Applications of differential geometry to econometrics

Edited by

PAUL MARRIOTT AND MARK SALMON

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge, CB2 2RU, UK www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

First published 2000

Printed in the United Kingdom at the University Press, Cambridge

Typeface 10/12pt Times *System* Advent 3B2 [KW]

A catalogue record for this book is available from the British Library

ISBN 0 521 65116 6 hardback

Contents

Figures and tables

Figures

Tables

1 An introduction to differential geometry in econometrics

Paul Marriott and Mark Salmon

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 either of differential geometric theory, or of all the applications that 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 that 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 to 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. Section 6 introduces key mathematical theorems involving statistical manifolds, duality, projection and finally the statistical application of the classic geometric theorem of Pythagoras. 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, although 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 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 that defines a *manifold* and shows how one manifold can be defined as a curved subfamily 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: 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 an 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, p. 4).

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.* in chapter 10 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

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

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

$$\theta' = \left(\frac{\beta'}{\sigma^2}, \ \frac{1}{-2\sigma^2}\right)$$

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

$$\mathbf{B}\mathbf{Y}_\mathbf{t} + \boldsymbol{\Gamma}\mathbf{X}_\mathbf{t} = \mathbf{U}_\mathbf{t},$$

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

$$\mathbf{Y}_\mathbf{t} = -\mathbf{B}^{-1}\boldsymbol{\Gamma}\mathbf{X}_\mathbf{t} + \mathbf{B}^{-1}\mathbf{U}_\mathbf{t},$$

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 reparameterisations 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,

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

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 Parameterisations

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 that are dependent on a particular choice of parameterisation and those that are independent of this choice. Indeed, one can define the geometry of a space to be those properties that 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

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

From equation (2) it follows that these parameters can be expressed as

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

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), p. 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

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

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

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

which obeys the following conditions:

1. the dimension of Ξ is less than that of Θ,
2. the mapping is one-to-one and 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

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

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

$$(\alpha, \beta) \to \big(\theta^1(\alpha, \beta), \ldots, \theta^n(\alpha, \beta)\big),$$

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

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

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

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

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 non-linear restriction will also be a curved exponential family. Its curvature is defined by a non-linear restriction in a family that 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 by 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 that lie outside this class and how this affects the geometric theory. We have space only to highlight the issues here but it is clear that families that 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 of these families is that the dimension of their sufficient statistics grows with the sample size. Although this does not make an exact geometrical theory impossible, it does considerably complicate matters.

Another property that 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, although 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, p. 16). In detail these conditions for a parametric family $p(x \mid \theta)$ 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

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

   are linearly independent,
3. moments of $\partial\ell/\partial\theta^i(\theta\,; x)$ 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 multi-dimensional 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 fully to 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 three-dimensional spaces regarding the geometry of the econometric models we want to consider.

The most familiar geometry is three-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

$$X \to X$$

$$x \mapsto x + v$$

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 two-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

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

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 that 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), p. 71).

In the sphere example it is Condition (A) that makes us exclude the poles and Condition (B) that 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 applied it only 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 such that these local subsets can be combined in a smooth way. A number of technical issues 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 as 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 that 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 $V = \{f(x) \mid f \in C^\infty(S, \mathbf{R})\}$.

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 that 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 a 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 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 subfamily of the n-dimensional multivariate normal model, where n is the sample size. This is the model that determines the innovation process. Thus it is a submodel of an n-dimensional full exponential family. In fact it lies in a three-dimensional subfamily of this full exponential family. This is the smallest full family that 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 that 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

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

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 and 1982) in a series of seminal papers that revived interest in the geometric aspects of statistical theory. This work was followed by a series of papers 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, as well as 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) and 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 that 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. 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 regularity condition 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 because 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), $\hat{\theta}$,

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

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-generation 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 matrix:

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

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 a translation that 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 study only manifolds that 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 this space to 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

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

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 that approximates the curve to first order (see Willmore (1959), p. 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 that 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 Willmore (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 that 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 an 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

$$\left\{\partial_i,\ i = 1, \ldots, r\right\};$$

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 set of the 1-forms defines the coordinate grid on the manifold given by the parameterisation θ.

3.2.4 The change of parameterisation 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