Differential Geometry and Statistics

Differential Geometry and Statistics

Michael K. Murray

Department of Pure Mathematics
The University of Adelaide
Adelaide
Australia

and

John W. Rice

School of Information Science and Technology
The Flinders University of South Australia
Bedford Park
Australia

CHAPMAN & HALL

London · Glasgow · New York · Tokyo · Melbourne · Madras

Published by Chapman & Hall, 2–6 Boundary Row, London SE1 8HN

Chapman & Hall, 2–6 Boundary Row, London SE1 8HN, UK

Blackie Academic & Professional, Wester Cleddens Road, Bishopbriggs, Glasgow G64 2NZ, UK

Chapman & Hall Inc., 29 West 35th Street, New York NY10001, USA

Chapman & Hall Japan, Thomson Publishing Japan, Hirakawacho Nemoto Building, 6F, 1-7-11 Hirakawa-cho, Chiyoda-ku, Tokyo 102, Japan

Chapman & Hall Australia, Thomas Nelson Australia, 102 Dodds Street, South Melbourne, Victoria 3205, Australia

Chapman & Hall India, R. Seshadri, 32 Second Main Road, CIT East, Madras 600 035, India

First edition 1993

Printed in Great Britain

ISBN 0 412 39860 5

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

Library of Congress Cataloging-in-Publication data available

∞ Printed on permanent acid-free text paper, manufactured in accordance with the proposed ANSI/NISO Z 39.48-199X and ANSI Z 39.48-1984

Preface

Several years ago our statistical friends and relations introduced us to the work of Amari and Barndorff-Nielsen on applications of differential geometry to statistics. This book has arisen because we believe that there is a deep relationship between statistics and differential geometry and moreoever that this relationship uses parts of differential geometry, particularly its 'higher-order' aspects not readily accessible to a statistical audience from the existing literature. It is, in part, a long reply to the frequent requests we have had for references on differential geometry! While we have not gone beyond the path-breaking work of Amari and Barndorff-Nielsen in the realm of applications, our book gives some new explanations of their ideas from a first principles point of view as far as geometry is concerned. In particular it seeks to explain why geometry should enter into parametric statistics, and how the theory of asymptotic expansions involves a form of higher-order differential geometry.

The first chapter of the book explores exponential families as flat geometries. Indeed the whole notion of using log-likelihoods amounts to exploiting a particular form of flat space known as an affine geometry, in which straight lines and planes make sense, but lengths and angles are absent. We use these geometric ideas to introduce the notion of the second fundamental form of a family whose vanishing characterises precisely the exponential families.

The second chapter, in which we introduce manifolds, should be most useful to statisticians who want to learn about the subject. The traditional theory starts with a heavy meal of the purest mathematics, (topological spaces, co-ordinate coverings, differentiable functions), before embarking on a treatment of calculus that is filled with multilinear algebra, and bears little relationship to anything one might have learned about several-variable calculus as an undergraduate. By contrast our treatment starts with calculus

on manifolds as a geometrical approach to the theory of rates of change of functions, treating it as though it were a first course on several variable calculus. We explain how the several-variable chain rule can be interpreted as dividing variations through a point into families with different velocities, how df is to be interpreted as the rate of change of f as a function of velocity, and what are vector fields (contravariant 1-tensors) and 1-forms (covariant 1-tensors). We give a brief discussion of the foundational concepts of differentiability and manifolds at the end of the chapter, but these are not really important for the application of differential geometry to statistics.

Our comment on the great divide between the so-called co-ordinate-free and index-laden approaches to differential geometry, is that we aim to be geometrical without being obsessed with freedom from co-ordinates. We have enormous interest in co-ordinates when it comes to calculations. However, it seems pointless to us to be in the position either to be able to calculate everything but explain nothing, or to explain everything but calculate nothing. So we explain geometrical concepts in co-ordinate-free terms, and we translate them into co-ordinate systems for calculations, with whatever debauches of indices they require.

Once the basic notions are in place, most notably the definition in Chapter 2 of the tangent space to a manifold, we begin an elaboration of the parts of differential geometry that are useful in statistics, illustrating them with statistical applications and examples. As the number of statistical applications is growing rapidly we have been unable to consider them all. However we believe that we have covered all the *concepts* from differential geometry that are needed at this point in time. Chapter 3 explains the idea of submanifold and the definition of a statistical manifold. We mention again the simplest statistical manifolds, the exponential families, and then consider the families with a high degree of symmetry, the transformation models.

The next two chapters introduce the concept of connections and their curvature, Amari's α-connections and the theory of statistical divergences. A connection defines the rate of change of vector fields. It therefore tells us which curves have constant tangent vector fields, that is which curves are straight lines or geodesics. Hence a connection defines a notion of geometry, or straight lines and the different connections define different geometries. Some

connections are essentially 'flat'. That is, the geometry they define is Euclidean. The curvature of a connection is a measure of its departure from flatness.

