &\rightarrow TM \\ p &\mapsto X(p) \in TM_p. \end{aligned}$$

We denote the set of all tangent fields by $\chi(M)$. In a similar way we can define the cotangent bundle T^*M by

$$T^*M = \{T^*M_p \mid p \in M\},$$

and a cotangent field X^* by

$$\begin{aligned} X^* : M &\rightarrow T^*M \\ p &\mapsto X^*(p) \in T^*M_p. \end{aligned}$$

We denote the set of all cotangent fields by $\chi^*(M)$.

4 Metric Tensors

Having carefully defined the tangent and cotangent spaces we are now able to define part of the underlying geometrical structure of a statistical manifold. Given that the tangent space can be viewed as a vector space, it is natural to ask if we have some way of measuring lengths of tangent vectors and angles between tangent vectors. We will see how this will enable us to define, amongst other things, lengths of paths. The tool that we use is called a *metric tensor*.

Since each tangent space is a vector space, the simpliest way to measure lengths and angles is to prescribe a quadratic form for each tangent space. However we would require that this quadratic form varies smoothly across the manifold. This gives the definition: a metric tensor, denoted by $\langle\ ,\ \rangle$, is a smooth function

$$\begin{aligned} <\ ,\ >: \chi(M) \times \chi(M) &\rightarrow C^\infty(M) \\ (X, Y) &\mapsto < X, Y > . \end{aligned}$$

It satisfies the properties that for all $X, Y, Z \in \chi(M)$, and for all $f \in C^\infty(M)$:

1. $\langle X, Y\rangle = \langle Y, X\rangle$ (symmetry),
2. $\langle fX, Y\rangle = f\langle X, Y\rangle$ and $\langle X + Y, Z\rangle = \langle X, Z\rangle + \langle Y, Z\rangle$ (bi-linearity),
3. $\langle X, X\rangle > 0$ (positive definiteness).

This definition is best understood when you consider what happens at the tangent space at a particular point p. On this vector space, T_pM, the metric $\langle\ ,\ \rangle_p$ is simply a quadratic form. The key to the above definition lies in ensuring that this quadratic form varies smoothly across the manifold with p.

Suppose now that we introduce a parametrisation, θ. Since the metric is a bi-linear form it can be written as an element of the tensor product of 1-forms, $T^*M \quad T^*M$. We can express it in terms of the basis for T^*M given by the parameterisation, i.e. relative to $\left\{d\theta^1, \ldots, d\theta^r\right\}$. Thus we have the expression

$$\langle\ ,\ \rangle = g_{ab} d\theta^a d\theta^b, \qquad \text{where } g_{ab} = \langle \partial_a, \partial_b \rangle.$$

Just as we write any tangent vector in terms of its components with respect to the basis $\{\partial_1, \ldots, \partial_r\}$, the metric is often simply denoted by it components $g_{ab} = g_{ab}(\theta)$ relative to the basis $\{d\theta^1, \ldots, d\theta^r\}$. It is important to then understand how this "component form" representation of the metric changes with a change of parameters. Suppose that η is another parameterisation, then we have

$$\langle\ ,\ \rangle = g_{ab} \partial_i \theta^a \partial_j \theta^b d\eta^i d\eta^j = \tilde{g}_{ij} d\eta^i d\eta^j,$$

where $\tilde{g}_{ij}$ is the coordinate form for the η-parameterisation. Hence the change of basis formula for the coordinate version of the metric is

$$\tilde{g}_{ij} = \partial_i \theta^a \partial_j \theta^b g_{ab}. \tag{8}$$

Remember that we are using the Einstein convention. From the above definition we can see that as long as the components of the metric obey the transformation rule in Equation 8, then lengths and angles defined by the metric will be invariant to changes in parametrisation.

The following example shows how the metric can be induced by the geometry on the embedding space.

Example 6. The Sphere example (continued) Consider the sphere embedded in $\mathbf{R}^3$ which is given parameterically by

$$\left\{ \left(\cos(\theta^1)\sin(\theta^2), \sin(\theta^1)\sin(\theta^2), \cos(\theta^2) \right) \mid \theta^1 \in (-\pi, \pi), \theta^2 \in (0, \pi) \right\}.$$

The tangent space at the point (θ^1, θ^2) is spanned by the vectors in $\mathbf{R}^3$

$$\begin{aligned} \partial_1 &= \left(-\sin(\theta^1)\sin(\theta^2), \cos(\theta^1)\sin(\theta^2), 0 \right) \\ \partial_2 &= \left(\cos(\theta^1)\cos(\theta^2), \sin(\theta^1)\cos(\theta^2), -\sin(\theta^2) \right). \end{aligned}$$

This surface is embedded in $\mathbf{R}^3$. Let us suppose we impose on $\mathbf{R}^3$ the standard Euclidean inner product, $\langle\ ,\ \rangle_{\mathbf{R}^3}$. Since $\partial_1, \partial_2 \in \mathbf{R}^3$ we can measure their lengths and angles with this inner product. Thus we can define

$$g_{ij} = \langle \partial_i, \partial_j \rangle_{\mathbf{R}^3},$$

and we have the metric

$$\begin{pmatrix} \sin^2(\theta^2) & 0 \\ 0 & 1 \end{pmatrix}.$$

In general if the embedding space N is not just an affine space, but also has a fixed inner product $\langle\ ,\ \rangle_N$, then this will induce the *embedding metric*. In component form, relative to a parameterisation θ, this will be

$$g_{ij} = \langle \partial_i, \partial_j \rangle_N.$$

We shall return to this notion in Section 4.2.

Having a metric tensor on the manifold not only defines the length of tangent vectors it also can be used to define the length of curves. Suppose that the curve

$$\begin{aligned} \gamma : (0,1)(\subset \mathbf{R}) &\rightarrow M \\ t &\mapsto \gamma(t) \end{aligned}$$

is a smooth map such that $\gamma(0) = p_1$ and $\gamma(1) = p_2$. For any $t \in (0,1)$ the corresponding tangent vector is $\frac{d\gamma}{dt}(t)$, with length

$$\sqrt{\langle \frac{d\gamma}{dt}(t), \frac{d\gamma}{dt}(t) \rangle_{\gamma(t)}}$$

and the length of the curve from p_1 to p_2 is defined as

$$\int_0^1 \sqrt{\langle \frac{d\gamma}{dt}(t), \frac{d\gamma}{dt}(t) \rangle_{\gamma(t)}}\, dt.$$

Due to the invariance properties of the metric this will be an invariant measure of arclength.

Given two such points on a manifold we can ask the natural geometric question: what is the shortest path connecting them? Although this is a natural question in general the complete answer is rather complex. There are examples where there will not be a shortest path in a manifold, or there may be many. There is a simpler answer, however, to the question: Is a path of *locally* minimum length, in a variational sense? That is, its length is a local minimum in the class of all paths joining p_1 and p_2. A path which satisfies this condition is called a *geodesic*.

Given a metric on a manifold its geodesics can be characterised by a set of differential equations. We will see this representation in Section 5.

4.1 Tensors

Tangent vectors, cotangent vectors and metrics are all examples of the general geometric notion of a *tensor*. All of these examples show how the careful definition of transformation rules allows definitions of invariant quantities to be made. This invariance to the choice of parameterisation is not only very important geometrically but also for statistical purposes. In general we would not want the tools of inference to depend on arbitrary choices in the model specification. Using a geometric approach will, almost by definition, avoid these problems, for

an example of this see Critchley, Marriott and Salmon (1996). For a good reference to where tensors have had a direct application to statistical theory, see McCullagh (1987).

A general tensor is made up of multivariate linear functions and functionals of the tangent fields. They come in two types *covariant* tensors, which on each tangent space are high dimensional generalisations of matrices. For example, the metric is simply a bilinear form which varies smoothly with the point of evaluation. The second type are called *contravariant* tensors which are simply products of tangent vectors when evaluated at each point. We have the following formal definitions;

A *covariant tensor* of degree r on T_pM is an r-linear real valued function on the r-fold product $T_pM \times \ldots \times T_pM$. A *contravariant tensor field* of degree r is an r-linear real valued function on the r-fold product $T_p^*M \times \ldots \times T_p^*M$. A tensor field is a set of tensors on T_pM which varies smoothly with $p \in M$.

4.1.1 Components of tenors

To work with these tensors in computations it is necessary to write them out in component form with respect to the bases which span TM_p and TM_p^*. Given a parameter system θ these bases will be

$$\{\partial_1, \ldots, \partial_r\}$$

and

$$\left\{d\theta^1, \ldots, d\theta^r\right\}.$$

Thus a k-covariant tensor can be written as

$$A_{i_1 i_2 \ldots i_k} d\theta^{i_1} d\theta^{i_2} \ldots d\theta^{i_k}$$

This is, by definition, an k-linear real valued function on the k-fold product $T_pM \times \ldots \times T_pM$ which acts on tangent vectors in the following way. If we have a set of k tangent vectors $v_1, \ldots, v_k$, which we write with respect to the basis as

$$v_i = v_i^j \partial_j,$$

then the tensor acts on the set of vectors as

$$A_{i_1 i_2 \ldots i_k} d\theta^{i_1} d\theta^{i_2} \ldots d\theta^{i_k} (v_1, \ldots, v_k) = A_{i_1 i_2 \ldots i_k} v_1^{i_1} \ldots v_k^{i_k},$$

recalling again the Einstein convention.

