GEOMETRIZATION OF STATISTICAL THEORY

Proceedings of the GST Workshop
University of Lancaster Department of Mathematics
28-31 October 1987

C T J DODSON *(EDITOR)*

ULDM Publications
Department of Mathematics
University of Lancaster,
Lancaster LA1 4Yl
England.

ULDM Publications
Department of Mathematics
University of Lancaster
LANCASTER LA1 4YL ENGLAND

Telephone: (0524) 65201
Telex: 65111 Lancul G
Electronic Mail: maa017 @ uk.ac.lancs.vax2

First published 1987
AMS Subject Classifications : 53B 53C, 62A, 62E

ISBN 0 901272 40 X

Printed for ULDM Publications by
Richardson Print
Riverside Park, Caton Road
LANCASTER. LA1 3NX

PROCEEDINGS OF THE GST WORKSHOP, LANCASTER 1987

GEOMETRIZATION OF STATISTICAL THEORY

CONTENTS

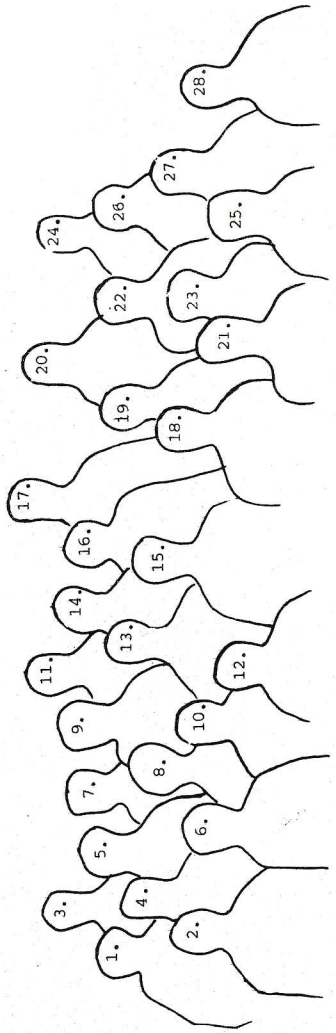
PREFACE

The idea for a Workshop on Geometrization of Statistical Theory arose from discussions with Wilfred Kendall at the Bernoulli Society Congress in Tashkent, September 1986, with subsequent encouragement from Peter Jupp and others. The time seemed ripe for a follow-up to the NATO Advanced Research Workshop, on Differential Geometry and Statistical Inference, organized by D.R. Cox at Imperial College 8-11 April 1984.

Our main objective was to bring together a significant number of specialists from geometry, statistics and analysis to stimulate discussion in an intensive workshop atmosphere, to describe current work and to identify future directions. The specialists responded admirably and their enthusiasm resulted in lively and general participation in the extensive discussion sessions. I am indebted to the Chairmen of these sessions for their hard work and for providing me with written reports within days of the end of the Workshop. The second objective was to publish the Proceedings without delay, while including a summary of salient points from the discussions; this has been achieved thanks to the cooperation of everyone concerned. We all appreciated the efforts of Jennifer Penwill for her efficient organization of the administrative details for the Workshop. Thanks are due to Sue Hubbard for the mathematical word processing and preparation of final copy for the printers, also to Stuart Wilkinson for assistance with proof reading. The London Mathematical Society and The Science and Engineering Research Council provided financial support, without which the Workshop would not have been possible. My only disappointment was that our colleagues N.N. Chentsov and Yu.V. Prohorov were not able to accept the invitation to participate in the Workshop.

The task of running the Workshop was enormously lightened by the good humour prevailing in quite intensive debates. Key phrases that recurr in my mind are 'beermat proof', 'do it in fifteen words', and the possible translation of 'empirical uncertainty' as 'unspeakability'. Whether derivative strings really are part of jet calculus, or conversely, remains to be seen! But, as will be apparent from these Proceedings, the geometrization of statistical theory continues apace, bringing the benefit of new ideas to geometry and statistics alike.

Kit Dodson
Lancaster
November 1987



1. M. HURLEY
2. J.M. PENWILL
3. G. TUNNICLIFFE WILSON
4. A. SCOTT
5. D. STASINOPOULOS
6. P.E. JUPP
7. S.L. LAURITZEN
8. K. BODWICK
9. W.S. KENDALL
10. M. GREEN
11. B.L. FOSTER
12. S.A. WILKINSON
13. P. SMITH
14. W.H. ROSS
15. D. WARREN
16. P. BLAESILD
17. O.E. BARNDORFF-WILSON
18. C.T.J. DODSON
19. M. MORA
20. P. ERIKSEN
21. P.K. BHATTACHARYYA
22. L. VERMEIRE
23. A.F.S. MITCHELL
24. R. FRYER
25. S-I. AMARI
26. B. PRUM
27. D. PICARD
28. B. HANZON

UNIVERSITY OF LANCASTER DEPARTMENT OF MATHEMATICS

WORKSHOP ON
GEOMETRIZATION OF STATISTICAL THEORY
28-31 October 1987

LIST OF PARTICIPANTS

S-I Amari
Department of Mathematical Engineering
Faculty of Engineering
University of Tokyo
Bunkyo-ku
Tokyo
Japan

O E Barndorff-Nielsen
Department of Theoretical Statistics
Institute of Mathematics
University of Aarhus
NY Monkegade
DK-8000 Aarhus C
Denmark

J Belcher
Department of Mathematic
Cartmel College
University of Lancaster

P K Bhattacharyya
School of Applicable Mathematics
Polytechnic of Central London
115 New Cavendish Street
London W1M 8JS

P Blaesild
Department of Theoretical Statistics
Institute of Mathematics
University of Aarhus
NY Monkegarde
DK-8000 Aarhus C
Denmark

K Bodwick
Department of Mathematics
Cartmel College
University of Lancaster

C T J Dodson
Department of Mathematics
Cartmel College
University of Lancaster

P Eriksen
Institute of Electrical Systems
Department of Mathematics & Computing
Strandvejen 19
DK-9000 Aalborg
Denmark

B L Foster
Department of Mathematical Sciences
University of Montana
Missoula
MT 59812
USA

B Francis
Centre for Applied Statistics
Cartmel College
University of Lancaster

R Fryer
Freshwater Biological Association
Ferry House
Ambleside
Cumbria LA22 0LP

M Green
Centre for Applied Statistics
Cartmel College
University of Lancaster

B Hanzon
Department of Mathematics
Delft University
P O Box 356
2600 AJ Delft
The Netherlands

M Hurley
Freshwater Biological Association
Ferry House
Ambleside
Cumbria LA22 0LP

P E Jupp
Department of Statistics
The Mathematical Institute
University of St Andrews
North Haugh
St Andrews KY16 9SS

W S Kendall
Department of Mathematics
University of Strathclyde
Livingstone Tower
26 Richmond Street
Glasgow G1 1XH

S L Lauritzen
Department of Pure Mathematics
and Mathematical Statistics
University of Cambridge
16 Mill Lane
Cambridge CB2 1SB

T J Lyons
Department of Mathematics
University of Edinburgh
James Clerk Maxwell Building
Mayfield Road
Edinburgh EH9 3JZ

A F S Mitchell
Department of Mathematics
Imperial College
Huxley Building
180 Queen's Gate
London SW7 2BZ

M Mora
27 Rue de la Chaine
3100 Toulouse
France

D Picard
Université Paris VII
Tour 4555
UER Mathematique et Informatique
5-eme etape, 2 Place Jussieu
75251 Paris
France

B Prum
Université Paris V
45 Rue des Saints Peres
75006 Paris
France

W H Ross
Department of Mathematics
Queen's University
Kingston
Ontario 798194
Canada K7L 3N6

A Scott
Centre for Applied Statistics
Cartmel College
University of Lancaster

P Smith
Department of Mathematics
Cartmel College
University of Lancaster

D Stasinopoulos

D Stott
Department of Mathematics
Cartmel College
University of Lancaster

M Thelwall
Department of Mathematics
Cartmel College
University of Lancaster

R Tucker
Department of Physics
University of Lancaster

G Tunnicliffe Wilson
Department of Mathematics
Cartmel College
University of Lancaster

L Vermeire
Katholieke Universiteit Leuven
Campus Kortijk
Faculteit Wetenschappen
B-8500 Kortrijk
Belgium

D Warren
Department of Mathematics
Cartmel College
University of Lancaster

S A Wilkinson
Department of Mathematics
Cartmel College
University of Lancaster

CONNECTION GEOMETRY

C.T.J.Dodson
Department of Mathematics
University of Lancaster

ABSTRACT

A mainly non-technical description is given of the role of connections in manifold geometry, aiming to motivate the definitions by developing an intuitive feel for what is needed. A differentiable structure is necessary to support geometrical theories, then choosing a connection serves to link up parallelism from point to point and provides an invariant gradient operator.

MANIFOLDS AND BUNDLES

The first 'complicated' geometrical spaces that we study are often encountered before we know much at all of mathematics, these are the surfaces constructed by folding sheets of paper. Such surfaces are 2-dimensional examples of **topological manifolds**, formed by a continuous pasting together of pieces of a Euclidean space. In applications of geometry, to navigation or to field theory, it turns out that, not only do we need to support continuity arguments, but we need also to provide for calculus.

It is fundamental that many real processes require a model that is sensitive to rates of change as well as to the values of quantities. A **differentiable manifold** is the least structure needed to support such a provision, it is formed by a differentiable pasting together of pieces of a Euclidean space. An introductory treatment of manifolds can be found in Dodson and Poston [12]; widely used more advanced texts are Bishop and Crittenden [1], Brickell and Clark [3], Lang [17], Spivak [22], Hirsch [14], Yano and Ishihara [23], Kobayashi and Nomizu [15,16]; for a fast overview of modern manifold geometry aimed at non-specialists, see [9] and for a detailed study with a view to applications see Crampin and Pirani [6].

Some formal operations (certain pullbacks, quotients and coproducts) are not possible in the category of manifolds (cf. Dodson [8], Hirsch [14], Lang [17]). However, such operations are possible in the category of vector bundles. Moreoever, some vector bundles always come as free superstructure on a smooth manifold, these express precisely the provision for differential calculus through the particular vector bundle of tangent vectors.

The **tangent bundle** TM to a smooth n-manifold M is itself a smooth manifold made from the set of all direction vectors available for all curves through points of M. There are many equivalent ways to effect this construction (cf. [12]) but the interesting thing is that TM need not look like $M \times \mathbb{R}^n$. For example, TS^1 does have the appearance of the cylinder $S^1 \times \mathbb{R}^1$, but TS^2 is not like $S^2 \times \mathbb{R}^2$, as is well known through the Hairy Ball Theorem.

We shall restrict attention for convenience to **smooth manifolds** and maps, so we always have differentiability of all orders. Consider a map

$$f : M \to N$$

it induces a map, the **derivative** of f,

$$Df : TM \to TN$$

between the corresponding tangent bundles. In components at $x \in M$, the map Df appears as the Jacobian matrix of partial derivatives of f, expressed in coordinates about x and f(x). In particular, Df is linear on each tangent vector space T_xM.

Bundles, like TM, allow for a generalization of the idea of a map. For, given any map

$$h : S^2 \to \mathbb{R}^2 \quad : \quad x \to h(x)$$

it is expressible through its 'graph' in $S^2 - \mathbb{R}^2$ space as a map

$$\Gamma h : S^2 \to S^2 \times \mathbb{R}^2 \quad : \quad x \to (x,h(x)).$$

Actually, this $\hat{h}$ is a **section** of the vector bundle $S^2 \times \mathbb{R}^2$ over S because it satisfies the identity

$$\pi \circ \hat{h} = 1_{S^2}$$

where $\pi : S^2 \times \mathbb{R}^2 \to S^2 : (x,v) \to x$ is the canonical bundle projectio and 1_{S^2} is the identity function on S^2. This we denote by $\hat{h} \in Sec(S^2 \times \mathbb{R}^2/S^2)$. Next consider the tangent bundle

$$\pi_T : TS^2 \to S^2$$

and a section of it, that is a tangent vector field on S^2,

$$\sigma : S^2 \to TS^2 \quad \text{with} \quad \pi_T \circ \sigma = 1_{S^2}.$$

So we have $\sigma \in Sec(TS^2/S^2)$. Now, in coordinates around a point in S^2, σ has the appearance of an $\mathbb{R}^2$-valued function simil to h above. But this is only locally true because we know th TS^2 is not globally like $S^2 \times \mathbb{R}^2$. Hence, σ gives us a family functions like h, one on each coordinate patch of S^2, and join up smoothly; but, unless σ is the zero tangent vector everywhere, will not be expressible over the whole of S^2 as a single functi like h. So, in general, a tangent vector field on an n-manifold M is not simply an $\mathbb{R}^n$-valued function, it has to accommodate t twisted way in which the tangent copies of $\mathbb{R}^n$ roll around t manifold. It is in this sense that sections of bundles generali ordinary functions.

The sections of any fixed vector bundle form a module over the ring of real valued functions on a manifold, so there always convenient algebraic structure to hand. For the particul case of the tangent bundle they also possess a Lie algebra structu under the commutator bracket. Indeed, although it is intuitively appealing to describe tangent vectors as representatives of curves through a particular point, their most powerfully effective representation is as partial differential operators on the ring of smooth real functions and it is in this role that their Lie algebra structure becomes important.

Given a tangent vector field $\sigma \in Sec(TM/M)$ then

$$\sigma : M \to TM$$

(for example, an east wind on the surface of the earth) we immediately have its derivative

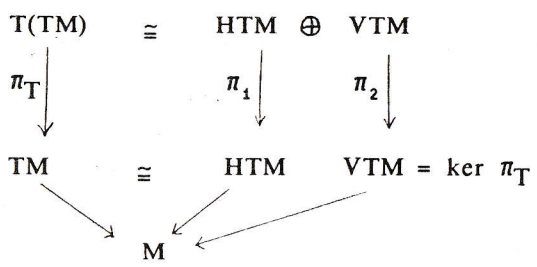
$$D\sigma : TM \to T(TM), \text{ and } D\sigma \in Sec\,(T(TM)/TM).$$

However, as a measure of the rate of change of σ over M it suffers from the disadvantage that it takes values in T(TM), not in TM like σ. This is perhaps the most convincing reason for introducing connections, they give unique ways to project $D\sigma$ back from T(TM) to TM while still preserving useful information about the full derivative. On a manifold a connection generalises the vector calculus notion of the gradient operator.

CONNECTIONS

Evidently, to provide an unambiguous way to project from T(TM) to TM we simply need to split T(TM) into two parts smoothly over TM, with one part (called **horizontal**) isomorphic to TM. The correct splitting in a vector bundle is by means of the Whitney sum, which is the bundle version of the direct sum of

vector spaces. We represent the process as follows:



or more elegantly as the split exact sequence

$$0 \to \mathrm{VTM} \hookrightarrow \mathrm{T(TM)} \underset{\Gamma}{\overset{\pi_T}{\rightleftarrows}} \mathrm{TM} \to 0.$$

Here Γ is a right inverse of π_T, and its image in T(TM) is precisely HTM in the previous diagram. Of course, Γ is then a linear fibred morphism over M, and it is one of many ways to characterise a **linear connection** on M. The problem of characterising rates of changes of tangent vector fields is now solved as follows using a linear connection Γ.

Take u,v $\in$ Sec(TM/M), then the **covariant derivative** of v with respect to u is denoted $\nabla_u v$ and defined to be the image in Sec(TM/M) of the horizontal part of Dv at u. In coordinates,

$$\nabla_u v = (u^i \partial_i v^k + \Gamma^k_{ij}\, u^i v^j)\, \partial_k$$

where the connection is represented by the Christoffel symbols Γ^k_{ij} defined on local basis fields ∂_i, ∂_i by

$$\nabla_{\partial_i} \partial_j = \Gamma^k_{ij}\, \partial_k.$$

Our description of a connection as a choice of splitting of T(TM) yields a connection in TM with appropriate linear properties. A similar construction would yield a connection in another vector bundle, or in any other fibre bundle by fixing appropriate nonlinear properties. For example, in principal G-bundles like the bundle of linear frames, invariance under G is required of the horizontal splitting. This, and general connections on fibred manifolds are studied in [4], [7], [15,16], [19,20] for example. Connections always exist but they do constitute extra structure, the exercise of some choice.

Returning to a connection Γ in TM, it induces the covariant derivative operator ∇ and this admits expression very elegantly in terms of a map which controls the definition of parallelism for the manifold.

Let

$$c : [0,1] \to M$$

be a curve in M with tangent vector

$$\dot{c} : [0,1] \to \mathrm{TM}.$$

Now consider the restriction to c of a tangent vector field w $\in$ Sec(TM/M), or let w be any tangent vector field along c. Then the covariant derivative of w along c is

$$\underset{\dot{c}}{\nabla} w = \left(\frac{dw^k}{dt} + \Gamma^k_{ij}\, w^i \dot{c}^j\right) \partial_k \qquad \text{where } \dot{c} = \dot{c}\,\partial_i,\ w = w^i \partial_i.$$

If $\underset{\dot{c}}{\nabla} w$ is zero along c then we say that w is a **parallel field** along c. Reversing the argument, we can be given c and a vector w_0 in $T_{c(0)}M$, the tangent space to M at c(0), then seek to develop

from w(o) a tangent vector field w along c such that $\nabla_{\dot{c}} w = 0$. For all sufficiently small t, the corresponding system of differential equations will admit a unique solution

$$\tau_t : T_{c(o)}M \to T_{c(t)}M : w_o \to w_t$$

This turns out to be a useful isomorphism called **parallel transport** along c; it relates to the covariant derivative through the limit

$$(\nabla_u w)_x = \lim_{t \to o} \frac{1}{t} [\tau_t^{-1} w \circ c(t) - w \circ c(o)]$$

where c is any curve through x = c(o) representing $u \in T_x M$ by $\dot{c}(o) = u$. The parallel transport map is to a general connection what a pair of parallel rulers is to a navigator.

A **geodesic** on a manifold with connection is a curve whose tangent vector is parallel along the curve. This implies satisfaction of the equation

$$\tau_t(\dot{c}(o)) = \dot{c}(t) \quad \text{or} \quad \nabla_{\dot{c}} \dot{c} = \underline{o}$$

which is always possible for any c(o), for a small enough interval of t.

Example Let $M=\mathbb{R}^2$ with standard coordinates and take the (non-standard!) connection with constant Christoffel symbols

$$\Gamma^1_{12} = \Gamma^1_{21} = 1, \quad \Gamma^2_{11} = \Gamma^2_{22} = \Gamma^1_{11} = \Gamma^2_{12} = \Gamma^2_{21} = 0.$$

So $\nabla_{\partial_1} \partial_1 = \nabla_{\partial_2} \partial_2 = \underline{o}$ but $\nabla_{\partial_1}(\partial_1 + \partial_2) \neq \underline{o}$.

Then the lines parallel to the x or y axes are geodesics but the line y=x is not a geodesic in this space. Along the parabolic curve

$$c : [0,1] \to \mathbb{R}^2 : t \to (t,t^2)$$

we can solve the parallelism equation $\nabla_{\dot{c}} w = \underline{o}$, taking each of ∂_1, ∂_2 from the standard basis for $T_{c(o)}M$ as initial condition in turn. Then, since parallel transport is linear, we deduce that it is given in matrix coordinates by

$$\tau_t : \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \to \begin{pmatrix} e^{-t^2} & k(t) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

where $k(t) = -e^{-t^2} \int_o^t e^{x^2}\, dx$.

In particular, we see that this parallel transport is not the usual Euclidean one because it sends the (unit) vector ∂_1 at the origin c(o) to $e^{-t^2} \partial_1$ at c(t); moreover, ∂_2 actually gets rotated as well.

The last remark highlights the detachment in general of a connection from any metric structure. It is clear what would be desirable if we have both : simply that the parallel transport isomorphism should be an isometry, and so preserve all lengths and angles. When this happens we say that the connection is **compatible** with the metric. Given a metric, it turns out that many connections have this property (in our example the Γ was not compatible with the standard Euclidean metric). However, there is only one among them for which the Christoffel symbols are **symmetric**

$$\Gamma^k_{ij} = \Gamma^k_{ji}$$

namely, the **metric** or **Levi-Civita connection**, and it is the one normally used in the presence of a metric tensor. This is so whether the metric tensor is positive definite, as in Riemannian geometry, or indefinite as in the Lorentz geometry of general relativity. Again, a metric is a choice of extra structure on a manifold but with it we always have a distinguished connection.

Geodesics are defined for any connection but they have a particular importance in Riemannian geometry. For there we can use as parameter the **length** of a curve,

$$\ell(t) = \int_0^t \|\dot{c}(t)\|,$$

where the norm of the tangent vector is given in terms of the metric tensor g by

$$\|\dot{c}(t)\|^2 = g(\dot{c}(t),\dot{c}(t)).$$

In fact, the name geodesic means 'divides the earth' - like great circles on S^2. Now we can define, for all $x \in M$ and some small enough neighbourhood N_x of the zero vector in T_xM, the **exponential map** at x

$$\exp_x : N_x \to M : v \mapsto c(1)$$

where c: [0,1] → M is the geodesic for which c(o) = x, $\dot{c}(o) = v$. Thus, $\exp_x$ sends $v \in N_x \subseteq T_xM$ to the point at unit distance along the geodesic through x which has tangent vector v at x. In fact, we can always choose N_x small enough to make $\exp_x$ a diffeomorphism onto its image. If (M,g) is a **complete** Riemannian manifold, that means every geodesic admits extension to arbitrary parameter values, then $\exp_x$ has domain T_xM; moreover, in that case every pair of points in a connected component of M can be joined by a geodesic that is minimal in length with respect to all nearby curves joining the points.

Suppose that at $x \in M$ the map $\exp_x$ restricts to a diffeomorphism f_x on $V \subseteq T_xM$ and denote by U the image $\exp_x V \subseteq M$. Then (U, f_x^{-1}) gives the very useful system of **normal coordinates** about x. With respect to normal coordinates the components of the metric tensor at x have the simple form

$$g_{ij}(x) = \delta_{ij} \text{ and } \partial_k g_{ij}(x) = o$$

moreover

$$\Gamma^k_{ij}(x) = o$$

for all i,j,k. It must be emphasized that this simplification happens only at x, not in a region (cf. Dodson and Poston [12] p 338 et seq. for more details).

The logical next step is to study the curvature and torsion of a linear connection. These objects are easily defined in terms of the machinery we have set up, as tensor fields:

Torsion T takes pairs of tangent vector fields u,v and yields a new tangent vector field

$$T(u,v) = \nabla_u v - \nabla_v u - [u,v]$$

where

$$[u^i\partial_i, v^j\partial_j] = (u^i\partial_i v^k - v^j\partial_j u^k)\partial_k$$

is the commutator of u and v.

Curvature R takes triples of tangent vector fields u,v,w and yields a new tangent vector field

$$R(u,v)w = \nabla_u\nabla_v w - \nabla_v\nabla_u w - \nabla_{[u,v]}w.$$

We can see immediately that the components of the torsion are given by

$$T(\partial_i,\partial_j) = \left(\Gamma^k_{ij} - \Gamma^k_{ji} \right) \partial_k$$

so it is zero precisely when Christoffel symbols are symmetric. It is a theorem that the curvature is zero precisely when, for all pairs of points, parallel transport between them is independent of the path followed. There is, however, great benefit in pursuing the detailed study of connection, curvature and torsion, by means of their corresponding differential forms and the exterior calculus (cf. [15], [21]). Then we find that in quite general cases a curvature is essentially the curl of its connection. Differential forms also provide the appropriate link between global differential geometry and algebraic topology (cf. Bott and Tu [2] and Mackenzie [18]).

It is natural to consider families of connections on a given manifold, for example in the study of the stability of connection - related properties like completeness (cf. [4], [7]). The problem immediately encountered is the infinite - dimensionality of the space of all connections on a given manifold. This can be circumvented rather elegantly by the idea of a **system of connections** which was introduced by Mangiarotti and Modugno (cf. [19], [20]). It allows significant finite-dimensional subspaces to be selected and admits a powerful geometrical calculus [11]. Its relevance to parametric models in statistical theory is proposed in [10]. The apparent importance of systems of connections lies in the fact that each carries its own universal connection, allowing a geometrical study of the whole family of connections in the system. This theory requires a generalization of the notion of a connection from that on a fibre bundle to one on a fibred manifold (cf. [7], [10] for more details). There the appropriate definition of a connection is as a section of the first jet bundle. Another powerful new formalism in the theory of connections uses the notion of **Lie algebroids**, and a detailed self-contained treatment has been given by Mackenzie [18]. This approach generalizes, for example, the integrability theory used on principal bundles which resolves when a differential 2-form is the curvature of some connection.

REFERENCES

1. R.L. Bishop and R.J. Crittenden. **Geometry of Manifolds** Academic Press, New York 1964.

2. R.Bott and L.W.Tu. **Differential Forms in Algebraic Topology.** Graduate Texts in Mathematics 82 Springer-Verlag, Berlin 1982.

3. F. Brickell and R.S.Clark. **Differentiable Manifolds.** Van Nostrand, London 1970.

4. D.Canarutto and C.T.J.Dodson. On the bundle of principal connections and the stability of b-incompleteness of manifolds. Math. Proc. Camb. Phil. Soc. 98, (1985) 51-59.

5. L.A.Cordero, C.T.J.Dodson and M.deLeon. Differential Geometry of Frame Bundles (book in preparation for Reidel 1988).

6. M.Crampin and F.A.E.Pirani. **Applicable Differentiable Geometry,** LMS Lecture Notes 59, CUP Cambridge 1986.

7. L.DelRiego and C.T.J.Dodson. Sprays, universality and stability. Math.Proc. Camb. Phil. Soc. 103 (1988) (in press).

8. C.T.J.Dodson. **Categories, Bundles and Spacetime Topology,** Shiva, Orpington 1980 (2nd Ed. Reidel 1988).

9. C.T.J.Dodson. Manifold geometry in **Encyclopedia of Physical Science and Technology Volume** 7, Academic Press,San Diego 1987 pp. 619-646.

10. C.T.J.Dodson. Systems of connections for parametric models. **Proc. GST Workshop** Lancaster 28-31 October 1987.

11. C.T.J.Dodson and M.Modugno. Connections over connections and a universal calculus. **Proc. VI Convegno Nazionale di Relativita Generale e Fisica della Gravitazione,** Florence 10-13 October 1984.

12. C.T.J.Dodson and T.Poston. **Tensor Geometry** Pitman London 1979.

13. W.Greub, S.Halperin and R.Vanstone. **Connections, Curvature and Cohomology Vol II,** Academic Press New York 1973, Ch. V11.

14. M.W.Hirsch. **Differential Topology,** Springer-Verlag, New York, 1976.

15. S.Kobayashi and K.Nomizu. **Foundations of Differential Geometry I,** Interscience, New York 1963.

16. S.Kobayashi and K.Nomizu. **Foundations of Differential Geometry II,** Intersicence, New York, 1969.

17. S.Lang. **Differentiable Manifolds.** Addison-Wesley, Reading, Massachusetts 1972.

18. K.Mackenzie. **Lie Groupoids and Lie Algebroids in Differential Geometry.** LMS Lecture Notes 124, CUP, Cambridge 1987.