In Chapter 6 we consider the theory of Riemannian manifolds. An initial impetus for introducing differential geometry into statistics was the observation of Rao that the Fisher information could be interpreted as a Riemannian metric on the space of parametrised probability distributions forming the statistical model.

Chapter 7 introduces the maximum likelihood estimator and considers some results in asymptotics, in particular the work of Amari. Here we begin to see the importance of Taylor series and the need for a higher-order geometry in statistics. The final Chapters 8 and 9 consider this higher order geometry: the theory of strings or phyla developed by Barndorff-Neilsen and Blæsild. Strings are generalisations of tensors. If we think of tensors in co-ordinates as functions with many indices transforming under a change of co-ordinates by the first derivative of the co-ordinate transformation, then a string has more indices and transforms by higher derivatives of the co-ordinate transformation. To consider strings from a co-ordinate-free point of view requires that we introduce in Chapter 8 the theory of principal and vector bundles, in particular the so-called infinite frame bundle and the infinite phylon group. Chapter 9 then applies this theory to Taylor expansions and co-ordinate strings and relates the theory of strings to the representation theory of the infinite phylon group.

A book is not just the result of the labours of its authors but also of the generosity of others. First and foremost we thank our families who had to live through this book's production; then our many statistical colleagues who have laboured to explain their subject to us. Our thanks and apologies for the places where despite your efforts we get it wrong. Particular thanks must go to Peter McCullagh for providing us with TEX macros for this book and to Peter Jupp for his amazingly thorough reading of our first manuscript. Of course any remaining errors and omissions are our responsibility.

Michael K. Murray
John W. Rice

Contents

CHAPTER 1

The geometry of exponential families

1.1 Geometry, parameters and co-ordinates

Parametric statistics concerns parametrised families of probability distributions $p(\theta)$, where the parameter $\theta = (\theta^1, \ldots, \theta^d)$ varies over some open set in $\mathbf{R}^d$. The most common example is the normal family, which is usually expressed as a family of densities

$$p(\mu, \sigma) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

The parameter θ in this case is the pair (μ, σ) which varies over the open subset of $\mathbf{R}^2$ determined by $\mu > 0$. The sample space is $\mathbf{R}$ and the densities are with respect to Lebesgue measure dx on $\mathbf{R}$, so that as a set of probability measures the normal family is

$$\mathcal{N} = \{p(\mu, \sigma)dx \mid \mu \in \mathbf{R}, \sigma > 0\}$$

Statistical inference concerns situations in which one knows or suspects that data are generated by sampling from a space according to a probability distribution which is a member of some known family $p(\theta)$. The problem is to infer facts about the distribution from the data. For example, one might want to know the parameter value of the distribution (point estimation), or simply whether or not this value lies in some particular set of parameters (hypothesis testing). If a given collection of numbers has arisen by sampling from a normal distribution then one might ask which normal distribution it is, i.e. what are the values of μ and σ for this particular normal distribution. On the other hand one might only want to test the hypothesis that the mean is greater than 1.

Many of these tests and much of the theory of statistical inference depends on the choice of parameters. This dependence on parameters usually comes about because the theory applies differential calculus to these parameters; differentiating them or perhaps Taylor series expanding some function with respect to them. It is important to know how the theory depends on the parameters, either because one suspects that it should not depend on the parameters at all or because one would like to know if a particular choice of parameters may simplify matters. That part of differential geometry which we have called 'Calculus on Manifolds' in Chapter 2, is concerned exactly with this question of how the differential calculus depends on co-ordinates. It obviously has immediate application to these problems.

Differential geometry however is more that just understanding how calculus depends on co-ordinates – it is also a theory of geometry or shape. Borrowing from the ideas of differential geometry we think of families of probability distributions as entities independent of any particular parametrisation, and able to support a variety of geometries. We seek to relate their statistical properties to these geometries.

We will motivate this use of geometry in statistics by considering in this first chapter the geometric significance of the exponential family. It has long been known that in seeking answers to many statistical questions the simplest type of family to deal with is the *exponential family*, i.e. one which can be parametrised in the form

$$p(\theta) = \exp\left(\theta^1 x^1 + \ldots + \theta^r x^r - K(\theta)\right) d\mu$$

where $x^1, \ldots, x^n$ are random variables and μ a measure on some sample space. If there is to be a meaningful relationship between statistics and geometry we must be able to discover some geometric significance to a family being exponential. Indeed we can. We shall show that the geometry of an exponential family is perhaps the simplest geometry possible, that is, affine geometry. This explains the geometric significance of the so-called *canonical parameters* θ in the exponential parametrisation. They are the affine co-ordinates arising from the affine geometry.