To see how the components of this tensor will transform under a change of parameters, we apply the same methodology as we did for the metric. Let θ be the original parameterisation and let the components of the tensor with respect to this parameterisation be $A_{a_1 a_2 \ldots a_k}$. Let η be the new parameterisation, with components $\tilde{A}_{i_1 i_2 \ldots i_k}$. This will give a transformation rule

$$\tilde{A}_{i_1 i_2 \ldots i_k} = \partial_{i_1} \theta^{a_1} \partial_{i_2} \theta^{a_2} \ldots \partial_{i_k} \theta^{a_k} A_{a_1 a_2 \ldots a_k}. \tag{9}$$

For a contravariant tenor the component form will be

$$A^{i_1 i_2 \ldots i_k} \partial_{i_1} \ldots \partial_{i_k}$$

which acts on k cotangent vectors $v^1, \ldots, v^k$ which we write as

$$v^i = v_j^i d\theta^j.$$

The action is then

$$A^{i_1 i_2 \ldots i_k \partial_{i_1} \ldots \partial_{i_k}} \left(v^1, \ldots, v^k\right) = A^{i_1 i_2 \ldots i_k} v_{i_1}^1 \ldots v_{i_k}^k,$$

and the transformation rule will then be

$$\tilde{A}^{i_1 i_2 \ldots i_k} = \partial_{a_1} \eta^{i_1} \partial_{a_2} \eta^{i_2} \ldots \partial_{a_k} \eta^{i_k} A^{a_1 a_2 \ldots a_k}. \tag{10}$$

4.1.2 Raising and lower indices

The general theory of dual spaces tells us that if V is a vector space and V^* its dual space, there will be a natural isomorphism between V and V^{**}, the dual of the dual space defined, by

$$\begin{aligned} V &\rightarrow V^{**} \\ v &\mapsto \alpha_v. \end{aligned}$$

where $\alpha_v : V^* \rightarrow \mathbf{R}$ is defined by $\alpha_v(A) = A(v)$ for all $A \in V^*$. However in general there will not be a *natural* isomorphism between V and V^* unless there is an inner product $\langle\ ,\ \rangle$ on V. In that case the isomorphism is defined by

$$\begin{aligned} V &\rightarrow V^* \\ v &\mapsto \langle v, . \rangle \end{aligned}$$

where we interpret $\langle v, . \rangle$ as an element of V^* since

$$\begin{aligned} \langle v, . \rangle : V &\rightarrow \mathbf{R} \\ w &\mapsto \langle v, w \rangle. \end{aligned}$$

For a manifold with a metric, of course, we do have an inner product on each of the vector spaces defined by the tangent space. This will by the above theory therefore allow us to define a *natural* isomorphism between tangent space TM and dual space T^*M. In more generality it will enable us to convert between covariant and contravariant tensors. Thus, if in component terms we have a 1-covariant tensor, given by A_i, then it can be converted to a contravariant tensor by using the metric. In detail we get $A^j = g^{ij}A_i$ where g^{ij} is the inverse to the metric g_{ij}.

4.2 Statistical metrics

Having seen the concept of a metric and a tensor in generality, we turn now to statistical applications. In this section we shall see that these tools are not new to statistical theory and that they are essential to understand the nature of invariance. We look first at three examples of metrics in statistics and their application.

4.2.1 Expected Fisher information

We have already seen the Expected Fisher information matrix in Section 3.1

$$I_{ij} = E_{p(x,\theta)}\left(-\frac{\partial^2 \ell}{\partial\theta^i \partial\theta^j}\right) = \mathrm{Cov}_{p(x,\theta)}\left(\frac{\partial \ell}{\partial\theta^i}, \frac{\partial \ell}{\partial\theta^j}\right).$$

These are the components of a covariant 2-tensor which can be easy checked by seeing how it transforms under a change of parameters using Equation (9). This is a simple exercise in the chain rule. Further it will, under regularity, be positive definite and a smooth function of the parameter. Hence it has the properties of the components of a metric tensor. In general we will abuse notation and refer to I_{ij} as being a metric, dropping reference to components and explicit parameter dependence.

For a full exponential family, in the natural parameterisation, the Fisher Information Matrix will be given by

$$I_{ij}(\theta) = \frac{\partial^2 \psi}{\partial\theta^i \partial\theta^j}(\theta).$$

Applying Equations (3) and (8) shows immediately that the components in the expected η-parameters will be

$$\left(\frac{\partial^2\psi}{\partial\theta^i\theta^j}(\theta(\eta))\right)^{-1}.$$

Using the convexity of ψ we see that the positive definiteness is assured, hence in the full exponential family case the Fisher Information Matrix is a metric tensor.

Although this has the formal properties of being a metric, we have to ask in what way does this act on the tangent space? In particular what statistical interpretation can be given? This will become clear if we use the interpretation of the tangent vectors as being random variables, elements of the space spanned by

$$\left\{\frac{\partial\ell}{\partial\theta^1},\ldots,\frac{\partial\ell}{\partial\theta^r}\right\}.$$

The Fisher information is the covariance for this space of random variables. Due to the additive properties of the log-likelihood, the Central Limit Theorem tells us that asymptotically linear combinations of score vectors will be Normally distributed. Equation (4) tells us that the mean will be zero, thus the metric completely determines the stochastic behaviour of the space, at least to first order. This interpretation, although being completely standard does need careful refinement in our geometric viewpoint. We will return to this issue below.

Having seen the Fisher information metric in the case of full exponential families it is natural to ask what form it takes in a curved exponential family? Given an (r,t)-curved exponential family

$$p(x\mid\xi)=\exp\left\{\theta^i(\xi)s_i-\psi(\theta(\xi))\right\}m(x),$$

the tangent space to this family will be spanned by the set

$$\left\{\frac{\partial\ell}{\partial\xi^1},\ldots,\frac{\partial\ell}{\partial\xi^t}\right\},$$

where

$$\frac{\partial\ell}{\partial\xi^i}=\frac{\partial\theta^j}{\partial\xi^i}\left(s_j-\frac{\partial\psi}{\partial\theta^j}\right)=\frac{\partial\theta^j}{\partial\xi^i}\partial\ell_j.$$

The construction of an embedding metric, as seen in Example 6, can be used here. The metric on the curved family is completely determined by that on the embedding full exponential family. In component form this is given by g_{ij}, where

$$g_{ij}=\frac{\partial\theta^k}{\partial\xi^i}\frac{\partial\theta^l}{\partial\xi^j}I_{kl}. \tag{11}$$

4.2.2 Applications

We now give two examples where the metric properties of the Fisher information have found application in Statistics. The first concerns the question of an invariant definition of length, and the second an invariant definition of angle.

The first example concerns the score test. Suppose we have an $(r,1)$-curved exponential family,

$$p(x\mid\xi)=\exp\{\theta^i(\xi)s_i(x)-\psi(\theta(\xi))\},$$

and we wish to test the null hypothesis $\xi=\xi_0$ using the score test. The components of the score vector will be

$$S(\theta(\xi_0))=\left(s_i-\frac{\partial\psi}{\partial\theta^i}(\theta(\xi_0))\right)\frac{d\theta^i}{d\xi}(\theta(\xi_0)).$$

The variance of the score is the Fisher information matrix, which can be calculated using Equation (11) as

$$V = \frac{d\theta^k}{d\xi}\frac{d\theta^l}{d\xi}I_{kl}.$$

Hence the score statistic is $S'V^{-1}S$. This is an invariant quantity for any reparameterisation of ξ. It has the geometric interpretation of the length of a cotangent vector which is a tensor. For an example of how this invariance is important in Econometrics see Critchley, Marriott and Salmon (1996).

The second example is where an angle is measured invariantly. Parameter orthogonality is a useful property when estimation is being made in the presence of nuisance parameters. For an example of this see Barndorff-Nielsen and Cox (1994, page 49).

4.2.3 Preferred point metrics

The Fisher information metric is not the only possible statistically based metric. The concept of a *preferred point geometry* and the related *preferred point metric* was introduced in Marriott (1989). This is a very natural geometric and statistical construction for the embedding given in Section 2.4.2. The embedding space can be viewed as a space of random variables. These random variables are simply functions of the data. It seems natural then that the data generating process (DGP) plays some role in defining the geometry. Let us suppose that we have a parametric family $p(x \mid \theta)$ and the data generating process is given by a member of this family, namely $p(x \mid \phi)$. The point ϕ, although unknown, plays a distinctive role in the family $p(x \mid \theta)$. It is called the *preferred point*. The tangent space at a general point θ is spanned by the random variables

$$\left\{\frac{\partial \ell}{\partial \theta^1}(\theta), \ldots, \frac{\partial \ell}{\partial \theta^r}(\theta)\right\}.$$

We can define the covariance matrix of this space relative to the DGP, i.e.,

$$I^{\phi}_{ij}(\theta) = \mathrm{Cov}_{p(x \mid \phi)}\left(\frac{\partial \ell}{\partial \theta^i}(\theta), \frac{\partial \ell}{\partial \theta^j}(\theta)\right). \tag{12}$$

Notice the important difference between this and the expected Fisher information matrix is that the covariance here is evaluated with respect to some fixed distribution for *all* tangent spaces. In the Fisher information case the evaluating distribution varies with tangent space.