19. L.Mangiarotti and M.Modugno. Fibred spaces, jet spaces and connections for field theories. **Proc. Int. Meeting on Geometry and Physics** Florence 12-15 October 1982, Pitagora Editrice Bolgona 1983 pp. 135-165.

20. M.Modugno. An introduction to systems of connections. Preprint. Istituto Mat. Applicata 'G.Sansone' Univ. Firenze 1986.

21. M.Spivak. **Calculus on Manifolds,** Benjamin New York 1965.

22. M.Spivak. **A Comprehensive Introduction to Differential Geometry.** Volumes 1 to 5, Publish or Perish, Boston 1970.

23. K.Yano and S.Ishihara. **Tangent and Cotangent Bundles** Marcel Dekker, New York 1973.

LIKELIHOOD

G. Tunnicliffe Wilson
Department of Mathematics
University of Lancaster

ABSTRACT

A brief introduction is presented, of the concept of the likelihood function derived from a family of probability distributions. The likelihood function is the basis of parametric statistical inference, providing for example, maximum likelihood estimates of parameters. The properties of these estimates, and some of the difficulties encountered in a two parameter model, are illustrated in the presentation.

LIKELIHOOD RATIOS

The simplest inference problem is to decide from which one o just two distinct distributions an observation has arisen. Fo example Figure 1 shows two binomial distributions, referred to $f_\theta(k)$ and $f_\phi(k)$.

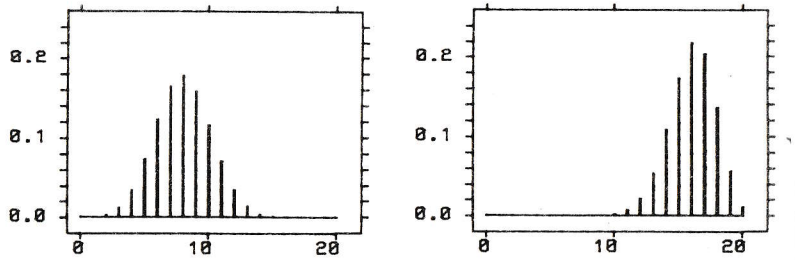


Figure 1

Graphs of the Binomial distributions for 20 trials and respective probabilities of success 0.4 and 0.8.

These are the probabilities of observing k successes in 20 trials, = 0. . . 20, when the probabilities of success in each trial ar θ=.4 and ϕ=.8 respectively. Suppose that the observed number o successes is x = 11, and assume that one of the two distributio is the true one from which x has arisen.

The likelihood principle states that our decision as to whether or ϕ is the correct parameter, should depend only on the relati magnitudes of $f_\theta(x)$ amd $f_\phi(x)$, which in this case are in th ratio of about 10:1. This principle may be deduced from ver plausible premises, and is well argued by Berger and Wolpert [3]. An important consequence of these arguments is that the decision should not depend on any other part of the displayed distributions than their values at k=x. In particular the 'tail areas' of the first distribution to the right of x, and of the second distribution to the left of x, are not relevant to the decision.

Nevertheless, the likelihood ratio value of 10:1 in this case is commonly thought to require some interpretation, since it does not mean for example that the respective probabilities of the two distributions are in that ratio, i.e. 91% and 9% approximately. Such a statement can be made however, if prior probabilities π_θ, π_ϕ can be given for the two possibilities, in which case Bayes' formula gives the posterior probabilities (i.e. taking the observed value of x into account) as being in the ratio $\pi_\theta f_\theta(x) : \pi_\phi f_\phi(x)$. Clearly then, if we are concerned to maximise the frequency with which we make correct decisions, we would choose the possibility favoured by these posterior odds, though if loss functions were relevant (i.e. differing costs for making an incorrect decision) this possiblility may be modified. Even then the decision depends on the observation x only via the ratio $f_\theta(x) : f_\phi(x)$.

It is possible however, to provide a 'sampling frequency' justification for the use of this ratio without relying on a Bayes procedure. The Neyman-Pearson lemma, commonly cited in statistical texts such as Cox and Hinkley [4] proves that

the following likelihood procedure has optimal statistica properties:

$$\left.\begin{matrix}\text{choose } \theta \\ \text{or } \phi\end{matrix}\right\} \text{ according to whether } \frac{f_\theta(x)}{f_\phi(x)} \left\{\begin{matrix}\geqslant \\ <\end{matrix}\right\} c$$

where c is some critical value

The optimality resides in the fact that no other procedure ca simultaneously enjoy smaller values of the two error probabiliti defined by :

$$\left.\begin{matrix}\alpha \\ \beta\end{matrix}\right\} = P \text{ (deciding on } \left\{\begin{matrix}\phi \\ \theta\end{matrix}\right\} \text{ when in fact } \left\{\begin{matrix}\theta \\ \phi\end{matrix}\right\} \text{ is true).}$$

The value selected for c simply provides a trade off between these errors. In our example (as in many others) the likelihood ratio procedure with the choice of c, is equivalent to th following procedure with a choice of k:

$$\text{choose } \left\{\begin{matrix}\theta \\ \phi\end{matrix}\right\} \text{ according the whether } x \left\{\begin{matrix}\leqslant \\ >\end{matrix}\right\} k.$$

In that case, the error probabilities α and β are the tail areas the right and left of k in the two respective distributions, an despite their irrelevance to the likelihood prinicple, are oft used to fix k and hence the procedure. Thus k (and implicitl c) may be chosen so that α=5%. This policy means that c woul vary from one decision problem to another, whereas overa statistical efficiency is maintained by using the same value of throughout.

There is some reason for choosing the 'natural' value of c=

since it may also be proved by Jensens inequality that under very general conditions,

$$E(\log \{f_\theta(x)/f_\phi(x)\}) > 0 \qquad (1)$$

when θ is true, and vice versa, where E is expectation with respect to the true distribution. In many applications the argument of this expectation will, by laws of large numbers, be close to its expected value, so that with probability close to one $\log(f_\theta(x)/f_\phi(x))$ will be greater than or less than 0, i.e. $f_\theta(x)/f_\phi(x)$ greater or less than 1, according to whether θ or ϕ is the true parameter. The correct decision is therefore highly probable. In the case of the example with which we started this section, the error probabilities consequent on using a value of c=1, are α = .0211 and β = .0321.

MAXIMUM LIKELIHOOD ESTIMATION

We continue with our example of an observation x from a binomial distribution, but suppose now that we are not restricted to two possible parameters; rather that the unknown parameter θ may have any value in the interval [0,1]. In this case we make an inference about θ from considering $f_\theta(x)$ as a function of θ, for the given value of x. This is the likelihood function, plotted in Figure 2a for our example, with x=11. Note that only relative values of this function are important - it is commonly scaled so that its maximum value is 1. Again we comment that

this may only be interpreted as a probability density for θ if it

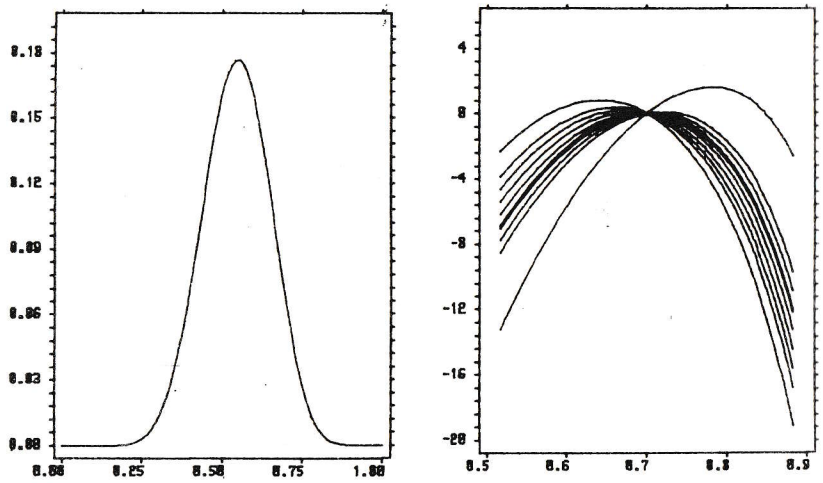


Figure 2

(a) The likelihood function for the probability parameter θ, given on observation of x=11 successes in 20 trials.

(b) Log likelihood functions for 8 different observations sampled from the Binomial (100,0.7) distribution. The heavy line is the expected log likelihood.

The maximum likelihood estimate is defined by the value $\hat{\theta}$ = arg max $f_\theta(x)$ - the value of θ for which $f_\theta(x)$ is a maximum. For both the theoretical investigation and practical application of this procedure, it is common to use $\ell(\theta) = \log f_\theta(x)$, the log-likelihood function. Let us now use θ_0 for the true, but unknown value of θ, so that the inequality (1) may be written

$$E\{\ell(\theta) - \ell(\theta_0)\} \leq 0 \qquad (2)$$

i.e. θ_0 = arg max $E\{\ell(\theta)\}$.

The importance of this result, due to Fisher, is that it is globally true, for all θ_0. The argument, in general terms, for use of the maximum likelihood estimator, is that we may commonly expect $\ell(\theta)$ to be close to $E(\ell(\theta))$, and hence $\hat{\theta}$ = arg max $\ell(\theta)$, to be close to θ_0= arg max $E(\ell(\theta))$.

To illustrate this, consider Figure 2b which shows graphs of $\ell(\theta)-\ell(\theta_0)$ for several different value of x, sampled from the true distribution which has now been taken as Binomial (n,θ) with n = 100 and θ_0=.7. Also graphed as a heavier line is $E\{\ell(\theta)-\ell(\theta_0)\}$. In practice of course $\ell(\theta_0)$ is not known, but additive constants (though possibly depending on x) do not affect inference using the log-likelihood. In particular, the position of $\hat{\theta}$ is unaffected, and the use of $\ell(\theta)-\ell(\theta_0)$ in this graph is designed to illustrate the sampling properties of the inference.

Note that although max $E\{\ell(\theta)-\ell(\theta_0)\}$ = 0 (at θ_0), we have always max $\{\ell(\theta)-\ell(\theta_0)\} \geq 0$ and $E(\max\{\ell(\theta)-\ell(\theta_0)\}) > 0$.

In cases such as the one illustrated, where the sample likelihood functions are well approximated by quadratics near their maximum, expansions of the likelihood function about the true θ_0 permit calculation of approximate distributional properties of $\hat{\theta}$

In particular, quantities such as

$$E\{(\ell^{(n)}(\theta_0))^m\}, \qquad (3)$$

the moments of derivatives of $\ell(\theta)$, are important. The most

important low order approximations are:

$$\hat{\theta} \simeq -\ell'(\theta_0)/\ell''(\theta_0) \qquad (4)$$

$$2\{\ell(\hat{\theta}) - \ell(\theta_0)\} \simeq -\ell'(\theta_0)^2/\ell''(\theta_0). \qquad (5)$$

There are hosts of identities relating the moments of the derivatives, such as:

$$E\{\ell'(\theta_0)\} = 0 \qquad (6)$$

$$E(\{\ell'(\theta_0)\}^2) = -E\{\ell''(\theta_0)\}. \qquad (7)$$

In many contexts the distribution of $\ell'(\theta_0)$ is well approximated by a Normal or Gaussian density, and the following approximations are justified:

$$\hat{\theta} \sim \text{Normal } (0, -1/E(\ell''(\theta_0))) \qquad (8)$$

$$2\{\ell(\hat{\theta}) - \ell(\theta_0)\} \sim \chi^2_1. \qquad (9)$$

The first of these approximations is sensitive to the parameterisation, and a suitable choice may considerably enhance its accuracy. Consider for example Figure 3a, which is similar to Figure 2b except that now the true parameters of the binomial distribution are n=500 and θ_0=.015. The asymmetry of the log likelihood functions, and of the distribution of their maxima, are quite clear. Figure 3b however shows the same graphs on the horizontal scale of $\theta^{1/3}$. The graphs now appear remarkably symmetric. The reason is that for this case, when the binomial distribution is close to a Poisson distribution, the transformation to $\theta^{1/3}$ annihilates the third derivative of the expected likelihood function at its maximum. This is again a global property, i.e. the transformation does not depend on the unknown true

parameter value.

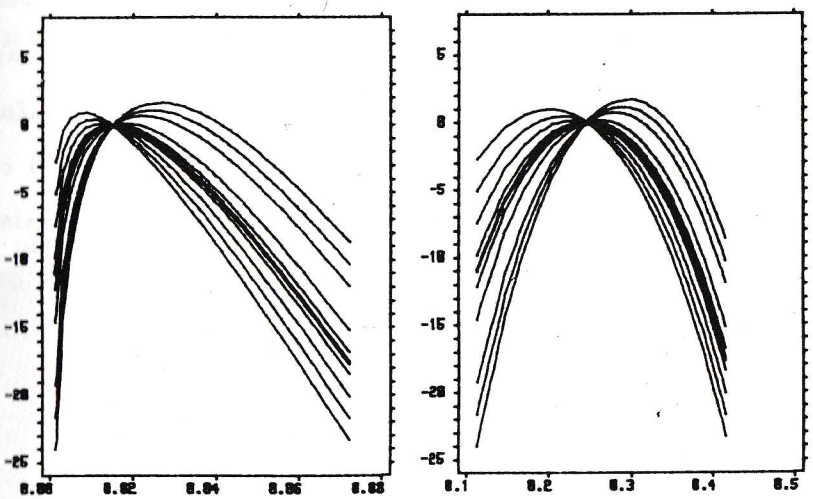


Figure 3

(a) Log likelihood functions for 9 different observations sampled from the Binomial (500,.015) distribution. The heavy line is the expected log likelihood.

(b) The same as (a) but on a scale of $\theta^{1/3}$.

The second of the above approximate distributions in (9) is particularly useful since the quantity $2\{l(\hat{\theta})-l(\theta_0)\}$ is parameterisation invariant. Knowing that a parameterisation exists for which the expected likelihood is quadratic (and the approximations therefore more justifiable) is reassuring, but knowledge of this parameterisation is not required. Now a χ^2_1 variate lies between 0 and 3.84 with 95% probability. Looking for example at Figure 2b, this means that the maximum value of the likelihood function will, with 95% probability, not exceed the

value at the true (but unknown) parameter, by more than 3.84/2=1.92. This may be rephrased by drawing across each likelihood function a line at a height of 1.92 below its maximum value. The interval about the maximum for which the function exceeds this 'contour' then has the property that it will contain the true parameter with a sampling probability of (approximately) 95%. This *likelihood* interval may therefore be in practice interpreted as an approximate confidence interval for the unknown parameter.

TWO PARAMETER LIKELIHOODS

There is no difficulty in principle in extending the use of maximum likelihood estimation, and the construction of likelihood regions, to problems with two or k parameters. In this case the level of the contour defining the region is derived from the chi-squared distribution with k degrees of freedom. The main difficulty is how to make a statement about just one of the parameters, for which purpose the remaining parameters are regarded as *nuisance* parameters. We illustrate two of the concerns which arise, by considering estimation for the location and scale parameters μ,σ for the Normal density, given a random sample of n observations. Thus from the density

$$f(x_1,\ldots,x_n: \mu,\sigma) \propto (1/\sigma^n) \exp\{-\tfrac{1}{2} \Sigma_1^n (x_i-\mu)^2/\sigma^2\}$$

we obtain, using the parameter $\phi = 1/\sigma^2$, and setting $\theta=\mu$ for continuity of notation, the log likelihood function

$$\ell(\theta,\phi) = \tfrac{1}{2}\{n\log \phi - \phi\Sigma(x_i - \theta)^2\} \quad .$$

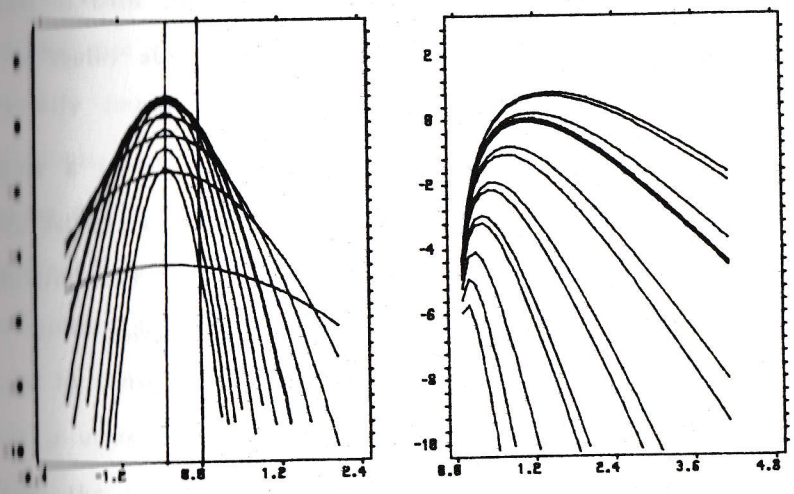


Figure 4

Sections through the two parameter log likelihood surface derived from 6 observations from a Normal density

(a) shown as a function of θ, for values of ϕ from 0.1 (flat section) to 4.1 (peaked section)

(b) shown as a function of ϕ, for values of θ from -2.1 to +2.1.

This is represented graphically in Figure 4 for the case of a particular sample of 6 values from the Normal density for which, without loss of generality, we take $\mu=0$ and $\sigma=1$. Figure 4a shows sections through the likelihood surface as a function of θ, for a selection of values of ϕ, and in Figure 4b the roles of the

parameters are interchanged. We therefore have an 'end-on' an
'side-on' view of the function. In Figure 4a the section for ϕ=
is emphasised, and similarly the section for θ=0 in Figure
These show the likelihood curves which should be used to mak
inferences about θ and ϕ, if, respectively, the true values of
and θ were known to be 1 and 0.

Without this knowledge a common way of extracting a singl
function of the parameter of interest, say θ, is to maximise ove
the nuisance parameter, ϕ, for each value of θ. This provide
what is known as the maximised log likelihood function:

$$p(\theta) = \max_{\phi} \ell(\theta,\phi) = \ell(\theta,\hat{\phi}(\theta))$$

also known as the log profile likelihood, since in Figure 4a it i
represented by the upper envelope of the sections, or the end-o
profile of the surface. The motivation is that ϕ is replaced b
$\hat{\phi}(\theta)$, its maximum likelihood estimate for each given value of
The justification is that, to the usual order of asymptoti
approximation, the results (8) and (9) remain valid when applie
to the profile function $p(\theta)$, and this is a very general result
Note that in the general case there will be some loss o
efficiency through the necessity to eliminate a nuisanc
parameter, and this would be apparent in the smaller expecte
value of the second derivative $-E\{p''(\theta_0)\}$, compared wit
$-E\{\ell''(\theta_0,\phi_0)\}$, where the derivatives are w.r.t. θ. In our example
these are the same, neglecting terms of order 1/n, but this is du

is an *orthogonality* condition, that at θ_0, ϕ_0.

$$E\{\partial^2 \ell(\theta_0,\phi_0)/\partial\theta\partial\phi\} = 0.$$

There are however, valuable improvements of order (1/n) which can be made in the use of the profile likelihoods.

In the case of $p(\theta)$ for our example, the approximation (9) is usefully improved by the Bartlett correction factor. The χ^2_1 distribution is simply scaled up by the factor $b = \{1+3/(2\vartheta)\}$ where ϑ = n-1. Thus in constucting the likelihood interval for a nominal 95% confidence, the contour at 1.92 below the maximum is replaced by one at 1.92b. Now for this example it is known how to construct for small n an exact 95% confidence interval for μ using the sample mean, standard deviation, and a value from the tabulated 't' distribution, in place of the asymptotic value of 1.96. The Bartlett correction factor in fact leads to an interval of the same structure, but with the value tabulated as h below, in place of the value of t.

ϑ	1	2	5	10	20
t	12.71	4.303	2.571	2.228	2.086
h	10.99	4.099	2.548	2.223	2.085

Table 1

For all practical purposes, the likelihood interval constructed using the correction factor is as good as the usual 't' interval.

For a recent note on the Bartlett adjustments see

Barndorff-Nielsen and Blaesild [2].

Consider now Figure 4b which reveals a different small sample effect (of order 1/n), a bias in $\hat{\phi}$. The profile log likelihood which we will call $q(\phi)$, is in this case a particular section, $\ell(\hat{\theta},\phi)$, because $\hat{\theta}$ does not depend on ϕ. It is easily demonstrated that $\hat{\phi}$ always lies to the right of the value $\hat{\phi}_0$ say, which would be obtained by using the section through θ_0. The likelihood intervals for ϕ would similarly be displaced and extended. This point may be summarised by the statements that ϕ_0=arg max $E\{\ell(\theta_0,\phi)\}$, but $\phi_0 \neq$arg max $E\{q(\phi)\}$. Thus the profile likelihood does not necessarily share the single parameter likelihood property (2). An associated loss is the basic property (6), i.e., using superscripts to denote derivatives,

$$E(q^{\phi}(\phi_0)) \neq 0.$$

Looking at Figure 4a, it is possible to see the extent to which moving from $\ell(\theta_0,\phi)$ to $\ell(\hat{\theta},\phi)$ causes the bias. The gain in the section increases as ϕ increases and has its source in the fact that $-\ell^{\theta\theta}$ is increasing, i.e. $-\ell^{\theta\theta\phi} > 0$ at θ_0,ϕ_0. Now the actual increase is also proportional to $\frac{1}{2}(\hat{\theta}-\theta_0)^2$, which is unknown, but whose expected magnitude may be well approximated by $-\frac{1}{2}\{\ell^{\theta\theta}\}^{-1}$ at θ_0,ϕ_0. This motivates a correction of q^{ϕ} to $\tilde{q}^{\phi}=q^{\phi} - \frac{1}{2}\{\ell^{\theta\theta}\}^{-1}\ell^{\theta\theta\phi}$ which at ϕ_0 ensures that the property

$$E(\tilde{q}^{\phi}) = 0$$

is now more nearly true, to a higher order of approximation. Now $\tilde{q}^{\phi}$ is the derivative of $\tilde{q} = q - \frac{1}{2}\log\{-\ell^{\theta\theta}\}$. In practice

$\ell^{\theta\theta}(\theta_0,\phi)$ is replaced by the known value $\ell^{\theta\theta}(\hat{\theta}(\phi),\phi)$ and $\tilde{q}(\phi)$ is then used in place of $q(\phi)$. For our example this leads to $\tilde{q}(\phi) = \{(n-1)\log\phi - \phi\Sigma(x_i-\bar{x})^2\}$, and thence to the modified maximum likelihood estimate of $\hat{\sigma}^2 = \{1/(n-1)\}\ \Sigma(x_i-\bar{x})^2$, i.e. with the 'degrees of freedom' divisor modified from n to n-1.

The statistical justification for using modified profile likelihood functions of the form $\tilde{q}$, is much more subtle than is presented here. The orthogonality between θ and ϕ is essential to the derivation, and the reader should consult Cox and Reid [5] for a recent treatment of this general problem. The modification as presented is not invariant to a reparameterisation of the nuisance parameters, so a further Jacobian term is seen in some expressions for the modified profile - see for example Barndorff-Nielsen [1].

CONCLUSION

The likelihood function plays a central role in statistical inference. This brief account of the topic covers much that is now classical. The intention has been to provide an introductory background for those who are not already familiar with the concept of likelihood inference. More recently the problems touched upon in this paper have benefitted from geometrical theory and procedures which are the subject of the central papers of this workshop.

REFERENCES

1. O.E.Barndorff-Nielsen. **On a formula for the distributio of the maximum likelihood estimato** Biometrika, 70, (1983) 343-365

2. O.E.Barndorff-Nielsen and P.Blaesild. **A note on th calculation of Bartlett adjustment** J.Royal. Statist.Soc. (B), 48, (1986) 353-358

3. J.O.Berger and R.L.Wolpert. **The Likelihood Principle,** Hayward Institute of Mathematical Statistic 1984

4. D.R.Cox and D.V.Hinkley. **Theoretical Statistic** Chapman and Hall, London, 1974

5. D.R.Cox and N.Reid. (1987). **Parameter orthogonali and approximate conditional inference.** J. Royal. Statist.Soc. (B), 49, (1987), 1-39

CONJUGATE CONNECTIONS IN STATISTICAL THEORY

Steffen L. Lauritzen

Department of Mathematics and Computer Science

Aalborg University, Denmark.

ABSTRACT

We review the basic facts about statistical manifolds i.e. Riemannian manifolds with a pair of mutually conjugate, in general non Riemannian, connections.

In particular we indicate a general abstract construction of a double embedding into a pair of affine spaces in duality, covering all known examples of such manifolds arising in statistical theory. We indicate how the method of quasi-likelihood can be conceived of as a geometric fitting procedure to be interpretable as projection in a statistical manifold. Finally, a condition due to Lauritzen and Picard for uniqueness of such a projection is given, involving a relation between distance to the submanifold and curvature of this sub-manifold.

1. INTRODUCTION

Statistical manifolds are Riemannian manifolds with pairs of conjugate torsion free connections. We shall review their basic definitions and properties in section 2, give a general framework describing their appearance in section 3, show how they lead to geometric fitting procedures of quasi-likelihood type in section 4. In section 5 we shall indicate a geometric result about uniqueness of projections, thus giving an example of applying geometric insight to discuss features of a wide range of statistical estimation problems.

We shall assume that the reader is familiar with modern differential geometry on the level of Boothby (1975), as well as the first developments in geometric statistical theory, as described e.g. in Amari (1985). Also some familiarity with more recent developments such as those described in the IMS Monograph. Vol X, containing, among others, papers by Barndorff-Nielsen (1987) and Lauritzen (1987). The latter reference has an ultrashort summary of some basic differential geometric concepts, in principle sufficient as background for reading this paper.

The paper is not a comprehensive survey of the differential-geometric approach to statistics but directs its attention primarily towards the consequences and roles played by the particular aspect of the conjugate pair of connections, occurring persistently in many statistical problems. Thus the bibliography at the end is by no means complete.

2. STATISTICAL MANIFOLDS

We shall briefly recall some basic facts concerning statistical manifolds as introduced in Lauritzen (1987), see also Amari (1982,1985), Eguchi (1985).

We let M be a smooth manifold, equipped with a Riemannian metric g. If ∇ is an affine connection on the tangent space of M, we define its *conjugate* ∇^* connection as

$$g(\nabla^*_X Y,Z) = Xg(Y,Z) - g(\nabla_X Z,Y). \qquad (2.1)$$

We have $(\nabla^*)^* = \nabla$ and the parallel transports associated with ∇ and ∇^* satisfies

$$g(\pi X,\pi^* Y) = g(X,Y),$$

where π and π^* are parallel transports w.r.t. ∇ and ∇^* along the same curve γ.

If ∇ is torsion free, we further have

$$\nabla^* \text{ is torsion free} \Leftrightarrow \bar{\nabla} = \tfrac{1}{2}(\nabla + \nabla^*),$$

where $\bar{\nabla}$ is the Levi-Civita connection associated with the metric g, see Lauritzen (1987).

To each such pair of conjugate, torsion free connections, a *skewness tensor* D is associated as

$$D(X,Y,Z) = g(\nabla^*_X Y,Z) - g(\nabla_X Y,Z).$$

The skewness tensor is symmetric in all three arguments.