If we regard a parametrised family of probability distributions as analogous to a surface with a co-ordinate system on it then individual probability distributions correspond to the points on the surface, and their parameter values are their co-ordinates.

After a fashion, this kind of interpretation is often made for the normal family by regarding (μ, σ) as the Cartesian co-ordinates of a point in the upper half plane in $\mathbf{R}^2$. This certainly sets up a correspondence between the probability distributions of the normal family and points on a surface, viz. the upper half plane. However, there is no good reason to think that such an *ad hoc* correspondence should be taken seriously, nor in particular that the flatness of the plane or any other of its geometric features should have any significance for the statistical properties of the normal distribution. On the other hand, as we shall see in this first chapter, if we use the exponential parametrisation to set up a correspondence between individual normal distributions and points of a plane then we should take its flatness very seriously indeed.

Ultimately we shall be considering a variety of geometries on any given family of probability distributions, which may be interpreted loosely as imposing a variety of shapes upon the 'surface' of probability distributions. For any given geometry, or shape, there may be co-ordinate systems which are closely tied to the geometry and co-ordinate systems which are not. For example, the Euclidean geometry of a plane, involving such concepts as distance and orthogonality, is well reflected by Cartesian co-ordinate systems, but not so well by polar or other kinds. The existence of an exponential parametrisation comes about because, in a certain well defined sense, the family of probability distributions is flat, and the exponential parameters or co-ordinates are the ones adapted to this flatness.

It is well known that a family can be exponential without it being immediately apparent. For example, the normal family is an exponential family since as well as its usual parametrisation it can be parametrised

$$p(\theta^1, \theta^2)(x) = \exp(x^2\theta^1 + x\theta^2 - K(\theta))$$

where

$$\theta^1 = \frac{-1}{2\sigma^2}, \quad \theta^2 = \frac{\mu}{\sigma^2} \quad \text{and} \quad K(\theta) = \frac{1}{2}\log(\frac{-\pi}{\theta^1}) - \frac{(\theta^2)^2}{4\theta^1}$$

The parameter $\theta = (\theta^1, \theta^2)$ is called the canonical parameter, and lies in the open subset of $\mathbf{R}^2$ defined by $\theta^1 < 0$.

The point is that one cannot say that a family of probability distributions is not an exponential family just because it does not

appear in exponential form. It has to be proved that the family cannot be reparametrised into exponential form. The question as to whether or not a family of probability distributions is an exponential family is therefore a question about reparametrisation of the family.

Because we are able to assign geometric meaning to a family being exponential we are able to produce an invariant of any family, its second fundamental form, whose vanishing characterises exactly the exponential families. In the case of a one-dimensional family this second fundamental form is closely related to Efron's statistical curvature Efron (1975).

1.2 Canonical co-ordinates

Let us begin to look for geometry in an exponential family

$$p(x,\theta) = \exp(\sum_{i=1}^{r} x^i\theta^i - K(\theta)) \qquad (1.2.1)$$

by considering the canonical parameters $\theta = (\theta^1,\ldots,\theta^r) \in \mathbb{R}^r$ which are obviously not arbitrary but have to be chosen so that p has this special form. We seek to relate the fact that an exponential family has this restricted set of parameters to some kind of geometry. It is important then to know how restricted this set of parameters is. That is 'how canonical are the canonical parameters ?' Is it possible to have another set of random variables $y^i(x)$, parameters $\phi^i(\theta)$, and a function $J(\theta)$ such that

$$p(x,\theta) = \exp(\sum_{i=1}^{r} y^i(x)\phi^i(\theta) - J(\theta))?$$

We should, of course, also allow the possibility that we have changed the measure relative to which these are densities, so we could have

$$p(x,\theta) = \exp(\sum_{i=1}^{r} y^i(x)\phi^i(\theta) - J(\theta) + f(x)) \qquad (1.2.2)$$

for some function f. Comparing (1.2.1) and (1.2.2) we see that we must have

$$\sum_{i=1}^{r} x^i\theta^i - K(\theta) = \sum_{i=1}^{r} y^i(x)\phi^i(\theta) - J(\theta) + f(x) \qquad (1.2.3)$$

and differentiating both sides of (1.2.3) with respect to x^i gives

$$\theta^i = \sum_{j=1}^{r} \frac{\partial y^j}{\partial x^i}\phi^j + \frac{\partial f}{\partial x^i}$$

for every $i = 1, \ldots, r$. In particular if we choose a point θ such that $\phi^i(\theta) = 0$ for all i we see that

$$\xi^i = \frac{\partial f}{\partial x^i}$$

must be a constant vector and

$$X_i^j = \frac{\partial y^j}{\partial x^i}$$

must be a constant matrix.