By the same argument as for the Fisher information, Equation (12)defines the components of a metric tensor, at least for all θ in a neighbourhood of ϕ. The properties of such preferred point metrics are explored in Critchley, Marriott and Salmon (1993, 1994 and in this volume).

We note further that we must reconsider Equation (4) in the preferred point context. The mean value of the score vector will only be zero in the tangent space of the preferred point. In general we define

$$\mu^{\phi}_{i}(\theta) = E_{p(x \mid \phi)}\left(\frac{\partial \ell}{\partial \theta^i}(\theta)\right). \tag{13}$$

This defines a covariant 1-tensor on the manifold.

For a full exponential family it is easy to check that the preferred point metric, in the natural parameters, is given by

$$I^{\phi}_{ij}(\theta) = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\phi).$$

Note that this is independent of θ and thus is constant across all tangent spaces. We return to this property later.

4.2.4 Observed information metrics

As mentioned in Section 3.1, Efron and Hinkley (1978) argue that in general the observed Fisher information

$$\mathcal{I}_{ij} = -\frac{\partial^2 \ell}{\partial\theta^i \partial\theta^j}(\hat{\theta}),$$

is a better measure of the covariance of the score vector at $\hat{\theta}$, since it reflects the conditional distribution of the score. This will also give a statistically based metric. Here though the change of parameterisation rule requires some care. By applying the chain rule to

$$\frac{\partial^2 \ell}{\partial\theta^i \partial\theta^j}(\theta)$$

twice we have the following formula

$$\frac{\partial^2 \ell}{\partial\eta^i \partial\eta^j}(\eta(\theta)) = \frac{\partial\theta^a}{\partial\eta^i}\frac{\partial\theta^b}{\partial\eta^j}\frac{\partial^2 \ell}{\partial\theta^a \partial\theta^b}(\theta) + \frac{\partial^2\theta^a}{\partial\eta^i \partial\eta^j}\frac{\partial \ell}{\partial\theta^a}(\theta).$$

The second term on the right hand side of this expression disappears when evaluated at $\hat{\theta}$ giving the correct covariant 2-tensor transformation rule. At all other values of θ however this will not transform as a tensor.

For a full exponential family the metric has components

$$\mathcal{I}_{ij}(\hat{\theta}) = \frac{\partial^2 \psi}{\partial\theta^i \partial\theta^j}(\hat{\theta}). \tag{14}$$

the same as the Expected Fisher metric. However for a curved exponential family the metrics will differ. The observed information metric will be a stochastic function, whereas the expected form will be deterministic.

The use of the observed information metric is best explained in the context of conditional inference. Let us suppose that we can write the minimum sufficient statistic, $(s_1(x), \ldots, s_r(x))$, for the curved exponential family $p(x \mid \xi)$ in the form $(\hat{\xi}(x), a(x))$ where a is any ancillary statistic for ξ and $\hat{\xi}$ is the maximum likelihood estimate. We have seen in Section 3.1 that the tangent space is invariant under conditioning on an exact ancillary statistic. Thus as an asymptotic approximation the observed metric will be the *conditional* covariance of the tangent vector and this will be the observed information matrix.

4.2.5 Bibilographic notes

The fact that the Expected Fisher information has the properties of a metric tensor was first observed by Rao (1945, 1987). The full theory of, so called, *expected* geometry was developed in a series of papers by Amari, see Amari (1987,1990). Preferred point metrics were introduced by Marriott (1989). The relationship between these two forms of geometry is explored in the companion chapter in this volume, Critchley, Marriott and Salmon. The *observed* geometry was developed by Bardorff-Nielsen (1987) and were developed to a more general theory of *yokes*, Blæsild (1987,1991).

5 Affine connections

In Section 2 it was stated that full exponential families can be thought of as affine subspaces, and that curved exponential families are curved sub-families of these affine subspace. In this section we shall see the formalisation of these statements. We have already studied

the construction of the manifold, tangent bundle and metric structure of a parametric family. There is one final structure which is needed before a satisfactory geometric theory for statistics can be developed. This is the idea of *curvature*, and the absence of curvature, *straightness* and *flatness*. The tool used in differential geometry to describe these features is called an *affine connection*.

Consider for motivation a one dimensional subfamily of a parametric family $p(x \mid \theta)$. We have already seen two ways in which we might consider this to be 'straight'. Firstly we have the affine structure of the embedding space. Affine subspaces are often considered to be straight. Secondly we have the concept of a geodesic with respect to a metric tensor. Loosely a geodesic is a curve of minimum length, and this is often intuitively thought of as being equivalent to straightness. However detailed investigation shows that for parametric families these two notions are quite different. A curve which is 'straight' according to this first criterion, will not be 'straight' according to the second. Amari (1990) realised that the way out of this seeming paradox lay in the careful definition of the correct connection structure for statistics.

In this section we shall first examine the general geometric theory of connections and then define the relevant statistical examples.

5.1 Geometric affine connections

We take as our fundamental notion of straightness for a curve that its tangent vectors are always parallel to some fixed direction. However to formalise this we must therefore have a way of measuring the rate of change of a tangent field. An affine connection gives us that tool. It can be thought of as a way to calculate the rate of change of one tangent field with respect to another. We would want the result of the differentiation itself to be a tangent field.

Formally we define a symmetric affine connection ∇ to be a function

$$\begin{aligned} \nabla : \chi(M) \times \chi(M) &\rightarrow \chi(M) \\ (X, Y) &\mapsto \nabla_X Y, \end{aligned}$$

which satisfies the following properties: For all $X, X_1, X_2, Y, Y_1, Y_2 \in \chi(M)$ and $f \in C^\infty(M)$, we have

1. $\nabla_{X_1+X_2} Y = \nabla_{X_1} Y + \nabla_{X_2} Y$,
2. $\nabla_X (Y_1 + Y_2) = \nabla_X Y_1 + \nabla_X Y_2$,
3. $\nabla_{fX} Y = f \nabla_X Y$,
4. $\nabla_X (fY) = f \nabla_X Y + X(f) Y$,
5. $\nabla_X Y - \nabla_Y X = XY - YX$.

Note that here we are explicitly using the derivative version of the tangent vector given in Section 3.2.2. Thus $X(f)$ is the directional derivative of the function f in the direction stipulated by the tangent vector X. Conditions $1, 2$ and 3 define the linearity properties of the connection. Condition 4 states that the connection is a derivative and hence satisfies the product rule. Condition 5 is the condition which make the connection symmetric. We shall only consider symmetric connections.

Just as in the case of a metric we can write a connection in component form. Let us suppose that the vector fields X and Y are given with respect to the basis defined by the parameterisation θ, i.e.,

$$\begin{aligned} X &= X^i \partial\theta_i \\ Y &= Y^i \partial\theta_i. \end{aligned}$$

By definition $\nabla_X Y$ is also a tangent field. If we define

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k$$

then by the above properties $1-5$ it must follow that

$$\nabla_X Y = \left(X^i \partial_i Y^k + X^i Y^j \Gamma_{ij}^k \right) \partial_k. \tag{15}$$

Thus the connection is completely determined by the 3-way array Γ_{ij}^k. These are called the Christoffel symbols for the connection. Note that condition 5 implies that for all i, j and k

$$\Gamma_{ij}^k = \Gamma_{ji}^k,$$

hence the connection is called symmetric.

It is important to note that the Christoffel symbols do not transform as tensors. If Γ_{ab}^c are the *Christoffel symbols* relative to the θ-parameters and $\tilde{\Gamma}_{ij}^k$ relative to the η-parameters, then the two sets of components are connected according to the rule

$$\Gamma_{ab}^c = \partial_k \theta^c \left(\partial_{ab}^2 \eta^k + \partial_a \eta^i \partial_b \eta^j \tilde{\Gamma}_{ij}^k \right). \tag{16}$$

One consequence of this not being a tensor rule, is that Christoffel symbols can be identically zero with respect to one parameterization, but non-zero in another.

Having defined a connection, we can see how they are able to define a straight line, or as it is called a *geodesic.* Let us define a path in terms of a parameterisation,

$$\begin{aligned} \gamma : [0,1] (\subset \mathbf{R}) &\rightarrow M \\ t &\mapsto \theta(t) \end{aligned}$$

Note that this is a slight abuse of notation, identifying $(\theta^1(t), \ldots, \theta^r(t))$ with the point in M with those parameters. The path will be a geodesic if it satisfies the equation

$$\frac{d^2\theta^k}{dt^2} + \frac{d\theta^i}{dt} \frac{d\theta^j}{dt} \Gamma_{ij}^k = 0. \tag{17}$$

If the tangent vector to this path is $X(t)$ this equation is the component form of the equation

$$\nabla_X X = 0. \tag{18}$$

5.1.1 Connections and projections

In order to understand the nature of a connection operator it is helpful to return to the sphere example and see how the most natural connection there operates on tangent fields.