The curvature tensors associated with ∇ and ∇^* are denoted by R and R* and satisfy the relation

$$R(X,Y,Z,W) = -R^*(X,Y,W,Z).$$

If in fact $R=R^*$, we say that the space is *conjugate symmetric*. Flat spaces, i.e. spaces with $R \equiv R^* \equiv 0$ are clearly conjugate symmetric.

If N is a regular submanifold of M, we define its embedding curvatures H and H* as the bilinear maps from the tangent space of N to its g-orthogonal complement, satisfying

$$g(H(X,Y),Z) = g(\nabla_X Y,Z)$$

$$g(H^*(X,Y),Z) = g(\nabla^*_X Y,Z),$$

for X,Y being *tangent* to N and Z *orthogonal* to N.

If $H \equiv 0$ or $H^* \equiv 0$ we say that N is ∇-*geodesic* or ∇^*-*geodesic*. If it is both, we just say that it is geodesic. We say that N is ∇-*geodesically convex* if any two points in N can be joined with a ∇-geodesic that lies entirely within N, (and analogously with ∇^*). Clearly this is stronger than requiring $H \equiv 0$ or $H^* \equiv 0$.

Examples of statistical manifolds are now plenty. The first manifolds of this type, appearing in statistical theory, are the geometries of Amari (1982), but also the geometries associated with estimation based on *contrast functionals* lead to structures of the same geometrical nature, see Eguchi (1985). Similarly the "observed" geometries associated with "maximum estimation", discussed by Barndorff-Nielsen (1986,1987), can be conveniently described as statistical manifolds.

The geometric structure has previously been studied to some extent by Norden (1945,1948) and Sen (1944,1945,1946), but also many of the considerations in Amari (1985) and Eguchi (1985) are really about general structures.

An important advantage of the geometrization of statistical theory is that many results can be derived to hold for all these structures simultaneously. An example is the following: Let x be a point in M and let N be a regular submanifold. A point $y \in N$ is said to be the ∇-*projection* of x onto N if there is a ∇-geodesic from x to y that intersects N orthogonally at y. The ∇-projection plays the role as an estimator possibly maximizing a suitable objective function in the statistical geometries mentioned above.

As shown in Amari (1985), *if the space itself is flat and N is ∇^*-geodesically convex, the ∇-projection is unique.*

This result is the general geometric formulation of the well-known result about uniqueness of the maximum likelihood estimate in linear exponential families.

An extension of this result relating the ∇-curvature of the space, the ∇^*-embedding curvature of N and the distance of x to N has recently been obtained by Lauritzen and Picard (1987), extending results of Efron (1978) to a general geometric set up. A special case is described in section 5 of this paper.

3. INDUCING STATISTICAL MANIFOLDS

In the following we shall give a general way of constructing statistical manifolds that in special cases lead to those previously mentioned. We hope thereby to get closer to an understanding of these structures as they appear in statistical theory.

Let V and W be finite-dimensional vector spaces that are mutually dual via a bilinear duality

$$V \times W \to \mathbb{R} \qquad (v,w) \to \langle v,w \rangle.$$

The spaces V and W are differentiable manifolds with standard parallelisms ∇' and ∇'' corresponding to usual parallel shifts.

At each point $v \in V$, we can identify the tangent space $T(V)_v$ with V itself, and similarly with W. The duality between V and W then induces a duality between the tangent spaces.

The spaces V and W are to be conceived of as the set of *functions* on a measure space and the set of *measures* on that space and the basic duality as

$$\langle f,\mu \rangle = \int f \, d\mu.$$

For the considerations to be exact the measure space in question has to be finite, but we shall sometimes ignore this and jump to conclusions about the general case by analogy.

Suppose that we have yet another differentiable manifold Ω, typically of much lower dimension than V and W, to play the role of a statisticians *parameter space*.

We now let λ and π denote embeddings of this Ω into V and W:

$$\lambda : \Omega \to V, \quad \pi : \Omega \to W,$$

in the 'classical' statistical context as discussed in Amari (1982), these are

$$\omega \xrightarrow{\lambda} \log f(\cdot,\omega); \quad \omega \xrightarrow{\pi} P_\omega,$$

where f is the density of P_ω w.r.t. a fixed measure μ. More 'refined' examples are in Eguchi (1985).

From any such pair of embeddings we can define a covariant tensor on the tangent space of Ω as

$$g(X,Y) = \langle \lambda_*(X), \pi_*(Y) \rangle \tag{3.1}$$

where λ_* and π_* are the *derivatives* of the maps λ and π and $\langle , \rangle$ is the duality on the tangent spaces of V and W.

In general g need not be a Riemannian metric, but in the classical case above, g is the Fisher-Rao information metric and in other cases we also get true metrics, see below.

But let us just first assume the bilinear g, to be *non-degenerate*, i.e.

$$g(X_\omega, Y_\omega)_\omega = 0 \text{ for all } Y_\omega \in T(\Omega)_\omega$$
$$\Rightarrow X_\omega = 0.$$

Then the tangent spaces $T(V)_{\lambda(\omega)}$ and $T(W)_{\pi(\omega)}$ can be decomposed as

$$V = T(V)_{\lambda(\omega)} = \text{im } (\lambda_*)_{\lambda(\omega)} \oplus [\text{im}(\pi_*)_{\pi(\omega)}]^\circ$$
$$W = T(W)_{\pi(\omega)} = [\text{im}(\lambda_*)_{\lambda(\omega)}]^\circ \oplus \text{im } (\pi_*)_{\pi(\omega)}.$$

Where "im" is the range space of a linear map and the *annihilator* A° of a subspace $A \subseteq V$ is the subspace of W given as

$$A^\circ = \{w \in W \mid \langle a,w \rangle = 0 \text{ for all } a \in A\}.$$

The parallelisms ∇' and ∇'' on V and W can therefore now be pulled back to Ω yielding parallelisms ∇ and ∇^* given by lifting, differentiating and projecting along annihilators to obtain

$$<\lambda_*(\nabla_X Y),\ \pi_*(Z)> = <\nabla'_{\lambda_*(X)}\lambda_*(Y),\ \pi_*(Z)>$$

$$<\lambda_*(X),\ \pi_*(\nabla^*_Y Z)> = <\lambda_*(X), \nabla''_{\pi_*(Y)}\pi_*(Z)>,$$

Recalling the expression for g (3.1), we see that ∇ and ∇^* connections are *conjugate* w.r.t. g, in the sense that

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z).$$

Which follows from the fact that ∇' and ∇'' are mutually conjugate (being trivial).

So if just g happens to be symmetric and positive, this construction, being very general, leads to a statistical manifold.

In all examples mentioned (Amari, Barndorff-Nielsen, Eguchi), λ and π are related to each other through a maximizing property:

$$<\lambda(\eta),\ \pi(\omega)> \leq <\lambda(\omega),\ \pi(\omega)> \qquad (3.2)$$

with equality if and only if $\eta=\omega$. This relation is, in the classical case known as "*the information inequality*".

Suppose now that η and ω are expressed in local coordinates and let

$$\lambda_i(\omega) = \frac{\delta}{\delta\omega^i}\,\lambda(\eta)\,/_{\eta=\omega}$$

$$\lambda_{ij}(\omega) = \frac{\delta^2}{\delta\omega^i\delta\omega^j}\,\lambda(\eta)\,/_{\eta=\omega}$$

and similarly with π. We would from (3.2) then in wide generality have

$$<\lambda_i(\omega),\ \pi(\omega)> = 0 \qquad (3.3)$$

and

$$\{<\lambda_{ij}(\omega),\ \pi(\omega)>\} \text{ is negative definite.} \qquad (3.4)$$

Differentiating (3.3) yields

$$<\lambda_{ij}(\omega),\ \pi(\omega)> + <\lambda_i(\omega),\ \pi_j(\omega)> = 0,$$

whereby from (3.4) we get

$$\{<\lambda_i(\omega),\ \pi_j(\omega)>\} \text{ is symmetric and positive definite.}$$

Realizing now that

$$\lambda_i(\omega) = \lambda_*(\frac{\delta}{\delta\eta^i})_\omega,\ \pi_j(\omega) = \pi_*(\frac{\delta}{\delta\eta^i})_\omega,$$

we see that *the fundamental relation (3.2), combined with assumptions of smoothness and regularity, eventually leads to the positive definiteness of* g .

We have described a way of inducing a statistical manifold through embeddings of a parameter space into vector spaces, satisfying a certain inequality. Are all statistical manifolds of this type? We do not know.

Eguchi (1985) gives a very wide range of examples of well known *contrast functionals* from statistics, all satisfying (3.2).

We shall conclude this section by showing how the observed geometries of Barndorff-Nielsen (1986,1987) fit into this framework.

The embedding λ is chosen to be a certain function to be maximized, the primary example being the likelihood function

$$\lambda(\omega) = \log f(.,\omega),$$

where f is a density of the unknown measure P_ω w.r.t. a fixed

measure μ.

An essential part of Barndorff-Nielsen's construction is the (local) identification of the *sample space* with the parameter space by choosing an *auxiliary statistic* a = a(t) such that the minimal sufficient statistitc t is in a 1-1 correspondence with a and $\hat{\omega}$, $\hat{\omega}$ being the maximum estimator i.e.

$$t \longleftrightarrow (\hat{\omega},a),$$

where $\hat{\omega}$ maximizes log f(t;ω) or a similar function. If t_0 is the observed value of t and we let $a_0 = a(t_0)$, we can now make the embedding

$$\pi_0(\omega) = \delta(\omega,a_0),$$

where δ denotes Dirac-measure. Thus this embedding depends heavily both upon the choice of the auxiliary function a and the particular *observed value*, a_0. The inequality (3.2) becomes

$$<\lambda(\eta),\ \pi_0(\omega)> = \log f((\omega,a_0);\eta) \leq \log f((\omega),a_0);\omega),$$

exactly expressing that ω is the maximum estimate if (ω,a_0) is observed. It is of interest to notice the remarkable fact that in the case of a full, regular exponential family the 'expected' geometries of Amari (1982) coincide with the observed geometries, a_0 being in these cases trivial.

4. QUASI LIKELIHOOD

In the previous sections we have seen how a range of statistical procedures from a geometric point of view can be studied within the same framework of a statistical manifold with a pair of mutually conjugate, torsion free connections.

Here we shall emphasize the possibility for a 'reverse' process: the geometric structures themselves can be used to provide procedures for data-fitting, that do not necessarily have a statistical counterpart, in a strict sense.

An example of this, which is well known in statistical literature is *quasi-likelihood* as introduced by Wedderburn (1974), with generalizations to the multivariate case attempted by McCullagh (1983).

The quasi-likelihood procedure concerns a data point $y \in \mathbb{R}^n$ to be fitted onto a smooth submanifold $N \subseteq \mathbb{R}^n$. This has sometimes a particular structure, which is not of our concern here, since this primarily relates to computational issues. The fitting is done by specifying an "inner product" g, to be thought of as an inverse covariance matrix and constructing the quasi-likelihood function as the solution to the partial differential equations

$$\frac{\partial \ell(y,\mu)}{\partial \mu^i} = g_{is}(\mu)(y^s-\mu^s), \tag{4.1}$$

where we have used the Einstein summation convention, implying that the right hand side has to be summed over s. The quantity g is assumed positive definite.

Provided this equation has a solution, the quasi-likelihood estimate is determined by maximizing the quasi-likelihood function. This value μ will in turn satisfy the equation

$$\frac{\partial \mu^i}{\partial \beta^\ell}\, g_{is}(\mu)\, (y^s - \mu^s) = 0, \qquad (4.2)$$

where $(\beta_1,\ldots,\beta_p)$ are local coordinates on N. Even if the equation (4.1) has a solution, it might not correspond to a proper likelihood function, see Jørgensen (1986,1987). Nevertheless, as we shall soon see, the procedure has a clear geometric interpretation within the context of statistical manifolds.

The basic manifold here is $M = \mathbb{R}^n$ and the parallelism ∇ that which makes straight lines in y geodesics. Thus the coefficients of this affine connection are equal to zero in the coordinate system determined by the data y.

The inner product g is nothing but a Riemannian metric on this manifold. Since the connection ∇ has coefficients equal to zero, we get for the conjugate connection that its coefficients satisfy

$$\Gamma^*_{ijk} = g(\nabla^*_{\partial_i}\, \partial_j, \partial_k) = \partial_i g_{jk} - \Gamma_{ijk} = \partial_i g_{jk},$$

where ∂_i denotes differentiation w.r.t. the i'th coordinate. Thus ∇^* is torsion free if, and only if, this satisfies

$$\partial_i g_{jk} \equiv \partial_j g_{ik}, \qquad (4.3)$$

a condition which is always fulfilled in the case where g is given by a diagonal matrix and $g_{ii}(\mu)$ depends on μ_i alone, this being the classical case discussed by Wedderburn (1972). The equation (4.2) expresses that μ is the ∇-projection of y onto N.

In general, *the condition (4.3) for the quasi-likelihood manifold to be of the type discussed is exactly that of* (4.1) *having a solution!*

The necessity of the condition (4.3) is seen by differentiating

(4.1) to obtain

$$\partial_i \partial_j \ell = \partial_j \partial_i \ell = \partial_j g_{is}(y^s - \mu^s) - g_{ij}$$

and the sufficiency is a bit more tricky (Frobenius' Theorem).

Summing up what we have learnt above, we see that the geometry of quasi-likelihood is *exactly* that of a flat statistical manifold with a conjugate pair ∇,∇^* and it follows immediately from the general result mentioned in section 2 that:

If N is ∇^*-geodesically convex, the quasi-likelihood estimate is unique, which was also obtained directly by Wedderburn (1974) in special cases. The existence of "canonical link-functions" etc. corresponds to dual coordinates as described by Amari (1985).

5. UNIQUENESS OF PROJECTIONS IN FLAT STATISTICAL MANIFOLDS

Consider a flat statistical manifold with a pair of conjugate connections ∇ and ∇^*, i.e. $R \equiv R^* \equiv 0$.

Let N be a regular submanifold of N and let $x \in M$ be fixed.

We shall here derive a recent result of Lauritzen and Picard (1987), relating the distance of x to N and the ∇^*-curvature of N to the problem of uniqueness of the ∇-projection of x onto N.

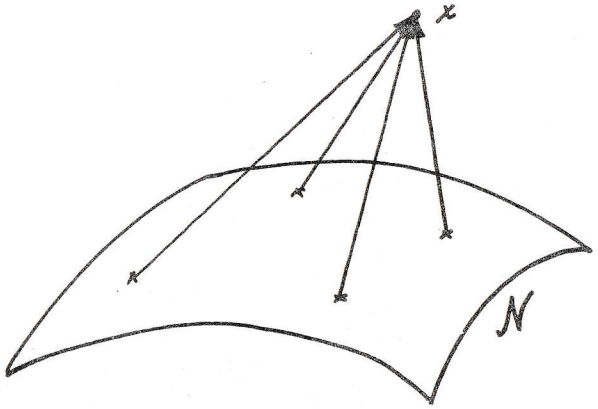
Recall that the projections are determined by ∇-geodesics from x, intersecting N in right angles.

We shall need some global properties of M and N and assume therefore that N is ∇-geodesically convex and that N is

${}_N\nabla^*$-geodesically convex, where ${}_N\nabla^*$ is the connection induced on N by ∇^* through g.

In more worldly terms, we assume that any two points in M can be joined by a ∇-geodesic and any two points on N by a ${}_N\nabla^*$-geodesic.

Since M is ∇-flat, a coordinate system exists where geodesics are straight lines. The assumption on M is then that M is convex. It this coordinate system, the picture below makes sense:



The arrows from points on N to the point $x \in M$ indicate both geodesics from points of N to x, but also a vector field X on N, determining these geodesic segments. A special feature of the vector field X is that for any vector field Y which is tangential to N, we have

$$\nabla_Y X = -Y \text{ on } N. \tag{5.1}$$

This follows from the fact that covariant and ordinary

differentiation coincide in this coordinate system, combined with the obvious interpretation of -X as "location vectors" of points on N in an affine *coordinate* system with x as origin. Further, $\|X\| = g(X,X)^{1/2}$ is the "distance" from x to N.

What we intend to show is that if either x is on "the right side" of N or it is close enough to N, the ∇-projection is unique. We therefore introduce the following measure of curvature of the submanifold N:

$$\gamma^*_x(y) = \max \{0, \sup_{\substack{\|E\| = \|F\|=1 \\ E \in T(N)^\perp_y,\ F \in T(N)_y \\ g(X,E) \geq 0}} g(E,H^*(F,F))\}$$

The scalar field γ^*_x is to be interpreted as the *maximal ∇^*-embedding curvature in the direction of* x. It is equal to zero if N curves away from x. Let Q_N be the orthogonal projection of X onto the normal plane $T(N)^\perp$.

We can now show the following:

THEOREM *If everywhere on N*

$$\|Q_N X\| < 1/\gamma^*_x, \tag{5.2}$$

then the ∇-projection of x onto N is unique.

PROOF:

The proof is indirect. Suppose y' and y" are both ∇-projections of x onto N. Then they can be joined by a ${}_N\nabla^*$-geodesic c(t), $t \in [0,1]$ with

$$y' = c(0),\ y'' = c(t).$$

If Y is the tangent to this curve, we thus have

$$\nabla^*_Y Y = H^*(Y,Y) \qquad (5.3)$$

Defining $f(t) = g(X_{c(t)}, Y_{c(t)})_{c(t)}$,

we have $f(0) = 0$, $f(1) = 0$ and therefore $f(t) = 0$ somewhere in the interval. But on the other hand we have

$$f'(t) = [Yg(X,Y)]_{c(t)}$$

and

$$Yg(X,Y) = g(\nabla_Y X,Y) + g(X,\nabla^*_Y Y)$$

$$= -\|Y\|^2 + g(X,H^*(Y,Y)),$$

where we have used (2.1), (5.1) and (5.3). From the definition of γ^*_X we now get

$$Yg(X,Y) \leq -\|Y\|^2 + \|Q_N X\| \cdot \|Y\|^2 . \gamma^*_X$$

$$< 0,$$

if the assumption of the theorem is fulfilled.

This establishes a contradiction and completes the proof. □

If the space M is not flat, (5.1) does not hold and the situation gets slightly more complicated.

Note that if $\gamma^*_X \equiv 0$, for example if N is ∇^*-geodesic the uniqueness follows, thus giving a previously mentioned result by Amari (1985) in a special case.

The condition (5.2) can be interpreted as a geometrically invariant condition for the quasi-likelihood function to be "concave". Usual concavity depends on the parametrisation chosen, whereas the condition (5.2) is geometrically invariant.

Thus it can be checked in *any* parametrization.

In the special case of expected geometries in exponential families, condition (5.2) reduces to something very similar to that given in Theorem 2 of Efron (1978).

REFERENCES

1. Amari, S. (1982). Differential geometry of curved exponential families - curvatures and information loss, **Ann. Statist. 10,** 357-382

2. Amari, S. (1985). **Differential - Geometrical Methods in Statistics.** Springer Lecture Notes in Statistics, No. 28, Springer, Berlin

3. Barndorff-Nielsen, O.E. (1986). Likelihood and observed geometries. **Ann. Statist. 14,** 856-873

4. Barndorff-Nielsen, O.E. (1987). Differential and integral geometry in statistical inference. In **Differential Geometry in Statistical Inference,** IMS Lecture Notes, X. Institute of Mathematical Statistics

5. Boothby, W.S. (1975). **An Introduction to Differentiable Manifolds and Riemannian Geometry.** Academic Press, New York

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

7. Eguchi, S. (1985). A differential geometric approach to statistical inference on the basis of contrast functionals. **Hiroshima Math. J. 15,** 341-391

8. Jørgensen, B. (1986). Some properties of exponential dispersion models. **Scand. J. Statist. 13,** 187-198

9. Jørgensen, B. (1987). Exponential dispersion models (with discussion). **J.R.Statist. Soc. B, 49,** 127-162

10. Lauritzen, S.L. (1987). Statistical manifolds. In **Differential Geometry in Statistical Inference.** IMS Lecture Notes, X. Institute of Mathematical Statistics

11. Lauritzen, S.L. and Picard, D. (1987). On the uniqueness of projections in statistical manifolds. Unpublished

12. McCullagh, P. (1983). Quasi-likelihood functions. **Ann Statist. 11,** 59-67

13. Norden, A. (1945). On pairs of conjugate parallel translations in n-dimensional spaces. **C.R. de l'Acad. de Sciences de l'URSS. Vol XLIX,** No. 9

14. Norden, A. (1948). On conjugate connections . (In Russian). **Trudy Seminara po Vektornomu i Tensornomu Analizu, Vol 8.** Moskva

15. Sen, R.N. (1944). On parallelism in Riemannian space. **Bull. Calcutta Math. Soc., 36,** 102-107

16. Sen, R.N. (1945). On parallelism in Riemannian space - II. **Bull. Calcutta Math. Soc. 37,** 153-159

17. Sen, R.N. (1946). On Parallelism in Riemannian space - III. **Bull. Calcutta Math. Soc. 38,** 161-167

18. Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. **Biometrika 61,** 439-447

19. Wedderburn,R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates for generalized linear models. **Biometrika 63,** 27-32

On Some Differential Geometric Concepts of Relevance to Statistics

O.E. Barndorff-Nielsen
Department of Theoretical Statistics
Institute of Mathematics
University of Aarhus
DK-8000 Aarhus C
Denmark

Abstract

The salient points of the recent theory of (derivative) strings are reviewed. As an application of strings, toward developping a 'geometric calculus' for statistical inference, a notion of invariant Taylor-like expansions is discussed and exemplified, one particular use being to the construction of exponential family approximations to an arbitrary parametric model.

1. Introduction

The present paper discusses some recent developments in the interactive field between statistics and differential geometry with which the author has been actively involved. Many interesting topics in this field will not even be touched upon here but a large fraction of these will undoubtedly be treated in other contributions to this conference. The emphasis will be on the mathematical aspects but some leads to the related statistical literature will be given. Further, a number of key papers not referred to in the text are included among the references.

The paper is mainly concerned with the theory of strings* which in the statistical context arose out of a seminal idea in a paper by McCullagh and Cox (1986), that paper being concerned with an invariant decomposition of a Bartlett adjustment of the likelihood ratio statistic. However, the idea of strings had, in fact, been introduced much earlier from a physical perspective, by Foster (1958, 1961, 1986), see also Foster (1987).

After a few preliminaries have been presented in section 2, we review the most important properties of strings in section 3, which also contains a number of examples. A main area of application of differential geometry to statistics is to the handling and interpretation of asymptotic expansions in statistical inference. For this it is of interest to develop a 'geometrical calculus' that is suited to the statistical context and which directly yields interpretable terms. A proposal along this line is discussed and exemplified in section 4. In particular, it provides invariant Taylor-like expansions. Finally, in section 5, it is noted how the same idea yields a method of approximating an arbitrary parametric model by an exponential model, in a parametrization invariant manner.

*The concept of strings discussed in this paper is quite different from several other similarly termed concepts which are of great current interest in physics and astronomy, see Campbell (1986). To avoid confusion we sometimes speak of the present concept

2. Preliminaries

Let M be a differentiable manifold of dimension d, let $\tilde{M}$ denote a copy of M, and let $\omega = (\omega^1,\ldots,\omega^d)$ and $\psi = (\psi^1,\ldots,\psi^d)$ indicate two alternative coordinate systems or <u>parametrizations</u> of M. For simplicity in exposition we mostly treat these, and other, coordinate systems as if they were global, rather than local, but the results to be discussed hold in fact without this restriction. We let generic coordinates of ω and ψ be indicated by ω^i, ω^j, ω^k,... and ψ^a, ψ^b, ψ^c,..., respectively. We shall freely refer to a point in M as ω or ψ or p, the latter being the coordinate-free notation. We write $\partial_i = \partial/\partial\omega^i$, $\partial_a = \partial/\partial\psi^a$, etc., and we think of these as elements of the tangent space TM of M. The analogous quantities on the copy $\tilde{M}$ of M are indicated by a tilde; thus, for instance, we write $\tilde{\omega}$, $\tilde{\psi}$ and $\tilde{p}$ for points in $\tilde{M}$ and $\tilde{\partial}_i$ for $\partial/\partial\tilde{\omega}^i$. Further, writing I_n for an index set $i_1\ldots i_n$, we introduce the notational conventions

$$f_{/I_n} = f_{/i_1\ldots i_n} = \partial_{i_1}\ldots\partial_{i_n} f \qquad (2.1)$$

for any $f\epsilon C^\infty M$, the space of infinitely differentiable functions on M, and

$$\omega^{I_n}_{/A_m} = \omega^{i_1\ldots i_n}_{/a_1\ldots a_m} = \sum_{A_m/n} \omega^{i_1}_{/A_{m1}}\ldots\omega^{i_n}_{/A_{mn}} \qquad (2.2)$$

where $A_m = a_1\ldots a_m$ and the summation is over all partitions of A_m into n blocks $A_{m1}, \ldots, A_{mn}$ such that the order of the indices in each of these blocks is the same as their order in A and such that for $p = 1,\ldots,n-1$ the first index in A_{mp} comes before the first index in A_{mp+1} as compared with the ordering within A_m. This is for $n \leq m$, and for $n > m$ both sides of (2.2) are interpreted as 0. Note that, as a particular case, we have

$$\omega^{i_1\ldots i_n}_{/a_1\ldots a_n} = \omega^{i_1}_{/a_1}\ldots\omega^{i_n}_{/a_n} .$$

We adopt the Einstein summation convention.

Let $X_1, X_2, \ldots, X_t$ be vector fields on M, write $\underline{t}$ for $\{1,2,\ldots,t\}$ and define an operator $X_{\underline{t}}: C^\infty M \to C^\infty M$ by

$$X_{\underline{t}} f = X_t(\ldots X_2(X_1 f)\ldots) . \qquad (2.3)$$

In ω coordinates the vector field X_1 has a representation of the form $X_1 = x_1^1\partial_1$, and letting

$$X^{l}_{\underline{t}} = X_t(\ldots X_2(x^{l}_1)\ldots) \qquad (2.4)$$

we have $X_{\underline{t}} = X^{l}_{\underline{t}}\partial_1$. Finally, for any $u = 1,\ldots,t$ let

$$X^{L_u}_{\underline{t}} = \sum_{\underline{t}/u} X^{l_1}_{\underline{t}_1}\ldots X^{l_u}_{\underline{t}_u} ; \qquad (2.5)$$

the summation is over all ordered partitions of $\underline{t}$ into u blocks $\underline{t}_1,\ldots,\underline{t}_u$, in complete analogy with the definition of (2.2). Also, for $u > t$ we interprete (2.5) as 0.