The two sets of canonical parameters are therefore related by

$$\theta^i = \sum_{j=1}^{r} X_j^i\phi^j + \xi^i \qquad (1.2.4)$$

The relationship in equation (1.2.4) between the canonical parameters of two exponential parametrisations is exactly the same as the relationship between Cartesian co-ordinate systems in the plane, but with the restriction in this latter case that the 2×2 matrix X_j^i must be a rotation matrix. In order to obtain general non-singular matrices X_j^i we must go beyond Cartesian systems to those determined by skewed axes with independent units of length. Such co-ordinate systems are not tied to the notions of length and angle, but they do reflect the notions of straightness and parallelism, or equivalently, as we shall explain, the notion of parallel translation in the plane.

As an example of the kind of geometry implied by this relationship between the canonical co-ordinates notice that we can use

the canical co-ordinates to define the notion of a straight line in an exponential family. We say that a subset L of an exponential family is a line if its image under some canonical co-ordinates is a line in $\mathbf{R}^r$. This is, in fact, independent of which particular canonical co-ordinates are chosen because the image of a line under an affine transformation is still a line. Similarly we can define an affine subspace of an exponential family to be a subset whose image under some (and hence all) canonical co-ordinates is an affine subspace of $\mathbf{R}^r$, that is, the translate of a vector subspace. To understand where all this geometry is coming from we have to introduce the concept of an *affine space*.

1.3 Affine spaces

An affine space can be thought of as a set which becomes a vector space by selecting a point to be the zero point. The plane is an important motivating example. It is not a vector space itself, and in particular no one point stands out as the zero element. However, having chosen an arbitrary point to play the role of an origin, and so to be the zero vector, all of the other points can be regarded as vectors. Specifically, points correspond to the tips of arrows based at the chosen origin, and they are added or multiplied by scalars according to the parallelogram rules applied to their corresponding arrows. Although it usually doesn't matter, the addition and scalar multiplication of points is completely different for different choices of origin, because of the different arrows to which points correspond.

A vector in the plane doesn't usually refer to a single arrow but rather to a whole family of arrows which are parallel translates of each other. In other words, we regard two arrows which are parallel translates of each other as instances of the same vector. Given a vector v and a point p in the plane we can consider the particular arrow based at p which corresponds to v. We call its tip $p+v$. From this point of view each vector v defines an operation on the plane sending each point p to the point $p + v$. We call this operation translation through v, and denote it by $+v$ applied to the right so that the value of $+v$ acting on p is given in the usual way as $v + p$. Notice that $(p + v) + w = p + (v + w)$, or in other words the composition of the operation $+v$ with $+w$ is the operation $+(v+w)$.

Moreover, given any two points p and q there is a unique vector v for which $q = p + v$, namely the vector corresponding to the arrow from p to q. This is just another way of saying that a choice of origin p sets up a one-to-one correspondence between points and vectors, viz. q corresponds to the vector v which translates p to q.

The same structure appears in three-dimensional space. We have three-dimensional vectors v, represented by arrows, which define translation operations $+v$ satisfying $(p+v)+w = p+(v+w)$ for any two vectors v and w and any point p. Moreover, for any two points p and q there is a unique vector v such that $q = p + v$, so that choosing p as an origin sets up a one-to-one correspondence between points and vectors. A general affine space is defined as a set X and a vector space V, each vector v of which corresponds to a transformation $+v$ from X to itself called translation by v. The translations have to satisfy the two rules described above, that is, $(p+v)+w = p+(v+w)$ for any point p and vectors v and w, and given any two points of X there must be a unique translation that moves one to the other.

Affine spaces have a fundamental geometic significance in that they are to be considered flat, like the plane and three-dimensional space. A characteristic of affine spaces is the presence of special co-ordinate systems called affine co-ordinates. As we shall show, exponential families are affine spaces, and their canonical parameters are affine co-ordinates.

In the plane a pair of linearly independent arrows based at an origin determines a co-ordinate system. In vector space terms the arrows v_1, v_2 form a basis for the space of arrows, so that every arrow v can be expressed uniquely in the form $\theta_1 v_1 + \theta_2 v_2$. The numbers (θ_1, θ_2) can be regarded as co-ordinates for the point corresponding to v. If the arrows are of unit length and at right angles, and v_1, v_2 are in anticlockwise order, then such a co-ordinate system is a Cartesian co-ordinate system. In general these kinds of co-ordinate systems are called affine co-ordinates.

For a general affine space, having chosen an origin o we choose an ordered basis $v^1, v^2, \ldots, v^r$ for the space of translations. This is like choosing a set of axes at the origin. We obtain co-ordinates for points by expanding their corresponding vectors in terms of the basis. Each vector can be expressed as

$$v = \theta^1 v^1 + \theta^2 v^2 + \ldots + \theta^r v^r$$