Example 6. Sphere example (continued). For the sphere embedded in $\mathbf{R}^3$ with parameterisation

$$\left\{ \left(\cos(\theta^1) \sin(\theta^2), \sin(\theta^1) \sin(\theta^2), \cos(\theta^2) \right) \mid \theta^1 \in (-\pi, \pi), \theta^2 \in (0, \pi) \right\},$$

consider a path

$$\left(\theta^1, \theta^2\right) = (t, \pi/2)\,.$$

In $\mathbf{R}^3$ it is the path

$$(\cos(t), \sin(t), 1)\,,$$

the equator of the sphere. Suppose we differentiate with respect to t, giving a vector in $\mathbf{R}^3$

$$X(t) = (-\sin(t), \cos(t), 0)\,.$$

This is the tangent vector to the curve. A connection allows us to differentiate tangent fields. The result of this differentiation should itself be a tangent vector, since the tangent space gives the set of all directions in the manifold. However differentiating $X(t)$ directly gives

$$\dot{X} = (-\cos(t), -\sin(t), 0).$$

This does not lie in the tangent space, which is spanned by

$$\partial_1 = (-\sin(t), \cos(t), 0) \quad \text{and} \quad \partial_2 = (0, 0, -1)\,.$$

If we want the derivative of the tangent field to be a tangent field we must project $\dot{X}(t)$ back into TM_p. Using the standard inner product on $\mathbf{R}^3$, $\langle\ ,\ \rangle_{\mathbf{R}^3}$ we find that $\dot{X}(t)$ is in fact orthogonal to the tangent space. Hence the rate of change of X, relative to the sphere will be zero. In other words in the sphere the curve is straight. This is completely consistent with intuition. The equator is a *great circle* which is well known to provide the shortest distance between two points on a sphere.

The intuition from the above example gives an interpretation of Equation 15. It can be viewed as a standard derivative operation combined with a projection onto the tangent space. The exact form of the projection will depend of the form of the Christoffel symbols. For details of this construction see Dodson and Poston (1991). We return to this view of the connection in Section 5.2.1.

5.1.2 Metric connections

Having seen how a connection defines a geodesic we return to the second definition of straightness. In Section 4 a geodesic was defined as a path of (locally) minimum length. It can be shown that given a metric there is a unique affine connection ∇^0, called the Levi-Civita connection, for which the two definitions of geodesic agree. That is to say, defining a metric automatically defines a particular connection, such that if a path satisfies Equation 18 it will be a path of minimum length. We quote here two important ways of characterising the Levi-Civita connection.

Theorem (i) For a manifold M with a metric $\langle\ ,\ \rangle_p$ the Levi-Civita connection ∇^0 is characterised as the only symmetric connection which satisfies

$$X\langle Y, Z\rangle_p = \langle \nabla^0_X Y, Z\rangle_p + \langle Y, \nabla^0_X Z\rangle_p, \tag{19}$$

for all tangent fields X, Y, Z.

(ii) If we introduce a parameter system θ the Christoffel symbols of the Levi-Civita connection are given by

$$\Gamma^{0\,k}_{\ ij} = g^{kl}\left(\frac{\partial g_{il}}{\partial\theta^j} + \frac{\partial g_{jl}}{\partial\theta^i} - \frac{\partial g_{ij}}{\partial\theta^l}\right), \tag{20}$$

where g_{ij} are the components of the metric. The proof of this theorem can be found in Dodson and Poston (1991).

5.1.3 Non-metric connections

Connections however do not have to be constructed in this way. Any connection which is not the Levi-Civita connection of some underlying metric will be called a non-metric connection. Formally all that is required is the definition of a set of Christoffel symbols which transform under a change of parameters according to Equation 16. We shall see in the following section that for statistical purposes non-metric connections play a more important role than metric ones.

5.2 Statistical connections

As we have previously noted we have two distinct notions of straightness or flatness for our statistical manifolds. Firstly the idea of flatness which comes from the affine structure of the embedding space, and secondly from the Levi-Civita connection of a statistically based metric. Initially we consider the case with the expected Fisher information as the metric tensor on the manifold which represents the parametric family.

5.2.1 The $+1$-connection

We wish to define a connection which is consistent with the affine structure of the embedding space. Relative to this embedding we have the tangent vector

$$\partial_i = \frac{\partial \ell}{\partial \theta^i},$$

and hence

$$\partial_i \partial_j = \frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta; x),$$

a random variable. This, as we have seen in Section 5.1.1, will not necessarily lie in the tangent space. The embedding structure gives a possible way of projecting into TM_θ. It is natural to use

$$\langle f, g \rangle = Cov_{p(x \mid \theta)}(f, g)$$

for all elements f, g of X. Using this we project $\partial_i \partial_j$ onto the tangent space, which is spanned by the set

$$\left\{ \frac{\partial \ell}{\partial \theta^1}(\theta), \ldots, \frac{\partial \ell}{\partial \theta^r}(\theta) \right\}.$$

This will then define a connection whose Christoffel symbols are

$$\Gamma^{+1}{}^k_{ij}(\theta) = I^{kl}(\theta) E_{p(x \mid \theta)} \left(\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta) \frac{\partial \ell}{\partial \theta^l}(\theta) \right). \tag{21}$$

This connection will be denoted by ∇^{+1}.

Calculating the Christoffel symbols for a full exponential family in its natural parameterisation and exploiting Equation 4, gives that for all $i, j, k \in \{1, \ldots, r\}$,

$$\Gamma^{+1}{}^k_{ij}(\theta) = 0.$$

We will return to this result later.

5.2.2 The 0-connection

In Section 5.1.3 it was shown that the metric tensor gives rise to the Levi-Civita connection. The Christoffel symbols are defined by Equation 20. If we take I, the expected Fisher information matrix as our metric, then the associated Levi-Civita connection will be given by

$$\Gamma^{+1}{}^{k}_{ij} = I^{kl}(\theta) E_{p(x \mid \theta)} \left(\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta) \frac{\partial \ell}{\partial \theta^l}(\theta) \right) + \frac{1}{2} I^{kl}(\theta) E_{p(x \mid \theta)} \left(\frac{\partial \ell}{\partial \theta^i}(\theta) \frac{\partial \ell}{\partial \theta^j}(\theta) \frac{\partial \ell}{\partial \theta^l}(\theta) \right).$$

For the full exponential family in the natural parameterisation the Christoffel symbols are given by

$$\Gamma^{0}{}^{k}_{ij} = \frac{1}{2} I^{kl} \frac{\partial^3 \psi}{\partial \theta^i \partial \theta^j \partial \theta^l}.$$

This connection will be denoted by ∇^0.

5.2.3 The -1-connection

Again considering connections as the combination of a derivative operation followed by a projection to the tangent space gives us yet another plausible connection.

As in Section 5.2.1 we wish to project the random variable

$$\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta)$$

into the tangent space. From Equation 4 it is clear that any member of the tangent space *must* have zero expectation. Hence this motivates the projection

$$\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta) \mapsto \frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta) - E_{p(x \mid \theta)} \left(\frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta) \right).$$

This operation is enough to define a connection, ∇^{-1}, and its Christoffel symbols will be

$$\Gamma^{-1}{}^{k}_{ij} = I^{kl} E_{p(x, \mid \theta)} \left(\left\{ \frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta) + \frac{\partial \ell}{\partial \theta^i}(\theta) \frac{\partial \ell}{\partial \theta^j}(\theta) \right\} \frac{\partial \ell}{\partial \theta^l}(\theta) \right).$$

In the full exponential family, again in the natural coordinates, this has the form

$$\Gamma^{-1}{}^{k}_{ij}(\theta) = I^{kl}(\theta) \frac{\partial^3 \psi}{\partial \theta^i \partial \theta^j \partial \theta^l}(\theta).$$

Notice the similarity to the ∇^0 case.

5.2.4 The α-connection

Amari (1987) pointed out that these different connections were all special cases of a large family of connections which have many statistical uses. They are called the α-connections, and denoted by ∇^α. They can be though of as a linear combination of any two of the above types. The general definition is that the Christoffel symbols are given by

$$\Gamma^{\alpha}{}^{k}_{ij} = I^{kl} E_{p(x, \mid \theta)} \left(\left\{ \frac{\partial^2 \ell}{\partial \theta^i \partial \theta^j}(\theta) + \frac{1-\alpha}{2} \frac{\partial \ell}{\partial \theta^i}(\theta) \frac{\partial \ell}{\partial \theta^j}(\theta) \right\} \frac{\partial \ell}{\partial \theta^l}(\theta) \right).$$

In the full exponential family this simplifies to be

$$\Gamma^{\alpha}{}^{k}_{ij}(\theta) = \frac{1-\alpha}{2} I^{kl} \frac{\partial^3 \psi}{\partial \theta^i \partial \theta^j \partial \theta^l}(\theta).$$

in the natural parametrisation.

5.2.5 Statistical manifolds

The development in the previous section concentrated on expected geometry. As we have seen this is not the only sensible or important geometrical structure for a parametric family of distributions. However the structure which was developed does seem to have a general applicability across most of the possible geometric structures. This was recognised by Lauritzen (1987). He defined a general structure which encompassed most forms of statistical geometry. A *statistical manifold* is defined as (M, g, T) where M is a manifold of distribution functions, g is a metric tensor and T is a covariant 3-tensor which is symmetric in its components and called the skewness tensor. For this structure there will always be set of affine connections which parallels the structure in the previous section.