3. Strings

A derivative string, or string for brevity, is a collection $\bar{H}$ of arrays

$$H^{I_r L_u}_{J_s K_t} , \qquad t=1,\ldots,T; \quad u=1,\ldots,U , \qquad (3.1)$$

satisfying the transformation law

$$H^{A_r D_u}_{B_s C_t} = \sum_{\tau=1}^{t} \sum_{\upsilon=u}^{U} H^{I_r L_\upsilon}_{J_s K_\tau} \omega^{K_\tau}_{/C_t} \psi^{D_u}_{/L_\upsilon} \psi^{A_r}_{/I_r} \omega^{J_s}_{/B_s} \qquad (3.2)$$

for $t = 1,\ldots,T$ and $u = 1,\ldots,U$. The derivative string $\bar{H}$ is said to be of tensorial degree (r,s) and of length (T,U), and we denote the class of all such strings on M by $\mathcal{S}^{rU}_{sT}$ or $\mathcal{S}^{rU}_{sT}M$. We do, in fact, allow that some of the numbers r, s, T and U are 0, in which case the relevant groups of indices do not occur in (3.1) and (3.2). In particular, if $U = 0$ and $T > 0$ we speak of (3.1) as a costring (or a (r,s) costring), while (3.1) is called a contrastring provided $U > 0$ and $T = 0$. We refer to the indices $i_1\ldots i_r$ and $j_1\ldots j_s$ of (3.1) as tensorial indices and to $k_1\ldots k_t$ and $l_1\ldots l_u$ as structural indices. We let $\mathcal{T}^r_s = \mathcal{S}^{r0}_{s0}$, i.e. $\mathcal{T}^r_s$ is the class of (r,s) tensors on M.

The transformation law (3.2) generalizes those for tensors, affine connections, and derivatives of scalars. Accordingly, members of the particular classes $\mathcal{S}^{10}_{0\infty}$ and $\mathcal{S}^{00}_{0\infty}$ are referred to, respectively, as connection strings and scalar strings.

Examples 3.1-3.3 below are taken from Barndorff-Nielsen, Blæsild and Mora (1987), to which we refer for details.

Example 3.1. Let I_r be a fixed set of indices and define an array sequence $\bar{\omega}^{I_r}$ by

$$(\omega^{I_r})_{C_t}(\psi) = \omega^{I_r}_{/C_t}, \quad t=1,2,\ldots . \qquad (3.3)$$

Then $\bar{\omega}^{I_r}$ is an element of $\mathcal{S}^{00}_{0\infty}$, i.e. $\bar{\omega}^{I_r}$ is a $(0,0)$ costring of infinite length.

Similarly, let J_t be a fixed set of indices and define an array sequence $/\bar{J}_t$ by

$$(/J_t)^{D_\tau}(\psi) = \psi^{D_\tau}_{/J_t}, \quad \tau=1,2,\ldots . \qquad (3.4)$$

Then $/\bar{J}_t$ is a $(0,0)$ contrastring of infinite length. □

Example 3.2. The Kronecker deltas $\delta^{L_u}_{K_t}$ constitute a $(0,0)$ string of length (∞,∞). □

Example 3.3. For any positive integer t and any $X_{\underline{t}} \in \mathrm{Sec}(TM^{\times t})$, where $X_{\underline{t}}$ is considered as the operator defined in section 2, the set of arrays

$$\{X^{L_u}_{\underline{t}} : u=1,\ldots,t\}$$

constitutes a $(0,0)$ contrasting of length t. □

Example 3.4. Let $\Gamma^i_{k_1k_2}$ be the Christoffel symbols of an affine connection ∇ on M.

The arrays $\delta^i_{k_1}$ and $\Gamma^i_{k_1k_2}$ together constitute a connection string of length 2 and this may be prolonged to a connection string $\bar{\Gamma} = \{\Gamma^i_{K_t} : t=1,2,\ldots\}$ of infinite length, i.e. an element of $\mathcal{S}^{10}_{0\infty}$, by defining recursively

$$\Gamma^i_{K_{t+1}} = \Gamma^i_{K_t//k_{t+1}}, \qquad t=2,3,\ldots , \qquad (3.5)$$

where $//k$ indicates covariant differentiation by ∇ and with respect to ω^k, in the sense defined in Barndorff-Nielsen and Blæsild (1987a). (Note that (3.5) is, in fact, true also for $t = 1$.) This string is termed <u>the canonical string generated by $\Gamma^i_{k_1k_2}$ or ∇</u>. These arrays $\Gamma^i_{K_t}$ are 'generalised Christoffel symbols' corresponding to repeated covariant differentiation by ∇ in the sense that

$$\nabla_{\partial_{k_t}}\nabla_{\partial_{k_{t-1}}}\cdots\nabla_{\partial_{k_2}}\partial_{k_1} = \Gamma^i_{k_1\ldots k_t}\partial_i . \qquad (3.6)$$

□

The following example 3.5 builds on the concept of <u>yokes</u>. For the defintion, notation and properties of yokes the reader is referred to Barndorff-Nielsen

(1987b) and Blæsild (1987). Note, in particular, the definition of mixed derivatives $g_{I_r;J_s}$ of a yoke g, as stated in Blæsild (1987) (which is included in this volume). Of special interest in statistics are the two instances of a yoke defined in terms of the log likelihood function l by

$$g(\omega;\tilde{\omega}) = E_{\tilde{\omega}}\{(l(\omega) - l(\tilde{\omega}))\}$$

$$= \int (l(\omega;x) - l(\tilde{\omega};x))p(x;\tilde{\omega})d\mu \qquad (3.7)$$

and

$$g(\omega;\tilde{\omega}) = l(\omega;\tilde{\omega},a) - l(\tilde{\omega};\tilde{\omega},a) . \qquad (3.8)$$

The geometric structures on M induced by (3.7) and (3.8) are called, respectively, <u>expected geometries</u> and <u>observed geometries</u>. In particular, the metric tensors determined from (3.7) and (3.8) as $g_{i;j}$ are the <u>expected information</u> and the <u>observed information</u>, respectively.

<u>Example 3.5.</u> Let the differentiable manifold M be equipped with a yoke g.

The sequence of arrays $g_{k_1\ldots k_t}$, $t = 1,2,\ldots$, constitutes a costring of tensorial degree $(0,0)$.

Furthermore, for any real α a costring of

tensorial degree $(0,1)$ is given by

$$g^{\alpha}_{jk_1\ldots k_t} = \frac{1+\alpha}{2} g_{k_1\ldots k_t;j} + \frac{1-\alpha}{2} g_{j;k_1\ldots k_t} .$$

Lifting the index j by means of $g^{i;j}$, the inverse matrix of $g_{i;j}$, this is transformed to the connection string

$$g^{\alpha i}_{k_1\ldots k_t} = \frac{1+\alpha}{2} g^{;i}_{k_1\ldots k_t} + \frac{1-\alpha}{2} g^{i}_{;k_1\ldots k_t} . \qquad (3.9)$$

In particular, for g as defined in (3.7) or (3.8) one obtains, respectively, the <u>expected</u> and the <u>observed</u> <u>α-connection strings</u>. □

Based on any connection string $\bar{\Gamma}$ with $\Gamma^i_j = \delta^i_j$ one can for every $r = 1,2,\ldots$ define a $(r,0)$ costring $\{\Gamma^{I_r}_{K_t} : t=1,2,\ldots\}$ by

$$\Gamma^{I_r}_{K_t} = \sum_{K_t/r} \Gamma^{i_1}_{K_{t1}} \ldots \Gamma^{i_r}_{K_{tr}} . \qquad (3.10)$$

Furthermore, for every $s = 1,2,\ldots$ let

$$G^{l}_{J_s} = \sum_{\pi=1}^{s-1} (-1)^{\pi} \Sigma^{*} \Gamma^{l}_{J_{(1)}} \Gamma^{J_{(1)}}_{J_{(2)}} \ldots \Gamma^{J_{(\pi-2)}}_{J_{(\pi-1)}} \Gamma^{J_{(\pi-1)}}_{J_s} , \qquad (3.11)$$

where Σ^* indicates summation over all cases for which the index sets $J_{(1)},\ldots,J_{(\pi-1)}$ satisfy

$$1<|J_{(1)}|<\ldots<|J_{(\pi-1)}|<s \ .$$

In case $s = 1$ we interprete (3.11) as $G^{l}_{j_1} = \delta^{l}_{j_1}$.

We may now define arrays $G^{L_u}_{J_s}$ by

$$G^{L_u}_{J_s} = \sum_{J_s/u} G^{l_1}_{J_{s1}} \ldots G^{l_u}_{J_{su}} \ , \qquad (3.12)$$

and we then have that $\{G^{L_u}_{J_s}: u=1,\ldots,s\}$ is a $(0,s)$ contrastring of length $(0,s)$. We refer to (3.11) and (3.12) as <u>the costrings and constrastrings generated by the connection string $\bar{\Gamma}$</u>. These are dual in the sense that

$$\sum_{\tau=u}^{t} G^{L_u}_{I_\tau} \Gamma^{I_\tau}_{K_t} = \delta^{L_u}_{K_t}$$

and

$$\sum_{t=r}^{s} G^{K_t}_{J_s} \Gamma^{I_r}_{K_t} = \delta^{I_r}_{J_s} \ .$$

Further examples of derivative strings are discussed in Barndorff-Nielsen and Blæsild (1987a,b) and Barndorff-Nielsen, Blæsild and Mora (1987).

The concept of derivative strings has been developed and studied (Barndorff-Nielsen (1986b), Barndorff-Nielsen and Blæsild (1987a,b,c), Barndorff-Nielsen, Blæsild and Mora (1987), Murray and Rice (1987), Murray (1987)) both for its potential usefulness in statistics and for its purely mathematical interest. Its root is in a particular construction of symmetric tensorial derivatives of log likelihood functions introduced by McCullagh and Cox (1986). Other statistical applications are discussed in the first two of the above-mentioned papers. However, the gist of this concept has been suggested independently, and with a view to its physical relevance, by Foster (1958, 1961, 1986), see also Foster (1987).

Various elementary algebraic operations on strings are discussed in Barndorff-Nielsen and Blæsild (1987a,b). As noted there, for any (r,s) and (T,U) the string class $\mathcal{S}^{rU}_{sT}$ is a real vector space. More interestingly, strings may be multiplied together to yield new strings by an operation termed <u>convolutive multiplication</u> or <u>*-multiplication</u>. Thus if $\bar{H} \in \mathcal{S}^{r0}_{s\infty}$ and $\bar{H}' \in \mathcal{S}^{r'0}_{s'\infty}$, the product $\bar{H} * \bar{H}'$ is an element of $\mathcal{S}^{r+s'0}_{s+s'\infty}$ given by

$$(\bar{H} * \bar{H}')^{I_r I'_{r'}}_{J_s J'_{s'} K_t} = \begin{cases} 0 & t=1 \\ \sum_{K_t/2} H^{I_r}_{J_s K_{t1}} H'^{I'_{r'}}_{J'_{s'} K_{t2}} & t=2,\ldots \end{cases}$$

Note that this product is noncommutative, and that the costring (3.10), generated by the connection string $\bar{\Gamma}$, may be considered as obtained by repeated convolutive multiplication of $\bar{\Gamma}$ with itself. Compare also (3.12).

Among the further operations that can be performed with strings those of intertwining and covariant differentiation are of some particular interest. In particular, intertwining provides a method for construction of tensors from strings, and covariant differentiation of string elements can be used to generate or prolong strings. Concerning the latter, the reader is referred to Barndorff-Nielsen and Blæsild (1987b), particularly section 8. Intertwining is discussed in subsection 3.1 below.

3.1. Intertwining

We shall consider certain sequences $\bar{N}$ of tensors and we therefore introduce the notation

$$\bar{\mathcal{T}}^{rU}_{sT} = \{\bar{N}=(N^{I_r I'_u}_{J_s J'_t}: t=1,\ldots,T,\ u=1,\ldots,U\}:\ N^{I_r I'_u}_{J_s J'_t} \in \mathcal{T}^{r+u}_{s+t}\}.$$

Theorem 3.1. Let $\bar{\Gamma} \in \mathcal{P}^{10}_{0\infty}$ with $\Gamma^i_j = \delta^i_j$ and let $(\Gamma^{I_r}_{K_t}: t=1,2,\ldots)$ and $(G^{L_u}_{J_s}: u=1,2,\ldots)$ be, respectively, the $(r,0)$ costring and the $(0,s)$ contrastring generated by $\bar{\Gamma}$. Furthermore, let $\bar{H}$ and $\bar{N}$ denote sequences of arrays.

Then

$$H^{I_r L_u}_{J_s K_t} = \sum_{\tau=1}^{t} \sum_{\upsilon=u}^{U} N^{I_r J'_\upsilon}_{J_s I'_\tau} \Gamma^{I'_\tau}_{K_t} G^{L_u}_{J'_\upsilon} \qquad (3.13)$$

for $t = 1,2,\ldots$ and $u = 1,2,\ldots,U$ if and only if

$$N^{I_r J'_\upsilon}_{J_s I'_\tau} = \sum_{t=1}^{\tau} \sum_{u=\upsilon}^{U} H^{I_r L_u}_{J_s K_t} \Gamma^{J'_\upsilon}_{L_u} G^{K_t}_{I'_\tau}\,. \qquad (3.14)$$

Moreover, provided $\bar{H}$ and $\bar{N}$ are related by (3.13) and (3.14), we have

$$\bar{N} \in \bar{\mathcal{T}}^{rU}_{s\infty} \Leftrightarrow \bar{H} \in \mathcal{P}^{rU}_{s\infty}\,. \qquad (3.15)$$

□

Proof. See Barndorff-Nielsen and Blæsild (1987b). □

In case $\bar{H}$ is a string we refer to the arrays (3.14) as the tensorial components of $\bar{H}$ with respect to $\bar{\Gamma}$.

According to this theorem all strings can be generated from a single connection string and the collection of all tensors on M.

We write the relations (3.13) and (3.14) succinctly as $\bar{H} = \bar{N} \square \bar{\Gamma}$ and $\bar{N} = \bar{H} \wedge \bar{\Gamma}$, respectively. There are rules of calculation such as

$$\bar{\Phi} = \bar{H} \wedge \bar{\Gamma},\ \bar{\Phi}' = \bar{H}' \wedge \bar{\Gamma} \Rightarrow \bar{\Phi} * \bar{\Phi}' = (\bar{H} * \bar{H}') \wedge \bar{\Gamma} ,$$

cf. Barndorff-Nielsen and Blæsild (1987a,b).

As a special instance of theorem 3.1, if $\bar{F}$ is a scalar string and $\bar{\Gamma}$ is a connection string, i.e. $\bar{F} \in \mathcal{S}^{00}_{0\infty}$ and $\bar{\Gamma} \in \mathcal{S}^{10}_{0\infty}$, then $\bar{\Phi} = \bar{F} \wedge \bar{\Gamma}$ is a sequence of covariant tensors. Assuming for simplicity that both $\bar{F}$ and $\bar{\Gamma}$ are symmetric, the first few of these tensors may be written

$$\Phi_{j_1} = F_{j_1}$$

$$\Phi_{j_1 j_2} = F_{j_1 j_2} - \Gamma^{i}_{j_1 j_2} F_i$$

$$\Phi_{j_1 j_2 j_3} = F_{j_1 j_2 j_3} - \Gamma^{i}_{j_1 j_2} F_{i j_3}[3] - (\Gamma^{i}_{j_1 j_2 j_3} - \Gamma^{i}_{j_1 k} \Gamma^{k}_{j_2 j_3}[3]) F_i .$$

In particular, when $\bar{F}$ consists of the successive derivatives of a scalar f, i.e. $\bar{F} = (f_{/K_t} : t=1,2,\ldots)$, we denote the elements of $\bar{F} \wedge \bar{\Gamma}$ by $f_{\Gamma K_t}$. (Thus, for instance, $f_{\Gamma k_1 k_2} = f_{/k_1 k_2} - \Gamma^{i}_{k_1 k_2} f_{/i}$.)

Example 3.6. In the statistical context, suppose $\bar{F}$ is the sequence of log likelihood derivatives, i.e. $F_{K_t} = l_{/K_t}$. The tensorial components of this with respect to the connection string $\overset{1}{\not{\partial}}$ derived, by (3.9), from the expected yoke (3.7) are the 'Möbius derivatives' of the log likelihood function, introduced by McCullagh and Cox (1986). □

3.2. Higher order differentiation

A coordinate-independent definition of derivative strings was established in Barndorff-Nielsen and Blæsild (1987c) under the assumption of structural symmetry i.e. invariance under permutation of the structural indices. See also Murray and Rice (1987) and Murray (1987). An

alternative approach is taken in Barndorff-Nielsen, Blæsild and Mora (1987), resulting in an axiomatic and coordinate-free treatment of strings, without the condition of structural symmetry. The results of this latter paper will now be summarized.

The paper introduces several concepts of higher order differentiation, the objects differentiated being functions or jet fields, vector fields and covector fields. Within this framework the higher order derivatives are not, in general, obtained by successive application of first order differental operators. Rather, for each of the objects concerned (functions, vector fields, etc.) we define, axiomatically and recursively, a sequence of operators $D^1,\ldots,D^n,\ldots$, say, such that $D^1,\ldots,D^n$ together determine the derivatives of order less than or equal to n.

We begin by discussing differentiation of jet fields or, equivalently, of sections of jet bundles. On account of the relation between jet fields and C^∞-functions, this concept implies one of differentiation of functions.

Let M be an arbitrary differentiable manifold and let D^n, $n = 1,2,\ldots$, denote a sequence of

$$D^n\colon \{0,1\}^n \times TM^{\times n} \times \mathrm{Sec}(J^n_0 M) \to C^\infty M$$
$$(\epsilon_{\underline{n}}, X_{\underline{n}}, s) \to D^n_{(\epsilon_{\underline{n}}, X_{\underline{n}})} s \,. \qquad (3.16)$$

Here $J^n_0 M$ denotes the zero-truncated jet space of order n on M, and $\epsilon_{\underline{n}} = (\epsilon_1,\ldots,\epsilon_n)$ is an arbitrary sequence of 0-s and 1-s. We say that

α and X_α are $\left\{\begin{matrix}\text{structural}\\ \text{tensorial}\end{matrix}\right\}$ according as ϵ_α is $\left\{\begin{matrix}1\\ 0\end{matrix}\right.$

($\alpha = 1,\ldots,n$). For simplicity we shall suppress $\epsilon_{\underline{n}}$ from the notation, thus writing $D^n_{X_{\underline{n}}}$ rather than $D^n_{(\epsilon_{\underline{n}}, X_{\underline{n}})}$, a partition of $X_{\underline{n}} = (X_1,\ldots,X_n)$ into structural and tensorial elements being understood.

Definition 3.1. A sequence of mappings D^n, $n = 1,2,\ldots$, of the type (3.16) is called a differentiation string if the following five axioms are satisfied for every $X_{\underline{n}} \in TM^{\times n}$, $s,s' \in \mathrm{Sec}(J^n_0 M)$ and $f \in C^\infty M$ $(n=1,2,\ldots)$:

$$D^n_{X_{\underline{n}}}(s+s') = D^n_{X_{\underline{n}}} s + D^n_{X_{\underline{n}}} s' \qquad (3.17)$$

$$D^n_{(X_1,\ldots,X_\alpha+X'_\alpha,\ldots,X_n)}{}^s$$

$$=D^n_{(X_1,\ldots,X_\alpha,\ldots,X_n)}{}^s + D^n_{(X_1,\ldots,X'_\alpha,\ldots,X_n)}{}^s$$

$$\text{for } X'_\alpha \in TM, \quad \alpha=1,\ldots,n\ , \tag{3.18}$$

$$D^n_{X_{\underline{n}}}(s\cdot s') = \sum_{\underline{n}//2} D^{|\underline{n}_1|}_{X_{\underline{n}_1}} s\ D^{|\underline{n}_2|}_{X_{\underline{n}_2}} s'\ , \tag{3.19}$$

the summation being over all partitions of $\underline{n}$ into two blocks one of which may be empty,

$$D^n_{X_{\underline{n}}}(fs) = fD^n_{X_{\underline{n}}} s\ , \tag{3.20}$$

and

$$D^n_{(X_1,\ldots,fX_\alpha,\ldots,X_n)}{}^s = \begin{cases} f\ D^n_{X_{\underline{n}}} s \\ \\ \underset{\underline{\beta}\subset\bar{\beta}(\alpha)}{\Sigma} X_{\underline{\beta}}(f)\ D^{n-|\beta|}_{X_{\underline{n}\setminus\underline{\beta}}}(s) \end{cases} \tag{3.21}$$

where the upper (lower) equality applies for α tensorial (structural) and in the last expression $\bar{\beta}(\alpha)$ denotes the set of structural indices among $\alpha+1,\ldots,n$. □

That differentiation strings do, in fact, exist as well-defined mathematical objects follows from theorem 3.2 below which characterizes differentiation strings in terms of a certain type of derivative strings.

As a preparation for that theorem we introduce some notational conventions. On the assumption that a partition of $1,2,\ldots$ into tensorial and structural elements has been given we let $Y_1,\ldots,Y_s$ and $Z_1,\ldots,Z_t$, where $s+t = n$, denote the elements of $X_1,\ldots,X_n$ whose indices are of, respectively, structural and tensorial type, the orderings of the Y-s and the Z-s being the same as within $X_1,\ldots,X_n$. Since the partition of $1,2,\ldots$ into tensorial and structural elements is considered as prescribed we may now, without ambiguity, write $D^n_{(Y_{\underline{s}},Z_{\underline{t}})}$ instead of $D^n_{X_{\underline{n}}}$.

Let

$$Y_s = y^k_s\partial_k \quad \text{and} \quad Z_t = z^j_t\partial_j$$

be the representations of Y_s and Z_t relative to the generic coordinate system ω, and let $Y^{K_\sigma}_{\underline{s}}$ and $Z^{J_\tau}_{\underline{t}}$ be defined by (2.5).

Theorem 3.2. A sequence of operators D^n, $n=1,2,\ldots,N$,

$$D^n: \{0,1\}^n \times TM^{\times n} \times \mathrm{Sec}(J_0^n M) \to \mathfrak{C}^\infty M$$
$$(\epsilon_{\underline{n}}, X_{\underline{n}}, s) \to D^n_{(Y_{\underline{s}}, Z_{\underline{t}})} s \qquad (3.22)$$

constitutes a differentiation string of length N if and only if for each $n=1,\ldots,N$ the quantity $D^n_{(Y_{\underline{s}}, Z_{\underline{t}})} s$ is of the form

$$D^n_{(Y_{\underline{s}}, Z_{\underline{t}})} s = Z_{\underline{t}}^{J_t} \sum_{\nu=1}^{n} \sum_{\sigma=1}^{s} s_{L_\nu} Y_{\underline{s}}^{K_\sigma} H^{L_\nu}_{J_t K_\sigma} \qquad (3.23)$$

where $H^{L_\nu}_{J_t K_\sigma} = 0$ if $\nu > \sigma + t$ and where, for each n,

$$\{H^{L_\nu}_{J_t K_\sigma} : \nu \leq n, \quad \sigma \leq s\} \qquad (3.24)$$

is a decomposable derivative string of tensorial degree $(0,t)$ and length (s,n), with $H^1_j = \delta^1_j$ and $H^1_k = \delta^1_k$. □

A string $H^{L_\nu}_{J_t K_\sigma}$ is said to be decomposable if

$$H^{L_\nu}_{J_t K_\sigma} = \sum_{J_t K_\sigma / \nu} H^{l_1}_{J_{t1} K_{\sigma 1}} \ldots H^{L_\nu}_{J_{t\nu} K_{\sigma\nu}} .$$

If Y_s and Z_t for every s and t are elements of the field of coordinate frames corresponding to ω then $Y_{\underline{s}} = \partial K_s = (\partial_{k_1}, \ldots, \partial_{k_s})$ and $Z_{\underline{t}} = \partial J_t = (\partial_{j_1}, \ldots, \partial_{j_t})$, and (3.23) takes the form

$$D^n_{(\partial K_s, \partial J_t)} s = \sum_{\nu=1}^{n} s_{L_\nu} H^{L_\nu}_{J_t K_s} .$$

In particular, if all the elements of $1,2,\ldots$ are tensorial then we have

$$D^n_{\partial J_n} s = \sum_{\nu=1}^{n} s_{L_\nu} H^{L_\nu}_{J_n} , \qquad (3.25)$$

while for all elements structural

$$D^n_{\partial K_n} s = \sum_{\nu=1}^{n} s_{L_\nu} H^{L_\nu}_{K_n} . \qquad (3.26)$$

The rules (3.25) and (3.26) specialize respectively to covariant differentiation and ordinary differentiation of functions by substituting $f \in C^\infty M$

for $s \in \mathrm{Sec}(J^n_0 M)$ and taking respectively $H^{L_v}_{J_n} = G^{L_v}_{J_n}$ and $H^{L_v}_{K_n} = \delta^{L_v}_{K_n}$ where $G^{L_v}_{J_n}$ indicates the contrastrings generated by an arbitrary affine connection $\Gamma^l_{k_1 k_2}$.

Having defined differentiation strings for functions and jet sections it is possible in a similar fashion axiomatically to introduce higher order differentiation of vector fields and of covector fields. In particular, to any connection string corresponds a sequence of operators $\nabla, \ldots, \nabla^n, \ldots$ satisfying three axioms, which are generalizations of those determining the concept of affine connections, and vice versa. The reader is referred to Barndorff-Nielsen, Blæsild and Mora (1987) for a detailed discussion.

4. Expansions

The concepts and techniques of differential geometry are of particular interest in connection with the handling and interpretation of the terms in a variety of - convergent or asymptotic - expansions that occur in statistics. We shall not attempt to give an exhaustive disucssion of this, but we will

briefly mention a number of points and problems of some current interest.

The expansions we have in mind allow of geometric interpretations of terms or groups of terms in the expansions. Up til quite recently this was discovered in retrospect only, by insightful rewriting and rearrangement of the terms of expansions, these latter having been derived by analytic calculations that are not of an inherent geometric nature. Thus there is a need to develop a 'geometric calculus' by which the appropriate geometric form and meaning of the expansions will appear automatically, as it were, and not as the result of a more or less fortuitous combination of ingenuity and luck. A step in that direction was taken, in the special context of Bartlett adjustments, by McCullagh and Cox (1986). This led to the introduction and study of strings (cf. section 3) and was also instrumental in connection with a paper by Skovgaard (1986) who - as do McCullagh and Cox - derives a decomposition of the expected Bartlett adjustment into six invariant terms but using covariant differentiation instead of the 'Möbius derivatives' of the log likelihood function introduced by the latter authors. It seems, however, that for many statistical purposes the route by

strings offers some advantages in terms of generality and simplicity.

We now discuss some further steps towards a geometrical calculus of the kind indicated above.

4.1. Invariant Taylor-like expansions

Taylor expansions lie at the root of most of the asymptotic results that are of importance in statistical inference and it is therefore natural first to look for reformulations in invariant terms of the idea of Taylor expansion. From an abstract viewpoint the existence of such invariant expansions is rather obvious (in this connection, see Murray and Rice (1987) and Murray (1987)), but as far as actual calculations are concerned it is desirable to devise a reformulation which is coordinate-based and of an algorithmic nature.

One possibility (Barndorff-Nielsen (1987b)) is to proceed as follows. Choose a yoke $g = g(\omega;\omega)$ and take the associated connection string $\not{g}^{-1}$ defined by (3.9). Furthermore, let

$$g^k = \not{g}^{k;j} g_j \qquad (4.1)$$

and

$$g^{K_t} = g^{k_1} \ldots g^{k_t} ,$$

and, for any scalar f on M, let $f_{\Vert K_t}$ denote the t-th order tensorial component of f with respect to $\Gamma = \not{g}^{-1} = (\not{g}^{i}_{;K_t} : t=1,2,\ldots)$ (as defined in subsection 3.1). Then f may be expanded around $\omega \in M$ as

$$f(\tilde{\omega}) = f(\omega) + \sum_{t=1}^{\infty} \frac{1}{t!} f_{\Vert K_t}(\omega)\, g^{K_t}(\omega,\tilde{\omega}) \ . \qquad (4.2)$$

The quantity g^r behaves as a contravariant vector in ω and as a scalar in $\tilde{\omega}$. Hence, as desired, each of the terms on the right hand side of (4.2) is a scalar and is of the same order in $\tilde{\omega} - \omega$ as the corresponding term in the ordinary Taylor series

$$f(\tilde{\omega}) = f(\omega) + \sum_{t=1}^{\infty} \frac{1}{t!} f_{/K_t}(\omega)\, (\tilde{\omega}-\omega)^{K_t} \ . \qquad (4.3)$$

To derive (4.2) we first note that, since $\not{g}_{i;j}$ is by assumption nonsingular, one may introduce $(g^1,\ldots,g^d)$ as a set of local coordinates around ω, alternative to $(\tilde{\omega}^1,\ldots,\tilde{\omega}^d)$. Letting g^α, g^β, g^γ, ... denote generic elements of $(g^1,\ldots,g^d)$ we

find, on applying (4.3) under the parametrization $(g^1,\dots,g^d)$,

$$f(\tilde{\omega}) = f(\omega) + \sum_{t=1}^{\infty} \frac{1}{t!} f_{/\alpha_1\dots\alpha_t} g^{\alpha_1}\dots g^{\alpha_t} .$$

The validity of (4.2) now follows from

Lemma 4.1. In the coordinate system $(g^1,\dots,g^d)$ we have

$$\not{g}^{\alpha}_{;\gamma_1\dots\gamma_t} = 0 , \qquad t = 2,3,\dots ,$$

and, hence,

$$f_{/\gamma_1\dots\gamma_t} = f_{\shortparallel\gamma_1\dots\gamma_t} , \qquad t = 1,2,\dots . \quad \square$$

Proof. Straightforward. $\square$

While (4.2) is coordinate-independent, it does depend on the yoke g. However, in applications there will often be a particular yoke which it is natural and convenient to adopt, cf. subsection 4.2 below.

4.2. Expansions for likelihood quantities

As an illustration we consider the first terms of an asymptotic expansion for the distribution of the maximum likelihood estimator $\hat{\omega}$ given an ancillary statistic a. Under rather mild regularity conditions the conditional probability density function of $\hat{\omega}$ possesses an asymptotic expansion of the form

$$p(\hat{\omega};\omega|a) = \varphi_d(\hat{\omega}-\omega-\mu;\not{j}^{-1})\{1-\tfrac{1}{2}\Gamma^{-1/3}_{ijk}h^{ijk}(\hat{\omega}-\omega-\mu;\not{j}^{-1})+\dots\} \qquad (4.4)$$

(cf. Barndorff-Nielsen (1983,1986a)). Here φ_d denotes the probability density function of the d-dimensional normal distribution; more specifically,

$$\varphi_d(x;\Sigma) = (2\pi)^{-d/2}\,|\Sigma|^{-1/2}\,e^{-\frac{1}{2}x\,\Sigma^{-1}x^*} .$$

Furthermore,

$$\not{j}_{ij} = \chi_{i;j}$$

$$\mu^r = -\tfrac{1}{2}\chi_{j;kl}\not{j}^{ij}\not{j}^{kl} = -\tfrac{1}{2}\Gamma^{-1}_{klj}\not{j}^{ij}\not{j}^{kl}$$

$$\Gamma^{-1/3}_{ijk} = \tfrac{1}{3}\chi_{ij;k} + \tfrac{2}{3}\chi_{k;ij}$$

and

$$h^{ijk}(x;\Sigma) = x^i x^j x^k - x^i \Sigma^{jk}{}_{[3]}, \qquad (4.5)$$

i.e. h^{ijk} is a Cartesian tensorial Hermite polynomial of order 3.

The connections $\overset{-1}{\Gamma}$ and $\overset{-1/3}{\Gamma}$ above are those induced by (3.9) from the yoke (3.8). Thus these two connections occur naturally in the asymptotic study of the conditional distribution of the maximum likelihood estimator $\hat{\omega}$.

The next order term in the expansion indicated by (4.4) comprises a large number of terms and it is expressible as a weighted sum of tensorial Hermite polynomials, the weights being determined by second, third and fourth order mixed derivatives of the yoke (3.8) (cf. Barndorff-Nielsen 1986a). A full understanding of the structure of this term requires not only classical geometric concepts, such as the α-connections, but also derivative strings in the sense of section 3. Mora (1987) discusses this problem.

Formula (4.4) was obtained by Taylor expanding the p^*-model (Barndorff-Nielsen (1980, 1983)) for the conditional distribution of $\hat{\omega}$ given a:

$$p^*(\hat{\omega};\omega|a) = c(\omega,a)|\hat{j}|^{\frac{1}{2}} e^{l-\hat{l}} . \qquad (4.6)$$

As noted in Barndorff-Nielsen (1987c), this expression may be transformed, by the usual rule for transformation of probability density functions, to a formula for the conditional distribution of the score vector $l_* = [l_i]$, the result being

$$p^*(l_*;\omega|a) = c(\omega,a))|\hat{j}|^{\frac{1}{2}}|l_{;}|^{-1} e^{l-\hat{l}} \qquad (4.7)$$

where $l_{;}$ is the matrix $[l_{i;j}]$.

In wide generality, (4.6) and (4.7) are either exactly equal to or close approximations to the actual conditional distributions of $\hat{\omega}$ and l_*, respectively.

Under conditions similar to those underlying (4.4) one may derive an asymptotic expansion of (4.7):

$$p^*(l_*;\omega|a) = \varphi_d(l_*;j)\{1 + \tfrac{1}{6} T_{ijk} h^{ijk}(l^*;j^{-1}) + \ldots\}; \qquad (4.8)$$

here T_{ijk} is the (0,3) skewness tensor derived from the observed likelihood yoke (3.8) as

$$T_{ijk} = {}^{-1}\Gamma_{ijk} - {}^{1}\Gamma_{ijk} ,$$

and

$$l^* = l_* \chi ,$$

while $h^{ijk}(l^*;\hat{\jmath}^{-1})$ is determined by (4.5).

Note that for every $v = 1,2,\ldots$ the Hermite polynomial for l^*

$$h^{i_1\cdots i_v}(l^*;\hat{\jmath}^{-1})$$

is a $(v,0)$ tensor, due to l_i being a $(0,1)$ tensor. By contrast, the Hermite polynomials occurring in (4.4) are not tensors (they are Cartesian tensors, though).

Thus the term $T_{ijk}h^{ijk}(l^*;\hat{\jmath}^{-1})$ is invariant, and so are all the higher order terms in the expansion (4.8). (The form of the next order term is derived and discussed in Mora (1987).)

A convenient way to derive (4.8) is by means of the invariant Ttylor technique discussed in subsection 4.1, using the observed likelihood yoke (3.8). Applying that we find

$$l-\hat{l} = -\tfrac{1}{2} l^i l^j \chi_{i;j} + \tfrac{1}{6} l^i l^j l^k T_{ijk} + \ldots$$

and

$$|\hat{\jmath}|^{\frac{1}{2}}|l_;|^{-1} = |\hat{\jmath}|^{-\frac{1}{2}}\{1 - \tfrac{1}{2} l^i \hat{\jmath}^{jk} T_{ijk} + \ldots\}$$

from which (4.8) follows.

5. Approximating exponential models

An exponential model has model function of the form

$$p(x;\omega) = a(\omega)b(x)\ \exp\{\theta(\omega)\cdot s(x)\}$$

where $\theta(\omega)$ and $s(x)$ are vectors of dimension $k \geq d$.

In view of the many important properties of exponential models as tools for statistical analysis it is of some interest to explore the possibilities of approximating an arbitrary statistical model by an exponential model of some suitable order k.

The technique of invariant Taylor-like expansion discussed in the previous section provides one class of such possibilities. In fact, by (4.2) any

parametric model function $p(x;\omega)$ may be approximated locally around an arbitrary parameter point ω as

$$p(x;\tilde{\omega}) \doteq a(\omega,\tilde{\omega})p(x;\omega)\ \exp\Big(\sum_{t=1}^{T}\frac{1}{t!}\,g^{K_t}(\omega,\tilde{\omega})l_{\pi K_t}(\omega)\Big) \tag{5.1}$$

where $a(\omega,\tilde{\omega})$ is a norming constant which ensures that the right hand side of (5.1) integrates to 1, l is the log likelihood function and g is an arbitrary yoke. The approximation (5.1) constitutes an exponential model with parameter $\tilde{\omega}$, $\theta(\tilde{\omega}) = [g^{K_t}(\omega,\tilde{\omega})/t!]$ and $s(x) = [l_{\pi K_t}(\omega)]$, a vector consisting of linear combinations of log likelihood derivatives at ω.

Example 5.1. Taking $t = 1$ and g as given by (3.8) with $\hat{\omega}$ substituted by $\tilde{\omega}$, the approximation (5.1) becomes

$$p(x;\tilde{\omega}) \doteq a(\omega,\tilde{\omega})p(x;\omega)\ \exp\{l_{k'}(\omega,\tilde{\omega},a)\jmath^{kk'}(\omega)l_{k}(\omega;\hat{\omega},a)$$

where $\jmath^{kk'}(\omega)$ is the inverse matrix of $-l_{kk'}(\omega;\omega,a)$. □

Another approach was suggested by Amari (1987) and is developed in Barndorff-Nielsen and Jupp (1987).

References

Amari, S.-I. (1982a): Differential geometry of curved exponential families - curvatures and information loss. Ann. Statist. 10, 357-385.

Amari, S.-I. (1982b): Geometrical theory of asymptotic ancillarity and conditional inference. Biometrika 69, 1-17.

Amari, S.-I. (1985): Differential Geometric Methods in Statistics. Lecture Notes in Statistics 28. Springer-Verlag, Heidelberg.

Amari, S.-I. (1987): Differential geometrical theory of statistics - towards new developments. Differential Geometry in Statistical Inference, IMS Monograph. (To appear).

Amari, S.-I. and Kumon, M. (1983): Differential geometry of Edgeworth expansion in curved exponential family. Ann. Inst. Statist. Math 35, 1-24.

Barndorff-Nielsen, O.E. (1983): On a formula for the distribution of the maximum likelihood estimator. Biometrika 70, 343-365.

Barndorff-Nielsen, O.E. (1986a): Likelihood and observed geometries. Ann. Statist. 14, 856-873.

Barndorff-Nielsen, O.E. (1986b): Strings, tensorial combinants, and Bartlett adjustments. Proc. Roy. Soc. London A 406, 127-137.

Barndorff-Nielsen, O.E. (1987a): Differential and integral geometry in statistical inference. Differential Geometry in Statistical Inference, IMS Monograph. (To appear).

Barndorff-Nielsen, O.E. (1987b): Differential geometry and statistics: some mathematical aspects. Indian J. Math. (To appear).

Barndorff-Nielsen, O.E. (1987c): Likelihood, ancillarity and strings. Proceedings of the First World Congress of the Bernoulli Society, VNU Science Press, Utrecht.

Barndorff-Nielsen, O.E. and Blæsild, P. (1987a): Strings: Mathematical theory and statistical examples. Proc. Roy. Soc. London A 411, 155-176.

Barndorff-Nielsen, O.E. and Blæsild, P. (1987b): Strings: Contravariant aspect. Proc. R. Soc. London A 411, 421-444.

Barndorff-Nielsen, O.E. and Blæsild, P. (1987c): Coordinate-free definition of structurally symmetric derivative strings. Adv. Appl. Math. (To appear)

Barndorff-Nielsen, O.E., Blæsild, P. and Mora, M.) (1987): Higher order differentiation. Research Report 155, Dept. Theor. Statist., Aarhus University.

Barndorff-Nielsen, O.E., Cox, D.R. and Reid, N. (1986): The role of differential geometry in statistical theory. Int. Statist. Rev. 54, 83-96.

Barndorff-Nielsen, O.E. and Jupp, P.E. (1987): Approximating exponential models. Research Report 161, Dept. Theor. Statist., Aarhus University.

Blæsild, P. (1987): Yokes: elemental properties with statistical applications. This volume.

Campbell, P. (1986): Strings, strings, and, ultimately, strings. Nature, Lond. 320, 679.

Efron, B. (1975): Defining the curvature of a statistical problem (with application to second order efficiency). (With discussion). Ann. Statist. 3, 1189-1242.

Foster, B.L. (1958): Differentiation on manifolds without a connection. Mich. Math. J. 5, 183-190.

Foster, B.L. (1961): Some remarks on tensor differentiation. Annali de Matematica Pura ed Applicata 54, 143-146.

Foster, B.L. (1986): Would Leibniz lie to you? (Three aspects of the affine connection). Math. Intelligencer 8, 34-40, 57.

Foster, B.L. (1987): Pre-geodesic equations. This volume.

Lauritzen, S.L. (1987): Statistical manifolds. Differential Geometry in Statistical Inference, IMS Monograph. (To appear).

McCullagh, P. (1987): Tensor Methods in Statistics. Chapman and Hall, London.

McCullagh, P. and Cox, D.R. (1986): Invariants and likelihood ratio statistics. Ann. Statist. 14, 1419-1430.

Mora, M. (1987): Geometric versions of likelihood expansions. Manuscript. (Title preliminary.)

Murray, M.K. (1987): Co-ordinate systems and Taylor series in statistics. Research Report, School of Mathematical Sciences, The Flinders University of South Australia.

Murray, M.K. and Rice, J.W. (1987): On differential geometry in statistics. Research Report, School of Mathematical Sciences, The Flinders University of South Australia.

Skovgaard, I. (1986): Differential geometric calculations of statistical quantities. Manuscript.

Differential Geometry and Parameters of Interest

P.E.Jupp

Department of Mathematical Sciences

Mathematical Institute

North Haugh

St Andrews KY16 9SS

United Kingdom

Abstract

The function taking the complete parameter of a statistical model to a parameter of interest can be considered as a submersion from the parameter manifold Ω to the manifold K of interest parameters. Such a submersion leads to a decomposition of the tangent bundle of Ω and of the Hilbert bundle over Ω. Some statistical applications of these decompositions are considered, mainly with reference to recent work (i) by Barndorff-Nielsen and Jupp (1987) on defining expected and observed metrics, α-connections etc. on K and (ii) by Amari and Kumon (1985) on "estimating functions" on K. Some open questions are mentioned.

1. Introduction

The basis of the differential-geometric approach to statistics is the association to each parametric statistical model $\mathcal{M} = (\mathcal{X}, p, \Omega)$ of various geometrical objects. In the model $\mathcal{M}$, $\mathcal{X}$ is the sample space, the parameter space Ω is taken to be a smooth finite-dimensional manifold and the model function $p: \mathcal{X} \times \Omega \longrightarrow (0,\infty)$ associates with each parameter value ω the probability density function $p(\cdot;\omega)$ with respect to some reference measure on $\mathcal{X}$. By abuse of notation, p can be regarded as a function $p: \Omega \longrightarrow P(\mathcal{X})$, where $P(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$. The geometrical objects include the expected information metric and skewness tensors and the α-connections studied by Chentsov (1972) and Amari (1982a,b, 1985, 1987), their observed counterparts introduced in Barndorff-Nielsen (1986a,1987), and the strings of Barndorff-Nielsen (1986b).

Submodels of $\mathcal{M}$ are obtained by restricting the parameter ω to lie in a submanifold Θ of Ω. The expected metric and skewness tensors of the submodel are induced from those of $\mathcal{M}$ as follows. If ι is the inclusion map of Θ in Ω then any Riemannian metric ϕ on Ω gives rise to a Riemannian metric $\iota^*\phi$ on Θ defined by $\iota^*\phi(X,Y) = \phi(\iota_* X, \iota_* Y)$ for X,Y $T\Theta_\theta$, or in terms of local coordinates $\theta^1, \ldots, \theta^p$ on Θ and $\omega^1, \ldots, \omega^{p+q}$ on Ω, $\iota^*\phi$ has components $\frac{\partial \omega^i}{\partial \theta^r} \phi_{ij} \frac{\partial \omega^j}{\partial \theta^s}$ (using the summation convention). More generally, ι induces a linear map $\iota^*: \mathcal{T}^0_s(\Omega) \longrightarrow \mathcal{T}^0_s(\Theta)$ of "covariant" tensor fields of order s. Furthermore, given a Riemannian metric ϕ on Ω, the corresponding orthogonal projection of $T\Omega_{\iota(\theta)}$ onto $T\Theta_\theta$ enables tensor fields on Ω to give rise to corresponding tensor fields on Θ. Thus ι induces a linear map of (r,s)-tensor fields, i.e. tensor fields of "contravariant" order r and "covariant" order s. Similarly, affine connections on Ω give rise to affine connections on Θ by means of the Gauss decomposition which is described for Riemannian connections in Chapter VII of Kobayashi and Nomizu (1969). The statistical importance of this construction is its use in defining the

curvature terms in the formulae of Efron (1975), Reeds (1975), L.T.Skovgaard (1984), and Amari (1982a,b,1985) for the mean squared error of a bias-corrected efficient estimator.

The use of an inclusion map $\iota:\Theta \longrightarrow \Omega$ (into Ω) when considering a submodel should be contrasted with the use of a "projection" map $\pi : \Omega \longrightarrow K$ (from Ω) when considering a parameter of interest. In many statistical contexts we may not wish or need to make inference about the parameter in its entirety but only about some function of it, say in $\kappa = \pi(\omega)$. Often we can write Ω as a product, $\Omega = K \times \Psi$ so that $\omega = (\kappa, \psi)$ where κ is the parameter of interest (or structural parameter) and ψ is the incidental (or nuisance) parameter. The question of how to make inference on parameters of interest has been the subject of considerable study.

One approach to this question is based on the use of pseudo-likelihood functions $K \times \mathcal{X} \longrightarrow \mathbb{R}$. Pseudo-likelihood functions which have been studied include

(i) marginal likelihood (based on a statistic satisfying some form of sufficiency for κ),

(ii) conditional likelihood (based on a suitable ancillary statistic),

(iii) integrated likelihood (the likelihood function being integrated over the fibres of π),

(iv) canonical likelihood (the likelihood $L(\kappa,\psi)$ being approximated by a function of the form $L_1(\kappa)L_2(\psi)$, Hinde and Aitkin, 1987),

(v) profile likelihood (the maximum over fibres of π of the likelihood function),

(vi) modified profile likelihood (Barndorff-Nielsen, 1983, 1985),

(vii) conditional profile likelihood (conditioning on $\hat{\psi}$, Cox and Reid, 1987).

These pseudo-likelihood functions can sometimes be used to construct geometrical objects on K.

This paper is concerned mainly with an alternative

(purely geometrical) approach to this statistical question by considering ways in which geometrical objects on Ω give rise to corresponding objects on K. In Section 2 horizontal lifts are introduced and are used to move geometrical objects from ω to $\kappa = \pi(\omega)$. Section 3 is concerned with cases in which this construction does not depend on the choice of ω in $\pi^{-1}(\kappa)$, so that well-defined transferred objects can be defined on K. Section 4 considers a construction in which the splitting of parameters into interest and nuisance parameters leads to a corresponding decomposition of Amari's (1987) Hilbert bundle. Finally some open problems are listed in Section 5.

2. Horizontal lifts

Let $\pi: \Omega \longrightarrow K$ be a differentiable function from one smooth manifold to another. We shall suppose that π is a submersion, i.e. at each point ω of Ω the tangent map π_* of π maps the tangent space $T\Omega_\omega$ onto the tangent space $TK_{\pi(\omega)}$, or in terms of local coordinates $\omega^1, \ldots, \omega^{p+q}$ on Ω and $\kappa^1, \ldots, \kappa^p$ on K the Jacobian matrix $\left(\frac{\partial \kappa^i}{\partial \omega^j}\right)$ has rank p. It follows that each fibre $\pi^{-1}(\kappa)$ is a q-dimensional submanifold of Ω. Further, if Ψ denotes one of these fibres, then small portions of Ω look like small portions of $K \times \Psi$ and π can be identified locally with projection of $K \times \Psi$ onto K.

In the tangent space $T\Omega_\omega$ to Ω at ω, the vertical subspace V_ω is defined by $V_\omega = \{X \in T\Omega_\omega : \pi_* X = 0\}$. Thus in terms of local coordinates $\psi^1, \ldots, \psi^q$ on the fibre Ψ, V_ω can be identified with the span of $\{\frac{\partial}{\partial \psi^i} : i=1, \ldots, q\}$. A smooth mapping which ascribes to each point ω of Ω a complementary subspace (the horizontal subspace) H_ω to V_ω in $T\Omega_\omega$ is called an (Ehresmann) connection on the submersion π. (An equivalent definition is given on pages 71-86 of Hermann, 1975). Such a connection can be used to lift vectors in $TK_{\pi(\omega)}$ to $T\Omega_\omega$ as follows. Given a tangent vector X to Ω at $\pi(\omega)$, the

horizontal lift $\bar{X}$ of X at ω is defined as the unique element $\bar{X}$ of H_ω satisfying $\pi_*(\bar{X}) = X$. A Riemannian metric ϕ on Ω gives rise to a connection on π by taking H_ω as the orthogonal complement of V_ω in $T\Omega_\omega$.

Horizontal lifts can be used to transfer geometry along π to K. For example, if ϕ is a Riemannian metric on Ω then an inner-product $\pi_\omega\phi$ on $TK_{\pi(\omega)}$ is defined by

$$\pi_\omega\phi(X,Y) = \phi(\bar{X},\bar{Y})$$

where $\bar{X}, \bar{Y}$ are the horizontal lifts at ω of X, Y in $TK_{\pi(\omega)}$. Let

$$\begin{pmatrix} \phi_{\kappa\kappa} & \phi_{\kappa\psi} \\ \phi_{\psi\kappa} & \phi_{\psi\psi} \end{pmatrix}$$

be the partitioned matrix of the expression for ϕ in terms of the local coordinates $\kappa^1, \ldots, \kappa^p, \psi^1, \ldots, \psi^q$ on Ω. Calculation shows that (if the connection on π is that given by ϕ) the corresponding matrix for $\pi_\omega\phi$ is

$$\phi_{\kappa\kappa.\psi} = \phi_{\kappa\kappa} - \phi_{\kappa\psi}\phi_{\psi\psi}^{-1}\phi_{\psi\kappa}.$$

More generally, if A is an (r,s)-tensor on Ω and ω is in Ω, then an (r,s)-tensor $\pi_\omega A$ at $\pi(\omega)$ is defined by

$$(\pi_\omega A)(X_1, \ldots, X_s) = (\otimes^r \pi_*)(A(\bar{X}_1, \ldots, \bar{X}_s)$$

where $X_i \in TK_{\pi(\omega)}$, $\bar{X}_i$ is the horizontal lift to ω of X_i for $i=1, \ldots, s$ and $\otimes^r \pi_* : \otimes^r T\Omega_\omega \longrightarrow \otimes^r TK_{\pi(\omega)}$ is the r-fold tensor product of the tangent map of π.

An alternative construction is based on the following identification of tangent vectors of Ω with cotangent vectors of Ω by means of the Riemannian metric ϕ. A tangent vector X to Ω at ω can be identified with the cotangent vector (1-form) $X^\flat$ defined by

$$X^\flat(Y) = \phi(X,Y) \qquad Y \in T\Omega_\omega.$$

Thus there is an invertible linear mapping

$\flat : T^1_0(\Omega)_\omega \longrightarrow T^0_1(\Omega)_\omega$. In terms of local coordinates, the tangent vector with components a^i is mapped by $\flat$ into the cotangent vector with components $\phi_{ij}a^j$, i.e. the mapping $\flat$ is the traditional "lowering of indices". The inverse of $\flat$ is $\# : T^0_1(\Omega)_\omega \longrightarrow T^1_0(\Omega)_\omega$, corresponding to "raising of indices" in local coordinates. More generally, ϕ gives rise to the invertible linear transformation

$\flat : T^{r+s}_0(\Omega)_\omega \longrightarrow T^r_s(\Omega)_\omega$ defined by

$$(X_1 \otimes \ldots \otimes X_r \otimes Y_1 \otimes \ldots \otimes Y_s)^\flat (Z_1 \otimes \ldots \otimes Z_s) = \left\{ \prod_{i=1}^{s} \phi(Y_i, Z_i) \right\} X_1 \otimes \ldots \otimes X_r$$

for $X_1, \ldots, X_r, Y_1, \ldots, Y_s, Z_1, \ldots, Z_s \in T\Omega_\omega$. In local coordinates the $(r+s,0)$-tensor with components $a^{i_1 \ldots i_r j_1 \ldots j_s}$ is mapped by $\flat$ to the (r,s)-tensor with components $\phi_{j_1 k_1} \ldots \phi_{j_s k_s} a^{i_1 \ldots i_r j_1 \ldots j_s}$. Note that any positive-definite symmetric $(2,0)$-tensor field is dual to a Riemannian metric and so gives rise to raising and lowering operators. Now the tangent map π_* of π induces a linear mapping "in the

direction of π", $\pi_* : T^r_0(\Omega)_\omega \longrightarrow T^r_0(K)_{\pi(\omega)}$. In coordinate terms, $a^{i_1 \ldots i_r}$ is mapped to

$$\frac{\partial \omega^{i_1}}{\partial \kappa^{k_1}} \ldots \frac{\partial \omega^{i_r}}{\partial \kappa^{k_r}} a^{k_1 \ldots k_r}.$$

The following diagram defines a linear mapping from $T^r(\Omega)_\omega$ to $T^r(K)_{\pi(\omega)}$

$$\begin{array}{ccc} T^r_s(\Omega) & \xrightarrow{\ \#\ } & T^{r+s}_0(\Omega) \\ & & \downarrow \pi_* \\ T^r(K)_{\pi(\omega)} & \xleftarrow{\ \flat\ } & T^{r+s}_0(K)_{\pi(\omega)} \end{array}$$

where $\flat$ is defined using the $(2,0)$-tensor $\pi_*(\phi(\omega)^\#)$. It can be shown that this mapping is in fact π_ω as defined above.