Theorem For any statistical manifold (M, g, T) there is a one dimensional family of affine connections defined by

$$\Gamma^{\alpha}{}_{ijk} = \Gamma^{0}{}_{ijk} - \frac{\alpha}{2} T_{ijk}$$

where ∇^0 is the Levi-Civita connection for the metric g.

In the case of expected geometry we have that

$$T_{ijk} = E_{p(x \mid \theta)} \left(\frac{\partial \ell}{\partial \theta^i} \frac{\partial \ell}{\partial \theta^j} \frac{\partial \ell}{\partial \theta^k} \right).$$

For the observed geometry we assume that the data can be written in the conditional resolution form $x = (\hat{\theta}, a)$ where a is ancillary. The likelihood can then be written as

$$\ell(\theta \ ; \ x) = \ell(\theta \ ; \ \hat{\theta}, a).$$

We use the following notation

$$\ell_i = \frac{\partial}{\partial \theta^i} \ell(\theta \ ; \ \hat{\theta}, a), \quad \text{and} \quad \ell_{,i} = \frac{\partial}{\partial \hat{\theta}^i} \ell(\theta \ ; \ \hat{\theta}, a). \tag{22}$$

In this notation it can be shown that $g_{ij} = -\ell_{ij} = \ell_{i,j}$ is the observed fisher information metric, and $T_{ijk} = \ell_{ijk}$ is the skewness tensor.

6 Three key results

In this section we look at the key mathematical theorems of statistical manifolds, the applications to statistics are shown in the last two sections. Throughout we shall assume that we have the structure of a statistical manifold (M, g, T) as defined in Section 5.2.5.

6.1 Duality

The concept of duality for a statistical manifold is tied up with that of *flatness*. A connection defines the geometry of a manifold, in particular it defines the curvature. There are many ways in which we can measure the curvature of a manifold. One issue of great interest is when there is no curvature which is measured using the Riemann curvature tensor. This is a covariant 2-tensor defined in component terms using a connection as

$$R_{ijkl} = g_{lm} \left(\partial_i \Gamma^m_{jk} - \partial_j \Gamma^m_{ik} \right) + \left(\Gamma_{iml} \Gamma^m_{jk} - \Gamma_{jml} \Gamma^m_{ik} \right). \tag{23}$$

If this tensor is identically zero for all θ and all $i, j, k, l \in \{1, \ldots, r\}$ then the manifold is said to be *flat*. It is an important theorem of differential geometry that if a manifold is

flat then there exists a parameterisation θ which has the property that the components of the Christoffel symbol in this parametrisation will be identically zero, see Section 7. If the connection which defines the Riemann curvature is the Levi-Civita connection of the metric g, then this theorem extends. In this case there exists a parameterisation which has the property that the components $g_{ij}(\theta)$ will be independent of θ. So that the metric will be a constant. In either of these cases the parameterisation is called an *affine parameterisation.*

The duality theorem for statistical manifolds then states

Theorem If (M, g, T) is a statistical manifold such that M is flat with respect to the connection ∇^{α} then it will be flat with respect to $\nabla^{-\alpha}$.

We say that the $-\alpha$-connection is *dual* to the α-connection. The 0-connection, which is the Levi-Civita connection derived from the metric will be self dual.

One important application of this theorem is to the full exponential family. In Section 5.2.1 it is shown that in the natural parameterisation the Christoffel symbols of the ∇^{+1}-connection are identically zero. Thus for this connection a full exponential family is flat and the natural parameters are affine. The theorem states that a full exponential family will also be flat with respect to the -1-connection. There is also therefore a set of parameters which is affine for this connection. These parameters are the expectation parameters defined in Section 2.1.2.

The relationship between a dual pair of connections is illuminated by the following result, which should be compared to the defining equation for the Levi-Civita connection, Equation (19).

Theorem If the metric is denoted by $\langle\ ,\ \rangle$ then for a statistical manifold (M, g, T)

$$X\langle Y, Z\rangle = \langle \nabla^{\alpha}_X Y, Z\rangle + \langle Y, \nabla^{-\alpha}_X Z\rangle. \tag{24}$$

6.2 Projection theorem

In the following two sections we concentrate on the statistical manifold structure which is defined by Amari (1990).

The geodesic distance in a manifold is one way to measure distances between points of the manifold. However for probability distributions there are other more statistically natural ways of defining a form of distance. Consider the following definition:

An α-divergence between $p(x)$ and $q(x)$, two density functions is defined as

$$D_{\alpha}(p, q) = \begin{cases} E_q\left(\log(q) - \log(p)\right) & \alpha = 1 \\ E_p\left(\log(p) - \log(q)\right) & \alpha = -1 \end{cases} \tag{25}$$

Note that these divergences are the Kullback-Leibler information from q to p when $\alpha = 1$ and from q to p when $\alpha = -1$. These are not formally distance functions as they do not obey the standard axioms for a distance, such as symmetry or the triangle inequality. For a clarification of their geometric role see Critchley, Marriott and Salmon (1994).

Amari's projection theorem connects these divergence functions with the geodesic structure. We give a simplified version of his theorem here, for exponential families. For more details see Amari (1990, page 89).

Theorem Let $p(x \mid \theta)$ be a full exponential family, and $p(x \mid \xi)$ a curved sub-family. Let θ_1 represent a distribution in the embedding family. Let ξ_{α} be an extremal point for the α-divergence, that is in Ξ, ξ_{α} is an extremal point of the function

$$D_{\alpha}(p(x \mid \theta_1), p(x \mid \xi)).$$

If ξ_α is an α extremal point then the α geodesic connecting $p(x \mid \theta_1)$ and $p(x \mid \xi_\alpha)$ cuts the family $p(x \mid \xi)$ orthogonally.

One natural application here is in the case of a misspecified model. Suppose that the model is incorrectly specified as being the curved exponential family $p(x \mid \xi)$, while the data generation process lies in the embedding family, and is denoted by $p(x \mid \theta_0)$. It is well known that the MLE on the misspecified model will converge to the value which is closest to $p(x \mid \theta_0)$ in terms of the Kullback-Liebler divergence. Hence it will be a -1-extremal point and the DGP and the pseudo-true estimate are connected by an -1-projection.

6.3 Pythagoras theorem

The final result presented in this section brings to the geometry of statistical families the classical geometric theorem of Pythagoras. We state the theorem for completeness.

Theorem Let θ_1, θ_2 and θ_3 be three points in the parameter space of a full exponential family. Suppose that the α-geodesic joining θ_1 and θ_2 is orthogonal to the $-\alpha$-geodesic joining θ_2 and θ_3. For this triangle we have the Pythagorean relationship

$$D_\alpha(\theta_1, \theta_2) + D_\alpha(\theta_2, \theta_3) = D_\alpha(\theta_1, \theta_3). \tag{26}$$

A more complete version of this theorem can be found in Amari (1990, page 86).

7 Inference in curved families

In final two sections of this paper we look at how the machinery previously set up enables us to understand both the problem of inference in curved families and how to assess the information content of a general statistic. This issue is taken up in several papers in this volume, in particular by Kees Jan Van Garderen and Tom Rothenberg

7.1 Curvature due to parameterisations

We have seen how the connection structure of a manifold enables us to define curvature, both of a curve and a set of parameters. We look first at the issue of parameter dependent curvature.

Example 7. Non-linear regression models Bates and Watts (1988, page 241) distinguish between two types of parameterisation which are important in the analysis of non-linear regression models. These curvatures are called the *parameter effects* and *intrinsic* curvatures. The first of these will be dependent on the exact form of the parameterisation and can be removed in principle by reparameterisation, the second will be independent of the parameterisation chosen. This distinction carries over from the nonlinear regression case to a general statistical manifold.

One example where parameter effects are important is the bias of the maximum likelihood estimate. In the case of a full exponential family, the first order bias of the maximum likelihood estimate is given by

$$b^i(\xi) = \frac{-1}{2n} g^{ij} g^{kl} \Gamma^{-1}_{jkl}. \tag{27}$$

Consider this formula in the context of linear models.

Example 1. The linear model (continued) There is a natural geometric interpretation for ordinary least squares (OLS) estimation, which connects Euclidean geometry and regression,

see for example Bates and Watts (1988). For presentational simplicity we consider the case where the variance parameter is fixed and known. We have the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

There is an n-dimensional space of possible response vectors Y and a $(k+1)$-dimensional sub-manifold of this affine space is defined by the expectation surface

$$\boldsymbol{\beta} \rightarrow X\boldsymbol{\beta} \subset \mathbf{R}^n.$$

This is an $(k+1)$-dimensional affine subspace. When σ is known the geometry of the model is simply that of Euclidean geometry.

In this model the natural parameters are given by β and the expectation parameters by

$$\frac{\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2}.$$

Thus there is a fixed linear relationship between the $+1$-affine and -1-affine parameters. Equation (16) implies that the Christoffel symbols for the ∇^{+1} and ∇^{-1} connections will be simultaneously zero in both the natural and expectation parameterisations. Equation (27) then implies that estimating $\boldsymbol{\beta}$ using maximum likelihood estimation will have zero bias in both those parameterisations. However in any nonlinear reparametrisation of the model the Christoffel symbols Γ^{-1} need not be zero. A simple example of this is given by a simultaneous equation model, when we parameterise by the parameters in the structural equations.

In general choosing a parameterisation can improve statistical properties. This was recognised by Hougaard (1982), and similar results are found in Amari (1990). In Section 6.1 we have seen that the Riemann curvature tensor can characterise the existence of a set of affine parameters. The following theorem formalises this idea.

Theorem For a manifold M with metric g and connection ∇ we have the following:

(a) Assume that the Riemann curvature tensor vanishes identically for all points in a sufficiently small open neighbourhood U. Then there exists an *affine* parameterisation θ covering U such that the Christoffel symbols for ∇ vanish identically. Further all ∇-geodesics in U are characterised by being affine functions of the θ-parameters.

(b) In the general case the Riemann curvature tensor will be non zero. For any point $p_0 \in M$ there will be a parameterisation θ around p_0 such that the Christoffel symbols for ∇ vanishe at p. Further the set of geodesics *through* p_0, in U, are all affine functions of the θ-parameters.

The parameterisation in case (b) is often called the *geodesic normal* parameterisation. For details of this and the related *exponential* map see Dodson and Poston (1991). When a manifold has an affine parameterisation the geometry becomes that of an affine space. All geodesics are the one dimensional affine subspaces. This is the simplest possible geometry. When the Riemann curvature is non zero the fundamental geometry of the manifold creates an obstruction to the existence of such an affine parameterisation. In this case the geodesic normal parameterisation is, in a sense, closest to being truly affine.

In the case where the connections come from the expected geometric statistical manifold then affine parametrisations, when they exist, and geodesic normal parameters, when they don't, have useful statistical properties. The following list summarises the properties of some of the more important of these. Each of these are global properties if the parameterisation is affine, otherwise they hold locally in the neighbourhood of a fixed point.

- +1-connection: The affine parameters for these are the natural parameters for the full exponential family.

- 1/3-connection: This is the quadratic log likelihood parametrisation. In this parametrisation the expectation of the third derivative of the log likelihood

$$E_{p(x \mid \theta)} \left(\partial_i \partial_j \partial_k \ell \right)$$

  vanishes. Thus the likelihood closely approximates that from a Normal family. This is the expected form of the *directed likelihood* parameterisation which has been shown to be a very effective tool in classical and Bayesian analysis, see Sweeting (1996) and references therein.

- 0-connection: This is the covariance stabilising parameterisation in which the Fisher information will be (locally) constant.

- $-1/3$-connection: This is called the *zero asymptotic skewness* parameterisation, because any first order efficient estimator will have zero asymptotic skewness in this parameterisation.

- -1-connection: We have already seen that in this parameterisation we have asymptotically minimum bias.

7.2 Intrinsic curvature

The distinction raised by Bates and Watts between parameter effects and intrinsic curvature is an important one in geometry. In this section we look at how intrinsic curvature is particularly important when undertaking inference on curved exponential families.

Suppose that N is the manifold representing the (r,t)-curved exponential family $p(x \mid \xi)$. N is embedded in the r-dimensional full exponential family, $p(x \mid \theta)$ which is denoted by M. In terms of manifolds N is an embedded sub-manifold of M, see Section 2.4.1. In the natural parameters M has many of the properties of an affine space, and intuitively it is useful to think of N as a curved subspace of an affine space. It is this curvature which will define the intrinsic curvature of the model.

We first define a set of parameters which is suited to describe the embedding of N in M.

Theorem For each $\xi_0 \in N$ there exists an open neighbourhood U of M, such that $p(x \mid \xi_0) \in U$ and U can be parameterised by

$$\begin{aligned} \Xi \times X (\subset \mathbf{R}^t \times \mathbf{R}^{r-t}) &\rightarrow U \subset M \\ (\xi, \chi) &\mapsto p(x \mid \theta(\xi, \chi)) \end{aligned}$$

where

$$U \cap N = \{ p(x \mid \theta(\xi, 0)) \mid \xi \in \Xi \} .$$

The parameters for the full family are split such that the first t components refer to the sub-manifold, while the remaining $(r-t)$ components fill out the rest of the space, in such a way that they they are zero on the sub-manifold. For notational convenience we will denote the first t components of such a system by indices $i, j, k, \ldots$, while the remaining $(r-t)$ indices will be denoted by $a, b, c, \ldots$.

Hence if we consider the tangent space to M at a point in N, denoted by p_0. In this space there will be directions which also lie in tangent space to N. In the above parametrisation these are spanned by the set $\{ \partial_i = \frac{\partial}{\partial \xi^i} \mid i = 1, \ldots, t \}$. There will remain those directions which

are not tangent to N. The set of tangent vectors $\{\partial_a = \frac{\partial}{\partial \chi^a} \mid a = 1, \ldots, (r-t)\}$ all have this property. We think of TN_{p_0} as being a subspace of TM_{p_0}.

In fact a parametrisation (ξ, χ) can be refined to have the property that

$$\langle \partial_i, \partial_a \rangle_M = \langle \frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \chi^a} \rangle_M = 0$$

for all $i = 1, \ldots, t$ and for all $a \in 1, \ldots, (r-t)$. Hence the splitting of the tangent space TM_{p_0} can be made into TN_{p_0} and its orthogonal component. It will be convenient to use such a parametrisation throughout this section.

Suppose now that there is a metric and a connection on M. These will induce a metric and a connection on the sub-manifold N in the following way. Let g_{ij} be the components of the metric on M relative to the (ξ, χ)-parameterisation. By definition the parameterisation has the property that

$$g_{ia} = 0$$

for all $i = 1, \ldots, t$ and for all $a \in 1, \ldots, (r-t)$. Relative to the ξ-parameterisation define the metric on N, $\tilde{g}$ to have components

$$\tilde{g}_{ij} = \langle \partial_i, \partial_j \rangle_M \tag{28}$$

Let X, Y be tangent fields on N. They can then also be thought of, slightly loosely, as tangent fields on M. Hence we can define the rate of change of Y relative to X using the connection on M. This gives $\nabla_X Y$ a tangent field in M. However a connection on N gives this rate of change as a tangent field on the sub-manifold. Hence we need to project $\nabla_X Y$ from TM to TN. We do this using the metric on M and an orthogonal projection. The connection $\tilde{\nabla}$ on N is defined by

$$\tilde{\nabla}_X Y = \Pi_N \left(\nabla_X Y \right) \tag{29}$$

where Π_N is the orthogonal projection of TM to TN.

The embedding curvature of N in M is defined as the difference between the result of using the connection relative to M and relative to N. Formally it is a 2-tensor H defined by

$$H(X, Y) = \nabla_X Y - \tilde{\nabla}_X Y. \tag{30}$$

$H(X, Y)$ will lie in the orthogonal complement to TN. It can be written with respect to the parameterisation defined above as

$$H(\partial_i, \partial_j) = \tilde{\Gamma}^a_{ij} \partial_a = H^a_{ij} \partial_a. \tag{31}$$

7.2.1 Auxiliary spaces

Suppose that we now wish to conduct inference in a curved exponential family N. An estimator is a function which maps the data x to a point in N. It is convenient to use the fact that in the full exponential family M, which embeds N, there is a bijection between x and $\hat{\theta}(x)$, the maximum likelihood estimate in M. We can therefore think of an estimator, geometrically, as a function

$$\begin{aligned} T : M &\rightarrow N \\ \hat{\theta}(x) &\mapsto T(\hat{\theta}(x)). \end{aligned}$$

We then study the properties of the estimator in terms of the geometry of this function. T is said to be *regular* if it is a smooth map with derivative of full rank and the restriction of T to the subspace, $T|_N$, is the identity.

Define the *auxiliary space* of the estimator to be the points in M given by the set of parameters

$$A(\xi) = \{\theta \mid T(\theta) = \xi\}. \tag{32}$$

We shall assume that for each point $\xi_0 \in N$ there exists an open subset U of M such that $A(\xi_0) \cap M$ is an $(r-t)$-dimensional sub-manifold of M. This assumption will be satisfied by all of the commonest estimators used in econometrics and in practice it will not be restrictive. We shall use the geometry of these sub-manifolds together with that of N to explore the behaviour of estimators.

A simple way of measuring the efficiency of an estimator is in terms of mean square error. We define an estimator to be k^{th} order efficient if the asymptotic expansion of its mean square error is minimal among all $(k-1)^{\text{th}}$ efficient estimators. The following theorems, due to Amari characterise this asymptotic efficiency.

Theorem A consistent estimator is *first order efficient* if the associated auxiliary space $A(\xi)$ cuts N orthogonally with respect to the expected Fisher information metric.