The above construction is useful in the statistical context when Ω is the parameter space of a statistical model, ϕ is the Fisher information i and K is the space of parameters of interest. Then $\pi_\omega i$ is known as the Fisher information about κ eliminating ψ or the orthogonalised Fisher information

(Amari, 1985). It has the important property that

$$(\pi_\omega i)_{jj} = \min_{\xi} E[(\frac{\partial l}{\partial k^j} - c_i \frac{\partial l}{\partial \psi^i})^2]$$

(where l denotes log-likelihood). See also Liang (1983) and Godambe (1984). Further, $(\pi_\omega i)^{-1}$ is the Cramer-Rao lower bound for the covariance matrix of an unbiased estimator of κ. Now suppose that the model $\mathcal{M}$ is a curved exponential family, so that the model function takes the form

$$p(x; \omega) = a(x) \exp\{ \theta(\omega).t(x) - c(\theta(\omega))\}$$

for some functions $t: \mathcal{X} \longrightarrow \mathbb{R}^k$ and $\theta: \Omega \longrightarrow \mathbb{R}^k$. Let $u: \mathbb{R}^k \longrightarrow K$ be an estimator of $\kappa = \pi(\omega)$ which assigns to a sample $x_1, \ldots, x_n$ the estimate $u(n^{-1}\sum_{i=1}^{n} t(x_i))$ of κ. Let ω_0 be the true value of the parameter, and put $\kappa_0 = \pi(\omega_0)$, $A(\kappa_0) = u^{-1}(\kappa_0)$ and $N(\kappa_0) = \theta(\pi^{-1}(\kappa_0))$. Amari (1985, Chapter 8) relates the higher order efficiency of u to the geometry of the submanifolds $\theta(\Omega)$, $A(\kappa_0)$ and $N(\kappa_0)$ of $\mathbb{R}^k$, the natural parameter space of $\mathcal{M}$. In particular, if u is

efficient,

$$\mathrm{Cov}(u^a, u^b) = n^{-1} g^{ab} + O(n^{-2})$$

where g^{ab} denotes the (a,b)-element of the inverse of $\pi_\omega i$. Further, Efron's (1975) formula for the mean squared error of the bias-corrected version of u extends to the case where nuisance parameters are present, the term of order n^{-1} containing a sum of various condensed second fundamental forms of the submanifolds $\theta(\Omega)$, $A(\kappa_0)$ and $N(\kappa_0)$.

A construction similar to that given above is used in semi-parametric models. In such models the density functions are indexed by $\omega = (\kappa, g) \in K \times G$ where K is a finite-dimensional manifold and G is a set which may be infinite-dimensional. For any statistical model $\mathcal{M} = (\mathcal{X}, p, \Omega)$, Amari (1987) defines the associated Hilbert bundle as

$$\mathcal{H}(\mathcal{M}) = \{(\omega, f) : E_\omega(f^2) < \infty, \; E_\omega(f) = 0\}$$

so that the fibre over ω is a subspace of $L^2_{p(\omega)}(\mathcal{X})$. The Hilbert bundle has metric tensor ϕ defined by

$$\phi(f_1, f_2) = E_\omega[f_1 f_2].$$

If Ω is finite-dimensional then the score function

$$a^i \frac{\partial}{\partial \omega^i} \longmapsto a^i \frac{\partial l}{\partial \omega^i}$$

identifies the tangent space to Ω at ω with a subspace of the fibre of $\mathcal{H}(\mathcal{M})$ over ω. Further, under this identification ϕ induces the expected information metric on Ω. For a semi-parametric model, the tangent space $T\mathcal{M}_\omega$ of $\mathcal{M}$ at ω is defined as the span in $L^2_{p(\omega)}(\mathcal{X})$ of all score functions at ω of one-dimensional parametric submodels of $\mathcal{M}$. Using the metric ϕ, we can define the horizontal subspace of $T\mathcal{M}_\omega$ as the orthogonal complement in $T\mathcal{M}_\omega$ of TG_g. The corresponding inner-product $\pi_\omega\phi$ on TK_κ is called the effective information on κ. It plays a central role in versions of semi-parametric Hajek-Le Cam formulations of the Cramer-Rao lower bound. See Begun, Hall, Huang and Wellner (1983) and Wellner (1985).

3. Transfer of geometries

In general, the inner-product $\pi_\omega\phi$ defined in Section 2 depends on the point ω to which the vectors X and Y in $TK_{\pi(\omega)}$ are lifted. In the special case in which $\pi_\omega\phi$ depends only on $\pi(\omega)$ we say that ϕ transfers to κ and define the transfer $\pi_!\phi$ of ϕ by

$$\pi_!\phi(\kappa) = \pi_\omega\phi \qquad \text{for any } \omega \in \pi^{-1}(\kappa).$$

Letting κ run through K, we obtain a Riemannian metric $\pi_!\phi$ on K. Transfer of tensor fields from Ω to K is defined similarly.

The transfer construction can be applied also to affine connections. Given an affine connection ∇ on Ω we can define $\pi_\omega\nabla$ by

$$(\pi_\omega\nabla)_X Y = \pi_*(\nabla_{\bar{X}}\bar{Y}(\omega))$$

where X, Y are vector fields in a neighbourhood of κ and $\bar{X}, \bar{Y}$

are their horizontal lifts to a neighbourhood of $\omega \in \pi^{-1}(\kappa)$. If the right hand side depends on ω only through $\pi(\omega)$ then ∇ transfers to K and the transfer $\pi_! \nabla$ is an affine connection on K. Simple calculations prove the following result which is useful in the statistical context.

Theorem Let ϕ and ∇ be respectively a Riemannian metric and an affine connection on Ω.

(i) If ϕ and ∇ transfer to K then

$$(\pi_! \nabla)(\pi_! \phi) = \pi_!(\nabla \phi).$$

(ii) If ∇ is the Levi-Civita connection of ϕ and ϕ transfers then ∇ also transfers and $\pi_! \nabla$ is the Levi-Civita connection of $\pi_! \phi$.

Formulae analogous to the Gauss-Codazzi equations of an immersion are given by O'Neill (1966) and yield the following

result.

Theorem Let ϕ be a Riemannian connection which transfers to K. Then the sectional curvatures of $\pi_! \nabla$ are greater than or equal to the corresponding sectional curvatures of the Levi-Civita connection ∇. Also, the following are equivalent:

(i) equality of sectional curvatures,

(ii) $R(\pi_! \nabla) = \pi_! R(\nabla)$,

(iii) the local coordinates κ, ψ can be chosen to be orthogonal.

Barndorff-Nielsen and Jupp (1987) consider in some detail conditions under which the expected and observed geometries (information metric, skewness tensor, α-connections) of a parametric statistical model transfer to the corresponding geometries on a manifold of interest parameters.

Important cases in which transfer occurs include the exponential models with τ-parallel or θ-parallel foliations of

Barndorff-Nielsen and Blaesild (1983), the extended class of generalised linear models considered by Barndorff-Nielsen (1983) and Jørgensen (1983) and composite transformation models (see Barndorff-Nielsen, Blaesild, Jensen and Jørgensen, 1982). In the latter models a group G acts on $\mathcal{X}$ and Ω so that

$$p(x;g\omega) = \chi(g,x)\, p(g^{-1}x;\omega)$$

for some function $\chi: G \times \mathcal{X} \longrightarrow (0,\infty)$. In this case, invariance under the group action ensures that the expected and observed geometries transfer to $K = \Omega/G$.

In many cases in which statistical geometries transfer, the transfer is ensured by a statistic $u: \mathcal{X} \longrightarrow \mathcal{Y}$ satisfying some form of sufficiency. That is, the diagram

$$\begin{array}{ccc} \Omega & \longrightarrow & P(\mathcal{X}) \\ \downarrow \pi & & \downarrow u_* \\ K & \longrightarrow & P(\mathcal{Y}) \end{array}$$

commutes and (in some sense) no loss of information about ω

is incurred by mapping x to $u(x)$. An appropriate form of sufficiency is L-sufficiency, introduced by Rémon (1984) and defined as follows.

Recall that profile likelihood $\tilde{L}$ is defined by

$$\tilde{L}(\kappa;x) = \sup_{\omega} L(\omega;x) = L(\omega_\kappa;x)$$

where $L(\omega;x) = p(x;\omega)$. Put $\tilde{l}(\kappa;x) = \log \tilde{L}(\kappa;x)$.

Definition A statistic $u: \mathcal{X} \to \mathcal{Y}$ is weakly L-sufficient for κ if $\frac{\partial \tilde{l}}{\partial \kappa}(\kappa;x)$ depends on x only through $u(x)$. If also the distribution of $u(x)$ depends only on κ then u is L-sufficient for κ.

Note that if u is L-sufficient for κ then the distribution of the m.l.e. $\hat{\kappa}$ depends only on κ. One consequence of this is the following asymptotic result for repeated random sampling.

Theorem If for every n there is a statistic $u_n: \mathcal{X}^n \to \mathcal{Y}_n$ which is L-sufficient for κ then the expected

information and skewness tensor transfer to K.

In cases where there is a statistic u making the above diagram commute, we can consider the marginal model $K \longrightarrow P(\mathcal{Y})$, the marginal log-likelihood $\breve{l}$ and the corresponding geometries. There are also geometries based on the profile likelihood $\tilde{l}$. Use of Barndorff-Nielsen's (1983,1985) modified profile log-likelihood $\tilde{\tilde{l}}$ gives rise to yet another set of geometries. Under repeated sampling from a composite transformation model the various geometries are closely related. For example, for the expected information from samples of size n,

$$\breve{i} = \pi_1 i + O(1).$$

4. Decomposing the Hilbert bundle

Recall from Section 2 that associated with a parametric model $\mathcal{M}$ is a "Hilbert bundle" $\mathcal{H}$ over the parameter space Ω and an inclusion of $T\Omega$ in $\mathcal{H}$. Given a submersion $\pi: \Omega \longrightarrow K$, the metric on $\mathcal{H}$ gives rise to a decomposition of $\mathcal{H}$ as the orthogonal direct sum

$$\mathcal{H} = V \oplus H \oplus (T\Omega)^{\perp} \qquad (4.1)$$

where $\mathcal{H}_\omega = T\Omega_\omega \oplus (T\Omega_\omega)^{\perp}$, $T\Omega_\omega = V_\omega \oplus H_\omega$, the latter forming the basis of Section 2. An interesting modification of this decomposition was introduced by Amari and Kumon (1985) in a study of estimating functions. See also Kumon and Amari (1984) and Amari (1987) as well as Amari's paper given at this workshop.

An "estimating function" for κ is a function γ from $\mathcal{X}$ to the set of 1-form fields on K. Abusing notation we write $\gamma: \mathcal{X} \times K \longrightarrow T^*K$. An estimating function γ yields an estimator which assigns to observations $x_1, \ldots, x_n$ the estimate $\hat{\kappa}$ of κ given by

$$\sum_{i=1}^{n} \gamma(x_i; \hat{\kappa}) = 0.$$

If $x_1, \ldots, x_n$ are observations on independent random variables $X_1, \ldots, X_n$ with X_i having p.d.f. $p(\cdot; \omega_i)$ where $\pi(\omega_i) = \kappa$,

consistency of the above estimator requires that

$$E_\omega[\gamma(x;\kappa)] = 0 \qquad (4.2)$$

whenever $\pi(\omega) = \kappa$. If (4.2) holds then (again abusing notation) we have

$$\gamma : \Omega \longrightarrow K \longrightarrow \pi^*T^*K \otimes \mathcal{H}$$

so that γ can be regarded as a "π-invariant" section of $\pi^*T^*K \otimes \mathcal{H}$.

This "π-invariance" of γ has a neat interpretation in terms of parallelism of γ along fibres of π. It is useful to consider the subbundle $\mathcal{F}$ of $\mathcal{H}$ defined by

$$\mathcal{F} = \{(\omega, f) : E_\omega(f) = 0, \quad Var_\omega(f) < \infty, \ \forall \omega' \in \Omega\}$$

We shall asume that γ is in fact a section of the subbundle $\pi^*T^*K \otimes \mathcal{F}$ of $\pi^*T^*K \otimes \mathcal{H}$. Amari (1987) introduced dual flat connections $\overset{e}{\nabla}$ and $\overset{m}{\nabla}$ on $\mathcal{F}$. Their parallel translation operators take $f \in \mathcal{F}_{\omega'}$ to

$$\overset{e}{\Pi}{}^{\omega}_{\omega'} f = f(\cdot) - E_\omega(f) \in \mathcal{F}_\omega$$

and

$$\overset{m}{\Pi}{}^{\omega}_{\omega'} f = \frac{p(\cdot;\omega')}{p(\cdot;\omega)} f(\cdot) \in \mathcal{F}_\omega .$$

[These connections on $\mathcal{F}$ induce respectively Amari's 1- and -1-connections on Ω.] Then "π-invariance" of a section of $\pi^*T^*K \otimes \mathcal{F}$ is equivalent to its being parallel along each fibre of π. This leads to the definitions

$$\mathcal{F}^T_\omega = \overline{\text{span}\{\overset{m}{\Pi}{}^{\omega}_{\omega'}(f) : f \in T_\omega\Omega, \ \omega \in \pi^{-1}(\pi(\omega))\}},$$

$$\mathcal{F}^N_\omega = \overline{\text{span}\{\overset{m}{\Pi}{}^{\omega}_{\omega'}(f) : f \in V_\omega, \ \omega \in \pi^{-1}(\pi(\omega))\}}.$$

Let $\mathcal{F}^A_\omega$ be the orthogonal complement of $\mathcal{F}^T_\omega$ in $\mathcal{F}_\omega$ and let $\mathcal{F}^I_\omega$ be the orthogonal complement in $\mathcal{F}^T_\omega$ of $\mathcal{F}^N_\omega$. Taking unions as ω runs through Ω we obtain bundles $\mathcal{F}^I$, $\mathcal{F}^N$ and $\mathcal{F}^A$ known respectively as the information, nuisance and ancillary bundles which form the orthogonal summands in the decomposition

$$\mathcal{F} = \mathcal{F}^I \oplus \mathcal{F}^N \oplus \mathcal{F}^A . \qquad (4.3)$$

Horizontal lifting, inclusion and projection combine to give a map $TK \longrightarrow T\Omega \longrightarrow \mathcal{F} \longrightarrow \mathcal{F}^I$ which then yields a section u^I of $\pi^*T^*K \otimes \mathcal{F}$.

Theorem There is an estimating function giving a consistent asymptotically normal estimator with variance of order n^{-1} if and only if dim $\mathcal{F}^{I}_{\omega} > 0$. If u^{l} is $\overset{m}{\nabla}$ - parallel along fibres of π then the corresponding estimator has minimum asymptotic variance.

There is a similar characterisation of uniformly informative estimators and of optimality within this class.

5. Open questions

Here are some miscellaneous open questions.

(i) Under what conditions do the strings of Barndorff-Nielsen (1986b) transfer? What is the statistical usefulness of this?

(ii) Are there any statistically useful connections on the submersion π other than those derived from the expected and observed metrics?

(iii) How do the various geometries compare in the case of stochastic processes?

(iv) What form of ancillarity of a statistic u: guarantees transfer of statistical geometries?

(v) Does L-sufficiency of u imply that expected geometry transfers?

References

Amari, S.-I. (1982a): Differential geometry of curved exponential families - curvatures and information loss. Ann. Statist. 10, 357-385.

Amari, S.-I. (1982b): Geometrical theory of asymptotic ancillarity and conditional inference. Biometrika 69, 1-17.

Amari, S.-I. (1985): Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics 28. Springer-Verlag, Berlin.

Amari, S.-I. (1987): Differential geometry of statistics - towards new developments. In Differential Geometry in Statistical Inference, IMS Monograph. To appear.

Amari, S.-I. and Kumon, M. (1985): Optimal estimation in the presence of infinitely many nuisance parameters - geometry of estimating functions. Research report METR 85-2. Dept. of Mathematical Engineering and Instrumentation Physics, University of Tokyo.

Barndorff-Nielsen, O.E. (1983): On a formula for the distribution of the maximum likelihood estimator. Biometrika 70, 343-365.

Barndorff-Nielsen, O.E. (1985): Properties of modified profile likelihood. In J. Lanke and G. Lindgren (eds.): Contributions to Probability and Statistics in Honour of Gunnar Blom, pp. 25-38. Lund.

Barndorff-Nielsen, O.E. (1986a): Likelihood and observed geometries. Ann. Statist. 14, 856-873.

Barndorff-Nielsen, O.E. (1986b): Strings, tensorial combinants, and Bartlett adjustments. Proc. Roy. Soc. Lond. A 406, 127-137.

Barndorff-Nielsen, O.E. (1987): Differential and integral geometry in statistical inference: some aspects. In Differential Geometry in Statistical Inference, IMS Monograph. To appear.

Barndorff-Nielsen, O.E. and Blaesild, P. (1983): Exponential models with affine dual foliations. Ann. Statist. 11, 753-769.

Barndorff-Nielsen, O.E., Blaesild, P., Jensen, J.L. and Jørgensen, B. 1982): Exponential transformation models. Proc. R. Soc. A379, 41-65.

Barndorff-Nielsen, O.E., Cox, D.R. and Reid, N. (1986): The role of differential geometry in statistical theory. Int. Statist. Rev. 54, 83-96.

Barndorff-Nielsen, O.E. and Jupp, P.E. (1987): Differential geometry, profile likelihood, L-sufficiency and composite transformation models. To appear in Ann. Statist.

Begun, J.M., Hall, W.J., Huang, W.M. and Wellner,J.A. (1983): Information and asymptotic efficiency in parametric -non parametric models. Ann. Statist. 11, 432-452.

Chentsov, N.N. (1972): Statistical Decision Rules and Optimal Conclusions. (In Russian.) Nauka, Moscow. Translation in English (1982) by Amer. Math. Soc., Providence, Rhode Island.

Cox, D.R. and Reid, N. (1987): Parameter orthogonality and approximate conditional inference (with discussion). J. R. Statist. Soc. B49, 1-39.

Efron, B. (1975): Defining the curvature of a statistical problem (with applications to second order efficiency). With discussion. Ann. Statist. 3, 1189-1242.

Godambe,V.P. (1984): On ancillarity and Fisher information in the presence of a nuisance parameter. Biometrika 71, 626-629.

Hermann, R. (1975): Gauge Fields and Cartan-Ehresmann Connections, Part A (Interdisciplinary Mathematics, Vol. X). Math. Sci. Press, Brookline, Mass.

Hinde, J. and Aitkin, M. (1987): Canonical likelihoods: A new likelihood treatment of nuisance parameters. Biometrika 74, 45-58.

Jørgensen, B. (1983): Maximum likelihood estimation and large-sample inference for generalized and nonlinear regression models. Biometrika 70, 19-28.

Kobayashi, S. and Nomizu, K. (1969): Foundations of Differential Geometry. Vol. II. Interscience, New York.

Kumon, M. and Amari, S.-I. (1984): Estimation of a structural parameter in the presence of a large number of nuisance parameters. Biometrika 71, 445-459.

Lauritzen, S.L. (1987): Statistical manifolds. In Differential Geometry in Statistical Inference, IMS Monograph. To appear.

Liang, K.-Y. (1983): On information and ancillarity in the presence of a nuisance parameter. Biometrika 70, 607-612.

Lindsay, B.G. (1982): Nuisance parameters, mixture models, and the efficiency of partial likelihood. Phil. Trans. Roy. Soc. A 296, 639-665.

O'Neill, B. (1966): The fundamental equations of a submersion. Michigan Math. J. 13, 459-469.

Reeds, J. (1975): Contribution to the discussion of Efron (1975).

Rémon, M. (1984): On a concept of partial sufficiency: L-sufficiency. Int. Statist. Rev. 52, 127-135.

Skovgaard, L.T. (1984): A Riemannian geometry of the multivariate normal model. Scand. J. Statist. 11, 211-223.

Wellner, J.A. (1985): Semi-parametric models: progress and problems. Proceedings of the 45th Session of the International Statistical Institute, Amsterdam.

Dual Connections on the Hilbert Bundles of Statistical Models

Shun-ichi Amari

Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, 113 JAPAN

Abstract

A fibre bundle is constructed on a finite dimensional parametric statistical model with a Hilbert space as the fibre space. The Hilbert space represents the tangent directions of the set of probability distributions in the function space. A pair of dual linear connections are introduced in the Hilbert bundle. This framework is useful for the study of inference in semi-parametric statistical models. The theory is applied to the statistical inference in the presence of an infinitely large number of nuisance parameters. It is also extended and applied to statistical inference in a semi-parametric model. These results show that a fundamental role is played by the notion of the exponential and mixture parallel transports and their duality.

1. Introduction

The dualistic theory of the α-geometry has been constructed on a regular parametric statistical model or a regular family of probability distributions (Chentsov [1972], Amari [1982a, 1985, 1987a], Nagaoka and Amari [1982], Eguchi [1983], Barndorff-Nielsen [1987], Lauritzen [1987], etc.). The higher-order asymptotic theory of statistical inference has been built in the framework of the curved exponential family of distributions by using the α-geometry (Amari [1982b, 1983, 1985], Amari and Kumon [1983], Kumon and Amari [1983, 1985]). In order to extend the theory to a more general statistical model including a semi-parametric or nonparametric model, it is desirable to construct the α-geometry of the function space of probability distributions. However, it is not an easy task to do so. We propose here a fibre bundle theory on a finite dimensional manifold, where the fibre space may have a functional degree of freedom (see also Amari [1987a], Kumon and Amari [1984], Amari and Kumon [1985]).

The fibre bundle theory can be applied to construct a theory of statistical inference on a model which includes an infinite number of nuisance parameters. It is expected to give an important tool for analyzing statistical problems such as robust inference in semi-parametric and non-parametric statistical models.

We first define a Hilbert bundle B(M) on a finite dimensional regular statistical manifold M. The fibre space is an infinite-dimensional Hilbert space, which includes the tangent space as a subspace. A one-parameter family of affine connections, which is called the α-connections, are introduced in the fibre bundle. The duality theory is also developed in the present framework. It is proved that the

fibre bundle is $\alpha = \pm 1$ curvature free, that is e- and m-curvature free. It includes the tangent bundle as a subbundle, and the α-geometry of the manifold M is naturally induced from that of B(M).

We then apply the fibre bundle theory to the problem of estimating a parameter of interest θ in the presence of an infinite number of nuisance parameters $\xi_1, \xi_2, \ldots$. More specifically, we consider a statistical model $M = p(x, \theta, \xi)$ with the parameter θ of interest and the nuisance parameter ξ, and the problem is to estimate θ from a large number of independent observations $x_1, x_2, \ldots$, where x_i is subject to the probability distribution $p(x, \theta, \xi_i)$. The sequence $\xi_1, \xi_2, \ldots$ is unknown, so that we have an infinite number of nuisance parameters. In order to perform statistical inference in this situation, parallel transports of fibre vectors along the ξ-axis play a fundamental role (Kumon and Amari [1984], Amari and Kumon [1985]). We study the characteristics of M-estimators which are given by solving the following estimating equations

$$\sum_i y(x_i, \theta) = 0.$$

An estimating function $y(x, \theta)$ is said to be universally optimal, if the estimator $\hat{\theta}$ derived therefrom gives the minimum asymptotic variance for any sequence of the nuisance parameter. We solve the following problems by using the fibre bundle theory: 1) When does a consistent M-estimator exist? 2) When does the universally optimal M-estimator exist and what is it?

When the sequence ξ_i of the nuisance parameter is assumed to be independent realizations from a common but unknown distribution $Z(\xi)$ of the nuisance parameter ξ, every x_i is regarded to be subject to the probability distribution

$$f(x, \theta, Z) = \int p(x, \theta, \xi)\, dZ(\xi),$$

parametrized by θ and the nuisance parameter $Z(\xi)$ of functional degrees of freedom. Such a statistical model is called a semi-parametric model (Begun et al. [1983]). We finally apply the fibre bundle theory to a semi-parametric statistical model. Even when there does not exist a universally optimal estimator, we can obtain the optimal M-estimator by using an adaptive estimator of $Z(\xi)$ (Bickel [1982], Lindsay [1985]). We can solve the following problems by using the fibre bundle theory: 1) When does a consistent M-estimator exists? 2) What is the amount of loss of information by using M-estimators and when is the optimal M-estimator fully efficient? 3) What is the optimal M-estimator?

The geometrical theory of statistical inference has been developed by many researchers in many directions (see Amari [1985, 1987a], Barndorff-Nielsen et al. [1986], Kass [1987]). Barndorff-Nielsen [1987] and his group have developed the theory from various points of view; the observed α-geometry, the theory of strings, the geometry of transformation models, etc. (See Barndorff-Nielsen's paper in this volume.) Mitchell and Krzanowsky [1985] studied the geometrical structure of elliptic models, which are also transformational models. Lauritzen [1987] and Picard [1987] studied some mathematical problems of the α-geometry.

There are two more interesting applications of the α-geometry. One is application to the time series analysis and families of linear control systems (Amari [1987b], see also Delchamps [1985]). The other is to information theory (Campbell [1986], Amari and Han [1987]). All of these applications together with the theory will open a new field of information science, which will be called information geometry.

7. Hilbert bundle of a statistical model

Let x be a (vector) random variable, and let $p(x, \theta)$ be probability density functions with respect to some dominating measure, parametrized by an n-dimensional vector parameter $\theta = (\theta^1, \theta^2, \ldots, \theta^n)$. The set of these probability distributions is called a statistical model. Under certain regularity conditions (Amari [1985]), the set M forms a differentiable manifold with an admissible coordinate system θ, where θ belongs to a set Θ.

Let F_0 be the set of random variables $r(x)$ having the second order moments with respect to any $p(x, \theta)$, i.e.,

$$F_0 = \{ r(x) \mid E_\theta[r(x)^2] < \infty, \quad \theta \in \Theta \},$$

where E_θ is the expectation with respect to $p(x, \theta)$. This F_0 is a linear space. The direct product $M \times F_0$ is a trivial fibre bundle.

Let F_θ be the subspace of F_0 defined by

$$F_\theta = \{ r(x) \mid E_\theta[r(x)] = 0, \quad r \in F_0 \}.$$

When the probability distribution $p(x, \theta)$ deviates to $p(x, \theta)\{1 + \varepsilon r(x)\}$, where ε is infinitesimally small, its relative deviation $r(x)$ satisfies $E_\theta[r(x)] = 0$. Hence every element $r \in F_\theta$ represents a direction in which $p(x, \theta)$ may deviate. The aggregate of all these F_θ's forms a fibre bundle with the base space M. This is a subbundle of F_0, and is called the Hilbert bundle B(M) associated with a

statistical model M. An element of B(M) is a pair (θ, r) with $\theta \in M$, $r \in F_\theta$. An inner product is naturally introduced in each F_θ by

$$\langle r(x), s(x) \rangle_\theta = E_\theta [r(x)s(x)]. \tag{2.1}$$

This shows that each F_θ is a Hilbert space.

Let $r(x, \theta)$ be a cross section of B(M). This implies that $r(x, \theta)$ is a vector field of M whose vector $r(x, \theta)$ defined at θ belongs to F_θ. Since F_θ depends on θ, the partial derivative $\partial_i r(x, \theta)$, where

$$\partial_i = \partial/\partial\theta^i,$$

does not represent the intrinsic change of r as θ changes. It is necessary to introduce a linear connection in the bundle B(M), which defines a linear correspondence between F_θ and $F_{\theta'}$ along a curve connecting θ and θ'. Equivalently, we may define a covariant derivative operator ∇.