Theorem The first order bias of a consistent estimator is defined to be

$$b^a(\xi) = \frac{-1}{2n} g^{ij} \left\{ g^{kl} \Gamma^{-1}_{ijk} + g^{ab} h^{-1}_{abj} \right\}, \tag{33}$$

where h is the embedding curvature of the sub-manifold $A(\xi)$ with respect to the -1-connection and is defined as

$$h^{-1}_{abj} = \langle H^{-1}(\partial_a, \partial_b), \partial_j \rangle.$$

Thus for a curved exponential family the bias term comes in two parts. The first, as we have seen before, is parameterisation dependent, however the second depends only on the intrinsic geometry of the auxiliary space. This depends on the exact form of the estimator but *not* on the parameterisation use. Thus we have a decomposition into parameter effects and intrinsic curvature.

Using this geometric characterisation the third order most efficient estimator for a curved exponential family can be found.

Theorem The biased correct maximum likelihood estimator is third order efficient for a curved exponential family.

All the results shown in this section are proved in Amari (1990), for a good treatment see Kass and Vos (1997, pp227).

8 Curvature and information loss

Once we move from the full exponential family case we lose the property that the maximum likelihood estimate (MLE) will be a sufficient statistic. Fisher (1925) was the first to suggest using the observed information as the best means of recovering information lost by reducing the data to just the maximum likelihood estimate. There is an elegant geometric extension to higher order recovery which provides a complete decomposition of the information in a minimal sufficient statistic. The geometry provides a natural framework for expressing the higher order calculations but also an explanation of why these terms successfully recover additional information. Having analysed the decomposition of the information in a statistic in geometric terms there remains the question of how to use this information for inference.

8.1 Information loss

Since the MLE is a sufficient statistic for a full exponential family standard inferential procedures will automatically make use of all available information in the data. We look at geometrical properties which make a family exponential, and hence ensures that the full information content in the data is exploited.

8.1.1 One dimensional case

First consider a one dimensional family as in Efron (1975). If we have an r-dimensional full exponential family defined by

$$p(x \mid \theta) = \exp\{\theta^i s_i(x) - \psi(\theta)\} m(x)$$

then any one dimensional affine map

$$\xi \mapsto \theta^i(\xi) = \alpha^i \xi + \beta^i$$

will define a one-dimensional full exponential family. In general we have the following result.

Theorem A one dimensional curved exponential family

$$p(x \mid \xi) = \exp\{\theta(\xi) s_i(x) - \psi(\theta(\xi))\} m(x) \tag{34}$$

is a full exponential family if and only if there exist constants $(\alpha^1, \ldots \alpha^r)$ and $(\beta^1, \ldots, \beta^r)$ such that

$$\theta(\xi) = \alpha^i f(\xi) + \beta^i$$

for some smooth monotone function f.

Hence we see that being a full exponential family depends on the family being defined as an affine subspace of the natural parameters of the embedding family. In Section 5.2.1 we saw that the ∇^{+1} connection is the appropriate tool for measuring curvature relative to the natural parameters. Efron (1975) approached this question more directly. He defined the *statistical curvature* of Equation (34) to be

$$\gamma = \gamma(\theta) = \frac{\langle \theta''(\xi)_N, \theta''(\xi)_N \rangle^{1/2}}{\langle \theta'(\xi), \theta'(\xi) \rangle} \tag{35}$$

where $\theta'(\xi) = \frac{d\theta}{d\xi}$ and $\theta''(\xi)_N$ denotes the component of the second derivative which is normal to the tangent direction $\theta'(\xi)$. The inner product is the expected Fisher information metric. The motivation for this comes directly from the definition of the curvature of a one dimension path in $\mathbf{R}^n$. The important result concerning statistical curvature is given by the following theorem.

Theorem The curved exponential family given by Equation (34) is a one parameter full exponential family if and only if its statistical curvature is zero.

One important property of the full exponential family is that the observed information equals the expected Fisher information. They both equal $\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}(\hat{\theta})$ in the natural parameterisation. The randomness of the observed information is purely a function of the MLE. In fact Murray and Rice (1993) characterise a full exponential family by the property that the second derivative of the log-likelihood lies in the span of the first derivatives. In a curved family the expected and observed information matrices will differ. The statistical curvature, defined above gives a useful measure of the amount of variability in the observed information, given the MLE. This comes from the following result, see Efron and Hinkley (1978).

Theorem In a regular curved exponential family if $\mathcal{I}(\hat{\xi})$ is the observed Fisher information and $I(\hat{\xi})$ the expected information for one observation, then

$$\frac{(\mathcal{I}(\hat{\xi}) - nI(\hat{\xi}))}{\sqrt{n}I(\hat{\xi})\gamma(\hat{\theta})} \to N(0,1) \tag{36}$$

where the convergence is in law as $n \to \infty$.

8.1.2 The general case

Suppose now that we are working in a (r,t)-curved exponential family. Again we wish to understand the amount of information lost if we only use the statistic $\hat{\xi}$. To do this we need to define the information contained in a general statistic $T(x)$. Using an asymptotic measure of information, the information in a statistic T is defined as being the expected Fisher information for the likelihood for ξ based on the statistic $T(x)$. That is

$$I^T(\xi) = E_{p(t \mid \xi)}\left(-\frac{\partial^2 \ell}{\partial \xi^i \partial \xi^j}(t\,;\,\xi)\right). \tag{37}$$

This measures, by the Cramer-Rao theorem, the variance of the best possible estimator based only on the statistic $T(x)$. Of course with this notation we would have $I^X(\xi) = I(\xi)$ the standard Fisher information. We can generalise the previous results on information loss to multidimensional families.

Theorem In a curved exponential family $p(x \mid \xi)$ then

$$nI_{ij} - I^{\hat{\xi}}_{ij} = g^{kl}\langle H^{+1}(\partial_i, \partial_k), H^{+1}(\partial_j, \partial_l)\rangle. \tag{38}$$

Where H^{+1} is the embedding curvature of the family relative to the +1-connection, and the inner product is relative to the expected Fisher metric.

For a proof of this result see Kass and Vos (1997, page 222). Thus the +1-embedding curvature plays the role of Efron's statistical curvature in higher dimensions.

8.2 Information recovery

Having seen that the geometric structure of a family allows the measurement of loss in information, we now ask in what way can this information be recovered and used in inference? Since the MLE, $\hat{\xi}$ will not in general be sufficient we need to add a further statistic which will recover (approximate) sufficiency. It is convenient to construct the *conditional resolution*. This is a decomposition of the sufficient statistic into $(\hat{\xi}, a)$, where a is (approximately) ancillary, combined with an (approximate) expression for $p(\hat{\xi} \mid \theta, a)$. For more details of this construction see Barndorff-Nielsen and Cox (1994).

8.2.1 Choosing ancillaries

One difficulty with using a conditional resolution structure is the problem of the non-uniqueness of the ancillary statistic, raised by Basu (1964). Solutions to this problem were considered by Barnard and Sprott (1971) and Cox (1971). We look at a geometrical construction defined in Kass and Vos (1997, page 222), which gives a way of ensuring that the ancillary constructed captures information to progressively higher asymptotic order.

First we use geometry to construct approximately sufficient statistics for a curved exponential family. These are then transformed into approximately ancillary statistics. Consider

the (r,t)-curved exponential family N given by $\exp\{\theta^i(\xi)s_i(x) - \psi(\theta(\xi)\}m(x)$. Clearly the statistic $(s_1(x),\ldots,s_r(x))$ is sufficient, but this can be very high dimensional. In fact as Examples 3, 4 and 5 show, in econometric examples this sufficient statistic can be of order n. It is therefore natural to see if we can reduce the dimension of the sufficient statistic. Using the definition of tangent vector and the affine embedding given in Section 3.2.1, the best local affine approximation to N at ξ_0 will be given by the t-dimensional full exponential family, $M_{[1]}$,

$$\begin{aligned} p(x \mid (\alpha^1,\ldots,\alpha^t)) &= \exp\left\{\left(\theta^i(\xi_0) + \alpha^j\partial_j\theta^i(\xi_0)\right)s_i(x) - \psi\left(\theta(\xi_0) + \alpha^j\partial_j\theta(\xi_0)\right)\right\}m(x) \\ &= \exp\{\alpha^j(\partial_j\theta^i(\xi_0)s_i(x)) - \psi(\alpha)\}m_1(x) \\ &= \exp\{\alpha^j\tilde{s}_j - \psi(\alpha)\}m_1(x) \end{aligned}$$

say, for a measure $m_1(x)$. The natural parameters are given by $(\alpha^1,\ldots,\alpha^t)'$. It follows from the theorems in Section 8.1.1 that this will be a full exponential family. Its sufficient statistic will equivalent to $\hat{\xi}$, the MLE for the curved family. This is defined by the equations

$$\begin{aligned} \frac{\partial\theta^i}{\partial\xi^j}(\hat{\xi}^i)\left(s_i - \frac{\partial\psi}{\partial\theta^i}\right) &= \tilde{s}_i - \frac{\partial\theta^i}{\partial\xi^j}(\hat{\xi}^i)\frac{\partial\psi}{\partial\theta^i}(\hat{\xi}) \\ &= 0. \end{aligned}$$

Due to the regularity conditions of Section 2.3 this equation will be invertible. Hence the information in $\hat{\xi}$ is precisely equivalent to $(\tilde{s}_1,\ldots,\tilde{s}_t)'$.