We introduce the α-connection, or the α-covariant derivative in B(M), where α is a scalar parameter. The α-covariant derivative in the direction of ∂_i, i.e., in the direction of the i-th coordinate curve θ^i, is defined by

$$\nabla_i^{(\alpha)} r = \partial_i r(x, \theta) - \frac{1+\alpha}{2} E_\theta[\partial_i r] + \frac{1-\alpha}{2} r u_i, \tag{2.2}$$

where

$$u_i(x, \theta) = (\partial/\partial\theta^i) \log p(x, \theta) \tag{2.3}$$

is an element of F_θ, because of $E_\theta[u_i(x, \theta)] = 0$. It is

easy to prove

$$E_\theta[\nabla_i^{(\alpha)} r] = 0,$$

because we have

$$E[\partial_i r + r u_i] = 0 \tag{2.4}$$

by differentiating $E_\theta[r] = 0$ with respect to θ^i. Hence, $\nabla_i^{(\alpha)} r$ again belongs to F_θ, i.e., it is again a cross section of B(M). The α-derivative in the direction of $A = A^i\partial_i$ is given by

$$\nabla_A^{(\alpha)} r = A^i \nabla_i^{(\alpha)} r, \tag{2.5}$$

where Einstein's summation convention is assumed.

We can represent the tangent vector $\partial_i = \partial/\partial\theta^i$ of M, i.e., the direction of the change of the distribution $p(x, \theta)$ as θ changes in the direction of θ^i, by the vector u_i in F_θ. Let T_θ be the linear subspace spanned by $u_1, u_2, \ldots, u_n$. It is a subspace of F_θ and is called the tangent space of M at θ, where the vector u_i is identified with ∂_i. Then, the aggregate of these T_θ is the tangent bundle T(M) of M. It is a subbundle of B(M).

When $r(x, \theta)$ belongs to T_θ for any θ, it is a cross section (vector field) of T(M). However, its α-derivative $\nabla_i^{(\alpha)} r$ does not necessarily belong to T_θ. Let $\tilde{\nabla}_i^{(\alpha)} r$ be the projection of $\nabla_i^{(\alpha)} r$ to the subspace T_θ by the use of the inner product $\langle , \rangle_\theta$. This $\tilde{\nabla}_i^{(\alpha)}$ defines a linear connection in the tangent bundle T(M). This is the geometry of M induced from that of B(M). Indeed, the metric tensor $g_{ij}(\theta)$ is naturally induced by

$$g_{ij}(\theta) = \langle u_i, u_j \rangle_\theta = E_\theta[u_i u_j], \tag{2.6}$$

which is the Fisher information metric. The components of the induced connection are given by the expansion

$$\nabla^{(\alpha)}_i u_j = \Gamma^{(\alpha)m}_{ij} u_m,$$

or explicitly by

$$\Gamma^{(\alpha)}_{ijk}(\theta) = \langle \nabla^{(\alpha)}_i u_j, u_k \rangle_\theta.$$

From (2.2), we have

$$\Gamma^{(\alpha)}_{ijk}(\theta) = E_\theta[\{\partial_i u_j + \frac{1-\alpha}{2} u_i u_j\} u_k]. \tag{2.7}$$

This implies that the induced geometry is exactly the α-geometry studied in Amari [1985]. Hence, this shows that the α-connection of B(M) is its natural generalization.

The α-connection defines a parallel transport of a vector $r(x) \in F_\theta$ from a point θ to another point θ' along a curve $\theta = \theta(t)$ connecting them. Let $r = r(x, t) = r(x, \theta(t))$ be the vector field defined on the curve $\theta(t)$. Then, the vector field $r(x, t)$ is said to be α-parallel along the curve, if

$$\nabla^{(\alpha)}_{\dot\theta} r(x, t) = 0$$

or

$$\dot r - \frac{1+\alpha}{2} E_{\theta(t)}[\dot r] + \frac{1-\alpha}{2} r \dot\theta^i u_i = 0 \tag{2.8}$$

is satisfied, where $\dot{}$ implies d/dt and

$$\dot\theta = \dot\theta^i u_i(x, \theta(t))$$

is the tangent vector of the curve. When the α-parallel field $r(x, t)$ is equal to $r(x)$ at θ and is equal to $r'(x)$ at θ', this $r'(x)$ is said to be the α-parallel transport of $r(x)$ along the curve. We denote it by

$$r' = \Pi^{(\alpha)}_{\theta(t)} r.$$

The $\alpha = 1$ parallel transport is called the exponential or briefly e-parallel transport. The $\alpha = -1$ parallel transport is called the mixture or briefly m-parallel transport.

The parallel transport in general depends on the curve along which the vector is transported. When the connection is curvature free, the transport does not depend on the curve. The α-curvature is defined by the operator

$$R^{(\alpha)}_{ij} = \nabla^{(\alpha)}_i \nabla^{(\alpha)}_j - \nabla^{(\alpha)}_j \nabla^{(\alpha)}_i, \tag{2.9}$$

which is a linear map from the space of the cross sections to itself. Direct calculations give, for any cross section $r(x, \theta)$,

$$R^{(\alpha)}_{ij} r(x, \theta) = \frac{(1-\alpha^2)}{4} \{u_i E_\theta[r u_j] - u_j E_\theta[r u_i]\}. \tag{2.10}$$

Theorem 1. The fibre bundle B(M) is curvature-free for $\alpha = \pm 1$-connections. The e- and m-parallel transports, which do not depend on the curve connecting two end points θ and θ', of $r(x) \in F_\theta$ from θ to θ' are explicitly given by

$$\Pi^{(e)\,\theta'}_{\theta} r(x) = r(x) - E_{\theta'}[r(x)], \tag{2.11}$$

$$\Pi^{(m)\theta'}_{\theta} = \frac{p(x, \theta)}{p(x, \theta')}\, r(x). \tag{2.12}$$

Proof. The relation (2.10) guarantees that B(M) is curvature free for $\alpha = \pm 1$. The parallel transports (2.11) and (2.12) are obtained by solving the corresponding differential equations given from (2.8).

Two connections ∇ and ∇^* of B(M) are said to be mutually dual, if

$$\partial_i \langle r, s \rangle_\theta = \langle \nabla_i r, s \rangle_\theta + \langle r, \nabla^*_i s \rangle_\theta$$

holds for any cross sections $r(x, \theta)$ and $s(x, \theta)$. The integral expression of the duality is

$$\langle r, s \rangle_\theta = \langle \Pi_{\theta(t)\,\theta}^{\theta'} r, \Pi^*_{\theta(t)\,\theta}{}^{\theta'} s \rangle_{\theta'}, \tag{2.13}$$

where Π and Π^* are, respectively, the parallel transports from θ to θ' along $\theta(t)$ by ∇ and ∇^*.

Theorem 2. The e-connection and the m-connection are mutually dual.

Proof. The relation

$$\langle r, s \rangle_\theta = \langle \Pi^{(e)\theta'}_{\theta} r, \Pi^{(m)\theta'}_{\theta} s \rangle_{\theta'}$$

is easily proved from (2.11) and (2.12). This shows that the e-connection and m-connection are dual.

Even though B(M) is e- and m-curvature-free, the model M itself is in general curved. This curvature is measured by the imbedding curvature of T(M) in B(M), which shows how the tangent subspace $T_\theta(M)$ changes in B(M) as θ changes.

The α-imbedding curvature at θ is given by $H^{(\alpha)}_{ij}(x) \in F_\theta$, defined by

$$H^{(\alpha)}_{ij}(x) = P^N_\theta \nabla^{(\alpha)}_i u_j, \tag{2.14}$$

where P^N is the projection operator which projects a vector of F_θ to the orthogonal complement of T_θ. That is explicitly given by

$$P^N r(x) = r(x) - \langle r, u_i \rangle_\theta g^{ij} u_j,$$

where g^{ij} is the inverse of g_{ij}.

The square of the curvature is given by the tensor

$$(H^{(\alpha)}_M)^2_{ij} = \langle H^{(\alpha)}_{im}(x), H^{(\alpha)}_{jk}(x) \rangle_\theta\, g^{mk}, \tag{2.15}$$

which plays an important role in the higher order asymptotics of statistical inference. The e- and m-Riemann-Christoffel curvature tensors of M are given by

$$R^{(e)}_{ijkl} = \langle H^{(m)}_{il}, H^{(e)}_{jk} \rangle - \langle H^{(e)}_{ik}, H^{(m)}_{jl} \rangle, \tag{2.16}$$

$$R^{(m)}_{ijkl} = \langle H^{(e)}_{il}, H^{(m)}_{jk} \rangle - \langle H^{(m)}_{ik}, H^{(e)}_{jl} \rangle. \tag{2.17}$$

These are generalizations of the Gauss equation, where the $\alpha = \pm 1$ flatness of B(M) is taken into account. (It is pointed out by P. Vos and by K. Murota (personal communication) that the Gauss equation (2.14) of Amari [1985] is incorrect. The correct one is

$$R_{abcd} = B^i_a B^j_b B^k_c B^m_d R_{ijkm} - g^{\varkappa\lambda}(H_{ac\varkappa} H^*_{bd\lambda} - H_{bc\varkappa} H^*_{ad\lambda}), \tag{2.18}$$

where H and H^* are the imbedding curvature tensors with respect to the dual ∇ and ∇^*, respectively.)

3. Estimation in the presence of infinitely many nuisance parameters.

Let us consider a statistical model

$$M = \{p(x, \theta, \xi)\},$$

which includes two (vector) parameters θ and ξ. We can construct the Hilbert bundle B(M) in the same way as before, where the base space M has a coordinate system (θ, ξ). When we have interest in estimating the value of θ but do not have interest in estimating ξ, the parameter ξ is called the nuisance parameter. Let us consider the following problem. Let $x_1, x_2, \ldots$ be an infinite sequence of independent observations such that x_i is subject to the distribution $p(x, \theta, \xi_i)$, where $\Xi = (\xi_1, \xi_2, \ldots)$ is an infinite sequence of ξ's. This implies that the value of ξ changes observation by observation, while θ is common. We do not know the values ξ_i. The problem is to search for an asymptotically best estimator $\hat{\theta}_N(x_1, \ldots, x_N)$ of θ from the N observations $x_N = (x_1, \ldots, x_N)$.

We study the case when θ is a scalar parameter. The present result can easily be generalized to a vector parameter case. The nuisance parameter ξ may be a vector parameter.

We study only the following type of estimators, which are obtained by solving

$$\sum_{i=1}^{N} y(x_i, \theta) = 0. \tag{3.1}$$

This type of estimators are called M-estimators. The function $y(x, \theta)$ is called an estimating function, and (3.1) is called the estimating equation. Let us analyze the asymptotic behavior of an M-estimator derived from (3.1), where N is assumed to be sufficiently large. By expanding

$$\sum y(x_i, \hat{\theta}) = 0$$

at the true value , we have

$$\sum \{y(x_i, \theta) + \partial_\theta y(x_i, \theta)\ (\hat{\theta} - \theta)\} = o_p\ (|\hat{\theta} - \theta|^2),$$

where $\partial_\theta = \partial/\partial\theta$.

We assume that $N^{-1}\sum y(x_i, \theta)$ and $N^{-1}\sum \partial_\theta y(x_i, \theta)$ converge to their expected values

$$\langle\langle y(x, \theta) \rangle\rangle = \frac{1}{N}\sum E_{\theta, \xi_i}[y(x_i, \theta)],$$

$$\langle\langle \partial_\theta y(x, \theta) \rangle\rangle = \frac{1}{N}\sum E_{\theta, \xi_i}[\partial_\theta y(x_i, \theta)],$$

respectively, where E_{θ, ξ_i} is the expectation with respect to $p(x, \theta, \xi_i)$ and the operator $\langle\langle\ \rangle\rangle$ implies

$$\langle\langle a(x) \rangle\rangle = \frac{1}{N}\sum E_{\theta, \xi_i}[a(x_i)]. \tag{3.2}$$

This operator depends on the true θ and the sequence Ξ. We have

$$\hat{\theta} - \theta = -\frac{\sum y}{\sum \partial_\theta y} + \text{higher order terms.}$$

Therefore, the estimator is consistent for any sequence Ξ of the nuisance parameter, iff $\langle\langle y \rangle\rangle = 0$ and $\langle\langle \partial_\theta y \rangle\rangle \neq 0$ for any Ξ. This is equivalent to the conditions

$$E_{\theta,\xi}[y(x, \theta)] = 0, \tag{3.3}$$

$$E_{\theta,\xi}[\partial_\theta y(x, \theta)] \neq 0, \tag{3.4}$$

for any ξ. This implies a consistent y belongs to $F_{\theta,\xi}$, the fibre space at (θ, ξ) for any ξ.

Under this condition, we define the asymptotic variance with respect to Ξ of the estimator of $\hat{\theta}$, or the estimating function y which gives $\hat{\theta}$, by

$$AV[y, \Xi] = \lim_{N\to\infty} N\, E[(\hat{\theta} - \theta)^2] .$$

When the central limit theorem holds for $(1/\sqrt{N}) \sum y(x_i, \theta)$, we have

$$AV[y, \Xi] = \frac{\langle\langle y^2 \rangle\rangle}{\langle\langle \partial_\theta y \rangle\rangle^2}. \tag{3.5}$$

An estimator $\hat{\theta}$ or its estimating function y is said to be asymptotically universally optimal, when its asymptotic variance $AV[y, \Xi]$ satisfies

$$AV[y, \Xi] \leqq AV[\bar{y}, \Xi]$$

for any estimating function $\bar{y}$ for all Ξ. We show that the fibre bundle approach is very useful for solving the following problems:

1) When does a consistent estimating function $y(x, \theta)$, which gives a consistent estimator for any Ξ, exist? How can we get one?

2) When does the universally optimal estimating function exists? How can we obtain it, when it exists?

Let u and $v_\varkappa$ be the tangent vectors along the coordinate curves θ and $\xi^\varkappa$ (the $\varkappa$-th component of ξ), respectively,

$$u = \partial_\theta \log p(x, \theta, \xi), \quad v_\varkappa = (\partial/\partial\xi^\varkappa) \log p(x, \theta, \xi). \tag{3.6}$$

Let M be a submanifold of M on which the value of θ is fixed to θ, i.e., M_θ consists of those points in M whose θ-coordinate is θ. The fibre space $F_{\theta,\xi}$ includes the subspace $T^N_{\theta,\xi}$ spanned by the vectors $v_\varkappa(x, \theta, \xi)$ which designates the directions of the nuisance parameters changing. The tangent space $T_{\theta,\xi}$ is spanned by u and $v_\varkappa$. Let us consider the subspace $F^N_{\theta,\xi}$ spanned by all the m-parallel transports along M from any ξ' to ξ of the vectors $v_\varkappa(x, \theta, \xi')$,

$$\Pi^{(m)\xi}_{\xi'} v_\varkappa(x, \theta, \xi').$$

Since the fibre bundle is m-flat, we do not specify the curves of parallel transports. The subspace $F^N_{\theta,\xi}$ is written as

$$F^N_{\theta,\xi} = \Pi^{(m)}_{M_\theta} T^N_{\theta,\xi} . \tag{3.7}$$

We define similarily the tangential subspace

$$F^T_{\theta,\xi} = \Pi^{(m)}_{M_\theta} T_{\theta,\xi}, \tag{3.8}$$

which includes $F^N_{\theta,\xi}$. Obviously, $F^N_{\theta,\xi}$ includes $T^N_{\theta,\xi}$ and $F^T_{\theta,\xi}$ includes $T_{\theta,\xi}$. Let $F^A_{\theta,\xi}$ be the orthogonal complement of $F^T_{\theta,\xi}$ in $F_{\theta,\xi}$. Then, we have the orthogonal decomposition

$$F_{\theta,\xi} = F^T_{\theta,\xi} \oplus F^I_{\theta,\xi}$$

so that

$$F_{\theta,\xi} = F^I_{\theta,\xi} \oplus F^A_{\theta,\xi} \oplus F^N_{\theta,\xi} . \tag{3.9}$$

We call $F^I_{\theta,\xi}$ the information subspace. Any element $r(x) \in F_{\theta,\xi}$ can be decomposed uniquely into

$$r(x) = r^I(x, \theta, \xi) + r^A(x, \theta, \xi) + r^N(x, \theta, \xi), \tag{3.10}$$

where $r^I \in F^I_{\theta,\xi}$, $r^A \in F^A_{\theta,\xi}$ and $r^N \in F^N_{\theta,\xi}$ are mutually orthogonal. Let P^I be the projection operator to $F^I_{\theta,\xi}$,

$$P^I r = r^I.$$

A consistent estimating function $y(x, \theta)$ satisfies

$$E_{\theta,\xi}\ [y(x, \theta)] = 0 \tag{3.11}$$

for all ξ. Moreover, it does not depend on ξ, so that

$$\nabla^{(e)}_\kappa y = \partial_\kappa y - E_{\theta,\xi}\ [\partial_\kappa y] = 0,$$

where $\partial_\kappa = \partial/\partial\xi^\kappa$. This implies that $y(x, \theta)$ is an e-parallel vector field on the submanifolds M_θ,

$$\Pi^{(e)\xi}_{\xi'}\ y(x, \theta) = y(x, \theta). \tag{3.12}$$

We first prove that a consistent estimating function $y(x, \theta)$ does not have the nuisance subspace component. By differentiating (3.11) with respect to ξ and by putting $\xi = \xi'$, we have

$$\frac{\partial}{\partial\xi^\kappa} E_{\theta,\xi}\ [y(x, \theta)] = \left\langle y,\ v_\kappa(x, \theta, \xi') \right\rangle_{\xi'} = 0.$$

From

$$0 = \left\langle y,\ v_\kappa \right\rangle_{\xi'} = \left\langle \Pi^{(e)\xi}_{\xi'}\ y,\ \Pi^{(m)\xi}_{\xi'}\ v_\kappa(x, \theta, \xi') \right\rangle_\xi$$
$$= \left\langle y, \Pi^{(m)\xi}_{\xi'} v_\kappa \right\rangle_\xi,$$

we see that y is orthogonal to $F^N_{\theta,\xi}$. Hence, we have the following decomposition

$$y(x, \theta) = y^I(x, \theta, \xi) + y^A(x, \theta, \xi),$$

where y^I and y^A may depend on ξ.

Conversely, any $y(x, \theta) \in F^I_{\theta,\xi}$ at a point (θ, ξ) satisfies

$$E_{\theta,\xi'}\ [y(x, \theta)] = 0$$

at all ξ'. Hence, it is a consistent estimating function, if (3.4) is satisfied. To check this last condition, we consider the information part of the score function

$$u(x, \theta, \xi) = \partial_\theta \log p(x, \theta, \xi)$$

given by

$$u^I(x, \theta, \xi) = P^I u(x, \theta, \xi). \tag{3.13}$$

We call u^I the information score function. A consistent estimating function $y(x, \theta)$ is decomposed as

$$y(x, \theta) = c(\theta, \xi)\ u^I(x, \theta, \xi) + y^*(x, \theta, \xi) + y^A(x, \theta, \xi), \tag{3.14}$$

where y^* is a vector in $F^I_{\theta,\xi}$ which is orthogonal to u^I. We

have, by differentiating (3.11) with respect to θ,

$$E_{\theta,\xi}[\partial_\theta y] = -\langle y, u\rangle = -c(\theta, \xi)\langle u^I, u^I\rangle,$$

because of

$$\langle u, y^*\rangle = \langle u^I, y^*\rangle = 0,$$

$$\langle u, y^A\rangle = \langle u^I, y^A\rangle = 0.$$

Therefore, we have the following theorem, which elucidates the structure of consistent estimating functions.

Theorem 3. A consistent estimating function exists if and only if the information score function is not null. Any consistent estimating function can be decomposed as

$$y(x, \theta) = c(\theta, \xi)\, u^I(x, \theta, \xi) + y^*(x, \theta, \xi) + y^A(x, \theta, \xi).$$

Conversely, at any fixed point ξ_0.

$$y(x, \theta) = c\, u^I(x, \theta, \xi_0) + y^* + y^A, \quad c \neq 0,$$

$y^* \in F^I_{\theta,\xi}$, $y^A \in F^A$, gives a consistent estimator.

We next search for the optimal estimator, provided the information score function u^I is not null. By using the decomposition (3.14) and the relation (3.5), the asymptotic variance given by a consistent $y(x, \theta)$ can be written as

$$AV[y, \Xi] = \frac{\ll (cu^I)^2 \gg + \ll (y^*)^2 \gg + \ll (y^A)^2 \gg}{\ll c(u^I)^2 \gg^2}.$$

Let us consider a special sequence $\Xi_0 = (\xi_0, \xi_0, \xi_0, \ldots)$. Then, the estimating function

$$y(x, \theta) = u^I(x, \theta, \xi_0)$$

has the minimum asymptotic variance for this special sequence Ξ_0. When $u^I(x, \theta, \xi)$ does not depend on ξ, it is easy to prove that

$$y(x, \theta) = u^I(x, \theta)$$

is indeed the optimal for any Ξ.

Theorem 4. When and only when $u^I(x, \theta, \xi)$ is not null and does not depend on ξ, the universally optimal estimating function exists and is given by $u^I(x, \theta)$.

4. Fibre bundles of semi-parametric statistical models

A statistical model

$$M = \{f(x, \theta, Z)\}$$

is called a semi-parametric model, when the probability density function $f(x, \theta, Z)$ is specified by a finite-dimensional parameter θ of interest and a nuisance parameter Z of functional degrees of freedom (Begun et al. [1982]). One typical example is given by the previous problem, if we assume that the nuisance parameter ξ is subject to an unknown probability distribution $Z(\xi)$ and that ξ_i is the i-th independent realization from $Z(\xi)$. In this case, every x_i is subject to the same probability distribution

$$f(x, \theta, Z) = \int p(x, \theta, \xi)\, dZ(\xi). \tag{4.1}$$

This is a typical example of the semi-parametric model.

Another example is the location model with an unknown distribution function $Z(x)$,

$$f(x, \theta, Z) = Z(x - \theta), \tag{4.2}$$

where $Z(x)$ satisfies

$$\int Z(x)dx = 1,$$

$$\int xZ(x)\, dx = 0,$$

and some other regularity conditions.

The fibre bundle theory can be effectively applied to these semi-parametric models, although its mathematical foundation might not be easy. The present section studies mainly the case (4.1) with infinitely many nuisance parameters. By this approach, even when there exist no universally optimal estimating functions, we can always obtain the optimal or semi-optimal estimating function, by choosing the estimating function $y(x, \theta)$ adaptively depending on $x_1, x_2, \cdots$.

Let us consider the fibre bundle over the statistical model M such that the fibre space at point (θ, Z) is

$$F_{\theta, Z} = \{r(x) \mid E_{\theta, Z}[r] = 0, E[r^2] < \infty \text{ for any distributions on M}\}.$$

Let $T^U_{\theta, Z}$ be the linear space defined at (θ, Z) spanned by $\partial_\theta \log f(x, \theta, Z)$ and let $T^N_{\theta, Z}$ be the linear space defined at (θ, Z) spanned by $\partial_Z \log p(x, \theta, Z)$, where ∂_Z denotes the Frechét derivative with respect to $Z(\xi)$. In the case of (4.1), $T^N_{\theta, Z}$ is spanned by

$$\frac{\partial_\xi p(x, \theta, \xi)}{f(x, \theta, Z)},$$

so that an element $r(x) \in T^N_{\theta, Z}$ is written as the linear combination,

$$r(x) = \int \frac{\partial_\xi p(x, \theta, \xi)}{f(x, \theta, Z)}\, dA(\xi)$$

by using the Stieltjes integral by an arbitrary function $A(\xi)$.

Let us define the tangential subspace $F^T_{\theta, Z}$ at (θ, Z) by one spanned by all the m-parallel transports from any Z' in Z over M of the vectors of $T^U_{\theta, Z'}$ and $T^N_{\theta, Z'}$. The m-parallel transport of $r(x)$ from Z' is Z is defined as before by

$$\prod^{(m)}{}^{Z}_{Z'} r = \frac{f(x, \theta, Z')}{f(x, \theta, Z)}\, r. \tag{4.3}$$

The nuisance subspace $F^N_{\theta, Z}$ is one spanned by all the m-parallel transports from any Z' to Z over M of the vectors in $T^N_{\theta, Z'}$.

The fibre space is decomposed into

$$F_{\theta, Z} = F^T_{\theta, Z} \oplus F^A_{\theta, Z} = F^I_{\theta, Z} \oplus F^A_{\theta, Z} \oplus F^N_{\theta, Z} \tag{4.4}$$

as before, where the ancillary subspace $F^A_{\theta, Z}$ is the orthogonal complement of $F^T_{\theta, Z}$, and the information subspace $F^I_{\theta, Z}$ is the orthogonal complement in $F^T_{\theta, Z}$ of $F^N_{\theta, Z}$.

We again treat a consistent estimating function $y(x, \theta)$ with the estimating equation

$$\sum y(x_i, \theta) = 0.$$

It is consistent, when and only when

$$E_{\theta, Z}[y(x, \theta)] = 0, \quad E_{\theta, Z}[\partial_\theta y] \neq 0. \tag{4.5}$$

The asymptotic variance is given by

$$AV[y] = \frac{E_{\theta, Z}[y^2]}{\{E_{\theta, Z}[\partial_\theta y]\}^2}, \tag{4.6}$$

where (θ, Z) are the true parameters of the distribution.

The theory can be constructed analogously to the previous section. The difference is that the nuisance parameter Z is infinite dimensional, but it is fixed. We study the following problems.

1) When does a consisting estimating function exist?

2) When does an M-estimator attain the Cramer-Rao bound given by the Fisher information? In other words, what is the amount of loss of information by considering the class of M-estimators?

3) How can one obtain the optimal estimating function?

From (4.5), by the same reasoning as before, we can decompose a consisting estimating function as

$$y(x, \theta) = c(\theta, Z)\, u^I(x, \theta, Z) + y^* + y^A,$$

where u^I is the information score function, $y^* \in F^I_{\theta, Z}$ orthogonal to u^I, and $y^A \in F^A_{\theta, Z}$. The relation

$$E_{\theta, Z}[\partial_\theta y] = c(\theta, Z)\langle u^I, u^I \rangle \tag{4.7}$$

holds. Therefore, the first question is answered.

Theorem 5. When and only when the information score u^I is not zero, there exists a consistent estimator.

The asymptotic variance of a consistent y is given by

$$AV[y] = \frac{c^2 E[(u^I)^2] + E[(y^*)^2] + E[(y^A)^2]}{c^2 E[(u^I)^2]^2}.$$

Therefore, when and only when $u^I(x, \theta, Z)$ can be written as

$$u^I(x, \theta, Z) = c(\theta, Z)\, u^I_0(x, \theta), \tag{4.8}$$

there exists a universally optimal estimating function not depending on Z, i.e., optimal for any Z. It is given by

$$y(x, \theta) = u^I_0(x, \theta).$$

When $u^I(x, \theta, Z)$ is not of the form (4.8), there does not exist the universally optimal function. However, if we know that the true nuisance parameter is Z,

$$y(x, \theta) = u^I(x, \theta, Z)$$

gives the best estimating function in this specific case. Therefore, if we can obtain an estimator $\hat{Z}$ of Z,

$$y(x, \theta) = u^I(x, \theta, \hat{Z}) \tag{4.9}$$

can be used as an approximation of the optimal estimator.