In general we can define a sequence of full exponential families $M_{[1]}, M_{[2]},\ldots,M_{[k]}$ where $M_{[k]}$ is the family

$$\begin{aligned} p(x \mid \alpha_i,\alpha_{ij},\ldots,\alpha_{i_1\ldots i_k}) = \exp\{(\theta^i(\xi_0) &+ \alpha^j\partial_j\theta^i(\xi_0) + \alpha^{j_1j_2}\partial^2_{j_1j_2}\theta^i(\xi_0) + \\ &+ \ldots + \alpha^{j_1\cdots j_k}\partial^k_{j_1\cdots j_k}\theta^i(\xi_0))s_i(x) - \psi(\theta(\alpha))\}m(x) \end{aligned}$$

The sufficient statistic for this full exponential family is given by

$$\left(\hat{\xi}, \partial_{i_1i_2}\ell(\hat{\xi}),\ldots,\partial^k_{i_1,\ldots,i_k}\ell(\hat{\xi})\right).$$

We therefore have constructed a sequence of full exponential families, which give progressively better approximations to the curved family N. These in turn give a sequence of statistics, which become progressively better approximations to the sufficient statistic for N.

It is necessary to check that this construction is geometrically well defined, so that the sufficient statistics generated will be independent of any choice of parameterisation. This follows since the linear spaces given by

$$T_{[k]} = span\{\partial_i\theta, \partial^2_{i_1i_2}\theta,\ldots,\partial^k_{i_1\cdots i_k}\theta\}$$

will be parameterisation independent, see Murray and Rice (1993).

The sequence of sufficient statistics $\left(\hat{\xi}, \partial_{i_1i_2}\ell(\hat{\xi}),\ldots,\partial^k_{i_1,\ldots,i_k}\ell(\hat{\xi})\right)$ can then be transformed to be approximately in the conditional resolution form $(\hat{\xi}, a)$. Define

$$h_{i_1\cdots,i_{k+1}} = P^{\perp}_{[k]}\partial^{k+1}_{i_1\cdots i_{k+1}}$$

where $P^{\perp}_{[k]}$ is the orthogonal projection into $T_{[k]}$. This orthogonalisation ensures that the terms are uncorrelated and asymptotically independent, further the expected Fisher information based on $h_{i_1\cdots,i_{k+1}}$ will be of asymptotic order n^{-k+1}. To achieve approximate ancillarity these statistics are adjusted to give zero mean and unit variance to the correct asymptotic order. For further details see Kass and Vos(1997). Note that the terms $h_{i_1\cdots i_k}$ are simply generalisations of the embedding curvature in Section 7.2.

8.2.2 The p^*-formula

Having used geometry in the construction of an approximately sufficient statistic of the from $(\hat{\xi}, a)$, the second part of a conditional resolution is to approximate the distribution $p(\hat{\xi} \,|\xi, a)$. Barndorff-Nielsen, in a series of papers, proposes that a very good higher order approximation, based on the saddlepoint method given by the so called p^*-formula. For a derivation of results in this section see Barndorff-Nielsen and Cox (1994, page 238). The p^*-approximation is given by

$$p^*(\hat{\xi} \mid \xi, a) = c|\hat{j}|^{1/2} e^{\bar{\ell}}. \tag{39}$$

where

$$\hat{j} = \frac{\partial^2 \ell}{\partial \xi^i \partial \xi^j}(\hat{\xi})$$

and

$$\bar{\ell}(\xi \mid \hat{\xi}, a) = \ell(\xi \mid \hat{\xi}, a) - \ell(\hat{\xi} \mid \hat{\xi}, a).$$

The constant c is defined to make the above density integrate to one. When this is not known the approximation

$$p^\dagger(\hat{\xi} \,|\theta, a) = (2\pi)^{-t/2} |\hat{j}|^{1/2} e^{\bar{\ell}}, \tag{40}$$

can be used. This version is accurate to order $O(n^{-1})$, whereas the p^*-formula is accurate to order $O(n^{-3/2})$. For a further discussion of this issue of the distribution of the Maximum Likelihood Estimator see the paper by Grant Hillier and Ray O'Brien later in this volume.

References

Amari S-I, (1987). Differential geometric theory of statistics. In *Differential Geometry in Statistical Inference* (eds. S.I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen and C.R. Rao), 217-240. Institute of Mathematical Statistics, Hayward, CA.

Amari S-I, (1990), *Differential-Geometric Methods in Statistics.* Springer, Lecture Notes in Statistics, **28**, Berlin.

Barndorff-Nielsen, O.E., (1978) *Information and Exponential Families in Statistical Theory.* Wiley, London.

Barndorff-Nielsen, O.E., (1987), Differential geometry and statistics: some mathematical aspects. *Indian J. Math.* **29**, Ramanujan Centenary Volume.

Barndorff-Nielsen, O.E., (1988) *Parametric Statistical Families and Likelihood.* Springer, New York.

Bandorff-Nielsen and Cox (1994),*Inference and Asymptotics* Monographs on Statistics and Applied Probability, **52**, Chapman and Hall, London

Barndorff-Nielsen, O.E. and Jupp, P., (1989), Approximating exponential models, *Ann. Statist. Math.* **41**, 247-267.

Barnard, G.A. and Sprott, D.A., (1971), A note on Basu's examples of anomalous ancillary statistics (with discussion). In *Foundations of Statistical Inference* eds V.P. Godambe abd D.A. Sprott). Holt, Rinehart and Winston, Toronto, pp. 163-76

Basu, D., (1964), Recovery of ancillary information, *Sankya A*, **26**, 3-16.

Bates, M. and Watts, D.G., (1988) *Nonlinear Regression Analysis & its Applications* Wiley, London.

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

Blæsild, P., (1991). Yokes and tensors derived from yokes. *Ann. Inst. Statist. Maths.* **43**, 95-113

Bröcker, T.H. and Jänich, K., (1982), *Introduction to Differential Geometry*, CUP, Cambridge.

Brown, L.D., (1986), *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Haywood, CA.

Cox, D.R., (1971), The choice between alternative ancillary statistics. *J. R. Statist. Soc. C,* **33**, 251-5.

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

Critchley, F., Marriott P.K., and Salmon, M., (1993), Preferred point geometry and statistical manifolds. *Ann. Statist.* **21**, 1197-1224.

Critchley, F., Marriott P.K., and Salmon, M. (1994) On the local differential geometry of the Kullback-Liebler divergence,*Annals Statist* **22** p1587-1602.

Critchley, F., Marriott P.K., and Salmon, M., (1996), On the differential geometry of the Wald test with nonlinear restrictions. *Econometrica* **64**, 1213-1222.

Critchley, F., Marriott P.K., and Salmon, M., (1998), An Elementary account of Amari's Expected geometry, *Applications of Differential Geometry in Econometrics* Ed. P.K. Marriott and M. Salmon, CUP, Cambridge

Dodson, C.T. and Poston, T., (1991), *Tensor Geometry: The Geometric Viewpoint and Its Uses*, 2nd ed. Springer-Verlag, New York.

Efron, B., (1975), Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3**, 1189-1217.

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

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

Efron, B. and Hinkley, D.V., (1978), Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information, *Biometrika* **65**, 457-481.

Fisher, R.A., (1925), Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society*, **22**, 700-725.

Hougaard, P., (1982), Parametrisations of nonlinear models. *J. Roy Statist. Soc. Series B Methodological* **44**, 244-252.

Kass R.E. and Vos P.W., (1997), *Geometrical Foundations od Asymptotic Inference*, Wiley series in Probability and Statistics, J. Wiley and Sons, New York.

Lauritzen, S.L. (1987), Statistical manifolds. In *Differential Geometry in Statistical Inference* (eds. S.I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen and C.R. Rao), 217-240. Institute of Mathematical Statistics, Hayward, CA.

Marriott, P.K., (1989), *Applications of differential geometry to statistics* Ph.D. dissertation. University of Warwick.

McCullagh, P., (1987), *Tensor Methods in Statistics*, Monographs on Statistics and Applied Probability, **29**, Chapman and Hall, London.

Murray, M.K. and Rice, J.W., (1993), *Differential Geometry and Statistics*, Monographs on Statistics and Applied Probability, **48**, Chapman and Hall, London.

Ravishanker, N (1994), Relative Curvature Measures of Nonlinearity for Time-Series Models, *Communications in Statistics–Simulation and Computation*, **23**, No.2, pp.415-430

Rao, C.R., (1945), Asymptotic efficiency and limiting information. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 531-545.

Rao, C.R., (1987), Differential metrics in probability spaces. In *Differential Geometry in Statistical Inference* (eds. S.I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen and C.R. Rao), 217-240. Institute of Mathematical Statistics, Hayward, CA.

Rudin, W., (1976), *Principles of Mathematical Analysis*, 3rd edition, McGraw Hill, New York.

Spivak, M. (1979), *A Comprehensive Introduction to Differential Geometry, 2nd Edition*, Publish and Perish, Boston.

Sweeting, T., (1996), *Bayesian Stats* Approximate Bayesian computation based on signed roots of log-density ratios (with discussion). *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.), Oxford: University Press, 427-444.

Willmore, T.J., (1959), *An Introduction to Differential Geometry*, Clarendon, Oxford.