It is a merit of an M-estimator that it is robust in the sense that, if we use a bad approximation $\hat{Z}$ of Z, the estimator is still consistent.

Now we search for the efficiency of the optimal estimating function. The asymptotic variance of the optimal M-estimator is given by

$$AV^{*}[Z] = \frac{1}{E[(u^{I})^{2}]}, \tag{4.10}$$

provided we can use the true Z. Let

$$g_{M}(\theta, Z) = E_{\theta, Z}[(u^{I})^{2}]. \tag{4.11}$$

Then, we have

$$AV[y] \geq g_{M}^{-1},$$

and the equality holds when and only when $y(x, \theta) = u^{I}(x, \theta, Z)$. We call g_{M} the amount of information of the M-estimation.

The Fisher information g is given by the expectation of the square of the orthogonalized score function, when there exists a nuisance parameter. This is given by the projection of the score function u to the subspace $T^{O}_{\theta,Z}$ orthogonal to the $T^{N}_{\theta, Z}$ in the tangent space $T_{\theta, Z}$.

$$g(\theta, Z) = E_{\theta, Z}[w^{2}], \tag{4.12}$$

$$w = P^{O} u(x, \theta, Z) \tag{4.13}$$

where P^{O} is the projection to the orthogonal complement of

$T^{N}_{\theta, Z}$ in $T_{\theta, Z}$.

Since $F^{I}_{\theta, Z}$ is orthogonal to $T^{N}_{\theta, Z}$, we have

$$P^{I} w = u^{I}. \tag{4.14}$$

Therefore,

$$g(\theta, Z) \geq g_{M}(\theta, Z)$$

holds. When the equality holds, no loss of information is caused if we use the best M-estimator. This occurs when and only when

$$F^{N}_{\theta, Z} = T^{N}_{\theta, Z}$$

holds.

A submanifold M_{θ} is said to be mixture flat or shortly m-flat, when its tangent directions are invariant under the m-parallel transports,

$$\Pi^{(m)}_{Z', Z} T^{N}_{\theta, Z'} = T^{N}_{\theta, Z}. \tag{4.15}$$

Obviously, when and only when M is m-flat, (4.15) holds and there is no loss of information.

Theorem 6. The optimal estimating function is fully informative, when and only when M is m-flat, i.e., $\{M_{\theta}\}$ defines a m-geodesic foliation.

It is important to see that M_{θ}'s in the two models (4.1) and (4.2) are m-flat. Hence, there is no loss of information.

Let (θ, Z_0) be the true parameter of distribution. When we use

$$y(x, \theta) = u^I(x, \theta, Z),$$

its asymptotic variance is given by

$$AV[y] = \frac{E_{\theta, Z_0}[u^I(x, \theta, Z)^2]}{\left\{E_{\theta, Z_0}[u^I(x, \theta, Z)\, u^I(x, \theta, Z_0)]\right\}^2}. \quad (4.16)$$

Let $\hat{Z}_N(\theta)$ be an adaptive estimator of Z where θ is assumed. If $\hat{Z}_N$ is a consistent estimator in the weak sense,

$$y(x, \theta) = u^I(x, \theta, \hat{Z}_N(\theta)) \quad (4.17)$$

gives the asymptotically optimal estimating function. However, this procedure is not easy. We can use instead a parametric model for Z

$$Z(\xi) = Z(\xi, \eta)$$

where η is a finite-dimensional parameter specifying the distribution $Z(\xi)$. Then, we can easily obtain the estimator $\hat{\eta}(\theta)$ of η. We can use this one to obtain.

$$y(x, \theta) = u^I(x, \theta, Z(\xi, \hat{\eta}(\theta))). \quad (4.18)$$

Even when the parametric model $Z(\xi, \eta)$ is incorrect, the above gives a good consistent estimator. This elucidates the method proposed by Lindsay [1985].

References

Amari, S. [1982a]. Differential geometry of curved exponential families --- curvatures and information loss. Ann. Statist., 10, 357-387

Amari, S. [1982b]. Geometrical theory of asymptotic ancillarity and conditional inference. Biometrika, 69, 1-17

Amari, S. [1983]. Comparisons of asymptotically efficient tests in terms of geometry of statistical structures. Bull. Int. Statist. Inst., Proc. 44th Session, Book 2, 1190-1206

Amari, S. [1985]. Differential-Geometrical Methods in Statistics. Springer Lecture Notes in Statistics, 28, Springer

Amari, S. [1987a]. Differential geometry of Statistics --- Towards New Developments. in "Differential Geometry in Statistical Inference". IMS Lecture Notes, vol. 10

Amari, S. [1987b]. Differential geometry of a parametric family of invertible linear systems --- Riemannian metric, dual affine connections and divergence. Math. Systems Theory, in press

Amari, S. and Han, T.S. [1987]. Statistical inference under multi-terminal rate ristrictions --- a differential geometrical approach. IEEE Trans. Inf. Theor., to appear

Amari, S. and Kumon, M. [1983]. Differential geometry of Edgeworth expansions in curved exponential family. Ann. Inst. Statist. Math., 35A, 1-24

Amari, S. and Kumon, M. [1985]. Optimal estimation in the presence of infinitely many nuisance parameters --- geometry of estimating functions. METR 85-2, Univ. Tokyo

Barndorff-Nielsen, O. [1987]. Differential and integral geometry in statistical inference. in "Differential Geometry in Statistical Inference". IMS L. N., vol. 10

Barndorff-Nielsen, O.E., Cox, D.R. and Reid, N. [1986]. The role of differential geometry in statistical inference, Int. Statist. Review, 54, 83-96

Begun, J.M., Hall, W.J., Huang, W.-M. and Wellner, J.A. [1983]. Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist., 11, 432-452

Bickel, P.J. [1982]. On adaptive estimation. Ann. Statist., 10, 647-671

Campbell, L.L. [1985]. The relation between information theory and the differential geometry approach to statistics, Inf. Sci., 35, 199-210

Chentsov, N.N. [1972]. Statistical Decision and Rules and Optimal Inference (in Russian). Nauka, Moscow, translated in English (1982), AMS, Rhode Island

Eguchi, S. [1983]. Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Statist., 11, 793-803

Kass R.E. [1987]. Introduction to "Differential Geometry in Statistical Inference. IMS L. N., vol. 10

Kumon, M. and Amari, S. [1983]. Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference. Proc. Roy. Soc. London, A387, 429-458

Kumon, M. and Amari, S. [1984]. Estimation of structural parameter in the presence of a large number of nuisance parameters. Biometrika, 71, 445-459

Kumon, M. and Amari, S. [1985]. Differential geometry of testing hypothesis: a higher order asymptotic theory in multiparameter curved exponential family, METR 85-1, Univ. Tokyo

Lauritzen, S. [1987]. Some differential geometrical notions and their use in statistical theory. in "Differential Geometry in Statistical Inference". IMS L. N. vol. 10

Lindsay, B.G. [1985]. Using empirical Bayes inference for increased efficiency. Ann. Statist., 13, 914- 931

Mitchell, A. and Krzanowski, W.J. [1985]. The Mahalanobis distance and elliptic distributions. Biometrika, 72, 464-467

Nagaoka, H. and Amari, S. [1982]. Differential geometry of smooth families of probability distributions, METR 82-7, Univ. Tokyo

Picard, D. [1987]. Invariance properties of Fisher-Rao metric and Chentsov-Amari connections. Prepubl. Univ. Pari Sad, No. 425

SYSTEMS OF CONNECTIONS FOR PARAMETRIC MODELS

C.T.J.Dodson
Department of Mathematics,
University of Lancaster.

ABSTRACT

The recently introduced differential geometric parametric models in statistical theory are shown to be examples of systems of connections on fibred manifolds. Moreover, they are subsystems of large natural systems of potential statistical importance. These unifying features can be exploited because on every system of connections there is a canonical connection which has the universal property that every connection for the parametric model is a pullback of the canonical connection. By this means the various models can be studied geometrically and globally through embeddings in spaces of connections. In particular, the notion of stability of geometrical properties under variation of the connection can be formulated in a precise way and some known results, for example, concerning incompleteness, may have relevance for statistical theory.

1. INTRODUCTION

A connection is arguably the most important single entity in differential geometry and it is natural to try to isolate its essential character. It turns out that this character is a differential operator which controls the lifting of differential equations from one space to a higher space, typically from a tangent space to its tangent space. Then it is of interest to know precisely what is the least structure needed to support the notion of a connection, and what is its most economical definition. Both questions appear to have been answered during the last thirty years and the answers continue to generate considerable research activity in geometry and in its applications to physical field theory. In this paper is given the basis for further applications, in the theory of parametric statistical models. The motivation here is the same as that for the purely geometrical developments and their physical applications, simply that there are important situations when it is advantageous to study a whole family of connections on a given space. Systems of connections give a way to select finite-dimensional subspaces of the infinite-dimensional space of all connections on a space.

2. SYSTEMS OF CONNECTIONS

The most commonly encountered situation is that of a linear connection on a manifold M ; this is a connection on the tangent bundle TM $\twoheadrightarrow$ M, or, equivalently, an invariant connection on LM $\twoheadrightarrow$ M the principal bundle of linear frames (i.e. ordered bases for tangent spaces) on M. The standard reference work is Kobayashi and Nomizu [13]; a brief descriptive account of connection geometry is given elsewhere in these Proceedings cf. [11].

In geometry, in applications to physics and very likely also in applications to statistical theory, more general situations arise. The least structure needed to support the notion of a connection is a fibred manifold and then the appropriate notion is of a section of its first jet bundle; we take these ideas in turn. For

an introduction to modern differential geometry with many examples see Dodson and Poston [11]. For an account of some of the most recent advances in connection theory using Lie algebroids, see Mackenzie [14], it contains a wealth of new ideas and is largely self-contained.

2.1 Sections of a Fibred Manifold

A **fibred manifold** is a (smooth) surjective submersion $p:E \twoheadrightarrow B$, so rank $p = \dim B$ everywhere.

Example $E = \mathbb{R}^2-\{0\}$, $B = \mathbb{R}$ with $p:E \twoheadrightarrow B : (x,y) \to x$,is a fibred manifold (but not a fibre bundle).

A **local section** of $p:E \twoheadrightarrow B$ over open set $U \subset B$ is a map

$$s : U \to E \text{ satisfying } p \circ s = 1_U$$

Example $s : \mathbb{R} \to \mathbb{R}^2-\{0\} : x \to (x,e^x)$.

Throughout we shall be concerned only with smooth (C^∞) maps.

Two local sections $s, t : U \to E$ of $p:E \twoheadrightarrow B$ are said to **agree up to first derivative** at $x \in U$ if $s(x) = t(x)$ and $s'(x) = t'(x)$ where

$$s'(x) : T_xB \to T_{s(x)}E : V^\lambda \to V^\lambda \frac{\partial s^i}{\partial x^\lambda}$$

is the derivative map (its components form the Jacobian matrix) of the section s and T_xB is the tangent space to B at x.

Here and elsewhere we use the summation convention.

2.2 First Jet Bundle

The **first jet bundle** of $E \twoheadrightarrow B$ is the fibre bundle $\eta : JE \twoheadrightarrow E : j_xs \to s(x)$ where j_xs denotes the equivalence class of local sections of $E \twoheadrightarrow B$ which agree up to first derivative at x; since they agree up to first derivative, they define the same linear map $T_xB \to T_{s(x)}E$, which we may also denote by j_xs.

Example Given the fibred manifold $p:E \twoheadrightarrow B$ where $p:\mathbb{R}^3 - [0] \to \mathbb{R}^2$

$: (x^1,x^2,x^3) \to (x^1,x^2)$ then a typical element of JE consists of : a point $y = (x^1,x^2,x^3) \in E$, and a matrix $[s^i_\lambda]$ at $p(y) = (x^1,x^2) \in B$ which we view as the coordinates of the linear map on tangent spaces

$$j_x s : T_{p(y)}B \to T_yE \quad : \quad \begin{bmatrix} v^1 \\ v^2 \end{bmatrix}_{p(y)} \to \begin{bmatrix} v^1 s_1^1 + v^2 s_2^1 \\ v^1 s_1^2 + v^2 s_2^2 \\ v^1 s_1^3 + v^2 s_2^3 \end{bmatrix}_y = (v^\lambda s^i_\lambda).$$

2.3 Connections

A **connection** on the fibred manifold $E \rightarrowtail B$ is a section $\Gamma : E \to JE$ of its first jet bundle.

Example A typical connection on the fibred manifold in Example (iii) is given by $\Gamma : \mathbb{R}^3\text{-}\{0\} \to J(\mathbb{R}^3\text{-}\{0\}) : y \to \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \Gamma^3_1 & \Gamma^3_2 \end{bmatrix}$

with components Γ^3_1, Γ^3_2

depending smoothly on position $p(y) = x \in B$.

Clearly, an element of JE can be viewed as a linear lift from the tangent space to B at p(y) up to the tangent space to E at y. Formally, we have that JE is an affine subbundle of the vector bundle $T^*B \otimes_B TE$. Then a connection Γ is a vector 1-form projectable onto the identity 1_{TB} viewed as the section $dx^\lambda \otimes \partial_\lambda$ of $T^*B \otimes_B TE$:

$$\Gamma \quad : E \to JE \subset T^*B \otimes_{TB} TE$$

$$: (x^\lambda,z^i) \to ((x^\lambda,z^i),[1_{TB},\Gamma^i_\lambda])$$

Briefly, $\Gamma = dx^\lambda \otimes \partial_\lambda + \Gamma^i_\lambda \; dx^\lambda \otimes \partial_i$.

2.4 Systems of Connections

A **system of connections** on a fibred manifold $E \rightarrowtail B$ consists of a pair (C,ξ) where $P_C : C \rightarrowtail B$ is a fibred manifold and $\xi : C \times_B E \to JE$ is a fibred morphism over E. We shall call C the **space of connections** of the system. Evidently, each section of $C \rightarrowtail B$ fixes a connection on $E \rightarrowtail B$.

This notion was introduced by Mangiarotti and Modugno [14]; a comprehensive study is available in Modugno [17]. The important cases for us are given in the following two equivalent formulations of linear connections, as vector connections on the tangent bundle and as principal connections on the frame bundle (cf. Del Riego and Dodson [8]).

Example (a) Tangent bundle system : $C_T \times TM \to JTM$

The system of all linear connections on a manifold M has one representation on the tangent bundle

$$E = TM \overset{\pi_T}{\rightarrowtail} M = B$$

with system space

$$C_T = \{v \in T^*M \otimes_M JTM \mid v \text{ projects onto } 1_{TM}\}.$$

Here we view 1_{TM} as a section of $T^*M \otimes TM$, which is a subbundle of $T^*M \otimes TTM$, with local expression $dx^\lambda \otimes \partial_\lambda$.

The fibred morphism for this system is given by

$$\xi_T : C_t \times_m TM \to JTM \quad T^*M \otimes_{TM} TTM$$

$$: (x^\lambda, v^\lambda_{\mu\vartheta}, y^\lambda) \to [(x^\lambda,X^\lambda) \to (x^\lambda,y^\lambda,X^\lambda,y^\mu v^\lambda_{\mu\vartheta} X^\mu)].$$

The local expression of the induced section of $T^*M \otimes_{TM} TTM$ is

$$\xi_T(x) = dx^\lambda \otimes (\partial_\lambda - y^\mu v^\nu_{\mu\lambda} \partial_\nu).$$

To each section of $C_T \twoheadrightarrow M$, such as

$$\bar{\Gamma} : M \to C_T : (x^\lambda) \to (x^\lambda, \gamma_{\mu\vartheta})$$

there corresponds the unique linear connection

$$\Gamma = \xi_\cdot(\bar{\Gamma} \circ \pi, 1_{TM}) \text{ with Christoffel symbols } \Gamma^\lambda_{\mu\vartheta}$$

Example (b) Frame bundle system : $C_F \times FM \to JFM$

A second way to view the system of all linear connections on a manifold M arises because a linear connection is actually a principal (i.e. group invariant) connection on a principal bundle. This latter is

$$E = FM \overset{\pi_F}{\twoheadrightarrow} M = FM/G$$

consisting of linear frames (ordered bases for tangent spaces) with structure group the general linear group, $G = G\ell(n)$; it is discussed at length in Kobayashi and Nomizu [13].

Here the system space is

$$C_F = JFM/G \subset T^*M \otimes_{TM} TFM/G,$$

consisting of G-invariant jets.

The system morphism is

$$\xi_F : C_F \times FM \to JFM \subset T^*M \otimes_{TM} TFM$$

$$([js_x], b) \to [T_xM \to T_bFM].$$

Its coordinate expression as a TFM-valued 1-form is

$$\xi_F(x) = dx^\lambda \otimes (\partial_\lambda - X^\mu \, \partial_\mu s^\lambda_\nu) \tilde{\partial}^\lambda_\nu$$

where $\tilde{\partial}^\lambda_\nu$ is the natural base on the vertical fibre of T_bFM induced by coordinates (b^λ_ν) on FM, that is

$$\tilde{\partial}^\lambda_\nu = \frac{\partial}{\partial b^\lambda_\nu}, \quad \text{corresponding to } \partial_i \text{ in 2.3.}$$

To each section of $C_F \twoheadrightarrow M$ that is projectable onto 1_{TM}, such as,

$$\Gamma : M \to C_F : (x^\lambda) \to (x^\lambda, [j\gamma_x]) \text{ with } \Gamma^\lambda_{\mu\nu} = \partial_\mu s^\lambda_\nu,$$

there corresponds the unique linear connection

$$\Gamma = \xi_F \circ (\Gamma \circ \pi_F, 1_{FM}), \text{ with Christoffel symbols } \Gamma^\lambda_{\mu\nu}.$$

There is a direct correspondence between the two systems representing linear connections by means of the linear diffeomorphism

$$\rho : TFM/G \to JTM \subset T^*M \otimes_{TM} TTM$$

$$: (x^\lambda, y^\lambda, B^\lambda_\mu) \to [(x^\lambda, X^\lambda) \to (x^\lambda, y^\lambda, X^\lambda, X^\mu B^\lambda_\mu)].$$

For it induces a smooth bijective correspondence of sections

$$\mathrm{Sec}(C_F/M) \leftrightarrow \mathrm{Sec}(C_T/M)$$

$$\Gamma \leftrightarrow (1_{T^*M} \otimes \rho) \circ \Gamma = \bar{\Gamma}.$$

2.5 The System of α-connections

Now we can identify the α-connections of Amari-Chentsov [6] (cf. Dawid [7], Amari [1], [2] and Barndorff-Nielsen, Cox and Reid [4]) as a subsystem of a system of linear connections. The original workers used the case of a parameter space that is an open submanifold of the Euclidean space $\mathbb{R}^d$. However, this can be treated as one coordinate patch on a parameter space that is a manifold and the theory generalizes in an obvious way. There may be situations where it is convenient to have a compact parameter space, perhaps with some algebraic structure, e.g. as a Lie group. For field theory in physics there is often advantage in studying solutions on compactified $\mathbb{R}^3$, i.e. S^3, for analytical convenience.

Take a smooth d-manifold Ω as the parameter space of a d-parameter family $M = \{p(\,;w) \mid w \in \Omega\}$ of probability density functions. Evidently M is also a smooth d-manifold. On M is induced a riemannian structure $g = i_{rs}\, dw^r \otimes dw^s$, the expected information metric, with

$$i_{rs} = -E\left[\frac{\partial^2 \ell}{\partial w^r dw^s}\right] = E\left\{\frac{\partial \ell}{\partial w^r}\,\frac{\partial \ell}{\partial w^s}\right\}$$

where $\ell = \log p(y;w)$ is the log likelihood function. This metric determines a unique torsion-free connection, the Levi-Civita connection, with coefficients $({}^{0}\Gamma^{t}_{rs})$. The α-connections consist of the 1-parameter family of linear connections $\{ {}^{\alpha}\Gamma \mid \alpha \in R\}$ with coefficients

$$ {}^{\alpha}\Gamma^{t}_{rs} = {}^{0}\Gamma^{t}_{rs} - \tfrac{1}{2}\,\alpha\, i^{kt}\, E\left\{\frac{\partial \ell}{\partial w^{r}}\frac{\partial \ell}{\partial w^{s}}\frac{\partial \ell}{\partial w^{k}}\right\} $$

Here (i^{kt}) represents the dual metric to (i_{kt}) and the expectation factor is the skewness tensor; what is being exploited is the geometrical fact that the difference of two connections is tensorial. Now consider the system

$$(M \times \dot{R}) \times FM \to JFM : (\alpha,b) \to {}^{\alpha}\Gamma(b)$$

where $\dot{R}$ is the real line with the discrete topology. So the system space $C = M \times \dot{R}$ consists of a stack of copies of M. Then any $\Gamma \in \mathrm{Sec}\,(C/M)$ is a constant real function on M, so defining precisely one α-connection. Hence the above system of connections coincides with the original Amari-Chentsov family of connections.

2.6 The Conformal System of Variable α-connections

There is a natural enlargement of the system of α-connections to the system

$$(M\times R)\times FM \to JFM : (\alpha,b) \to {}^{\alpha}\Gamma(b)$$

where $\mathcal{C} = M\times R$ is the usual product manifold of M with the standard real line. Then any $\Gamma \in \mathrm{Sec}\,(\mathcal{C}/M)$ is a smooth real function on M and hence determines a 'variable-α' α-connection.

A further enlargement is to the system $(M\times R^2) \times FM \to JFM : (\alpha,\varphi,b) \to {}^{\varphi}\Gamma(b)$ where ${}^{\alpha}\Gamma$ is the α-connection arising from the Levi-Civita connection of the metric $e^{\varphi}i$. In this case the system space is $C = M\times R^2$ and $\Gamma\in\mathrm{Sec}\,(C/M)$ yields a pair (α,φ) of smooth real functions on M. One yields the variable α, the other yields the positive conformal factor e^{φ} which multiplies the expected information metric. In coordinates the metric so formed is given by

$$ i^{c}_{rs} = e^{\varphi} i_{rs}. $$

The effect of the conformal factor is to allow metric weighting to be applied differentially over the parameter space; the choice of uniform weighting, $\varphi = 0$ everywhere, recovers the original expected information metric structure. There will be a parallel result to 2.6 for a system arising from the observed information metric [3].

2.7 Conformal stability of incompleteness about an information metric

Every connected riemannian manifold (M,g) is a metric space with distance function the infimum of curve length between pairs of points. The Hopf-Rinow theorem states that this metric space is Cauchy complete if and only if (M,g) is geodesically complete, that is if and only if all geodesics admit extension to infinite parameter values. Using systems of connections we can show that if a parametric statistical model is geodesically incomplete then this incompleteness persists for a 1-parameter family of metrics conformal to the expected information metric. This is a direct consequence of the construction given in Canarutto and Dodson [5] because b-incompleteness coincides with geodesic incompleteness in the riemannian case (cf. Dodson [9], Del Riego and Dodson [8]).

From its definition it is clear that each α-connection is torsion-free.

Now, connections of this kind are in smooth bijective correspondence with sprays. Namely, the unique spray of a linear torsion-free connection Γ on M is precisely the horizontal lift map from TM to TTM and each spray is a tangent vector field on TM having as image the horizontal distribution of a unique torsion-free linear connection. A discussion of properties of sprays is given in [8]. Williams [18] showed that the topology, induced on the torsion-free subspace by the Whitney topology on the space of linear connections on M, is discrete. His difference-tensor-topology on the space of connections seems more appropriate and

with respect to it the symmetrization map is a retraction [8]. When M is compact then the space of all torsion-free linear connections is simply-connected in the difference-tensor-topology [18]. In [8] it is shown how each spray generates Lie subalgebras of the Lie algebra of tangent fields on TM and there are universal analogues. These developments provide topological and algebraic tools for studying parametric statistical models.

3. Universal connections

Interestingly, at the same time that Chentsov [6] introduced his α-connections, García [12] introduced the idea of universal connections on principal bundles. Now, on the principal bundle FM of linear frames of a manifold M a principal connection corresponds precisely to a linear connection on the tangent bundle TM. So in the context of parametric models the construction of García applies directly. However, recently Modugno and his coworkers (cf. [15],[17] and its bibliography) extended the notion of a universal connection to any system of connections; the result of García then becomes the special case for systems of principal connections. To date, the role of differential geometry in parametric models has used mainly the special case of linear connections on a tangent bundle but already more general situations, effectively using fibred manifolds, are being pursued (cf. Amari [2], Barndorff-Nielsen and Jupp [3]). The more general setting for connections in the context of fibred manifolds is very natural; it may be a useful way to attack nonregular problems such as estimating the support points of a density function, when no information metric is defined (cf. open question (e) in [4]).

3.1 Universal connection on a system of connections

Let $p:E \to B$ be a fibred manifold on which is given a system of connections

$$\xi : C \underset{B}{\times} E \to JE : (x^\lambda, c^a, y^i) \to (x^\lambda, y^i, \xi^i_\lambda).$$

Then on the fibred manifold

$$\pi_1 : C \times_B E \to C$$

there is a connection

$$\Lambda : C \times_B E \to J(C \times_B E) \subset T^*C \underset{E}{\otimes} T(C \times_B E)$$

$$: (x^\lambda, c^a, y^i) \to dx^\lambda \otimes \partial_\lambda + dc^a \otimes \partial_a + \xi^i_\lambda \, dx^\lambda \otimes \partial_i .$$

The connection Λ is universal in the following sense. If Γ is any section of $C \to B$ then the induced connection $\Gamma = \xi \circ (\Gamma \circ p, 1_E) : E \to JE$ coincides with restriction of Λ, on the embedding by $(\Gamma \circ p, 1_E)$ of E in $C \times_B E$. So Γ is a pullback of Λ.

Example (a) Tangent bundle system

In the case of a tangent bundle, $E = TM \to M = B$. Here we know that C_T is an affine subbundle of $T^*M \otimes_M JTM$ projectable onto 1_M. Then the universal connection is given by

$$\bar{\Lambda} : C_T \times_M TM \to J(C_T \times_M TM) \subset T^*C_T \otimes T(C_T \times_M TM)$$

$$: (x^\lambda, v^\lambda_{\mu\nu}, y^\lambda) \to [(X^\lambda, V^\lambda_{\mu\nu}) \to (X^\lambda, V^\lambda_{\mu\nu}, Y^\mu V^\lambda_{\mu\nu} X^\nu)]$$

briefly, $\bar{\Lambda} = dx^\lambda \otimes \partial_\lambda + dv^a \otimes \partial_a + y^\mu v^i_{\mu\nu} \, dx^\nu \otimes \partial_i$.

Example (b) Frame bundle system

In the case of a frame bundle FM, with structure group $G = Gl(n,R)$ then $E = FM \to M = B$.

Here the system of linear connections has $C = JFM/G$, namely the space of 1-jets of G-invariant local sections of $FM \to M$. The universal connection is

$$\Lambda : JFM/G \underset{M}{\times} FM \to J(JFM/G \underset{M}{\times} FM) \subset T^*JFM/G \underset{FM}{\otimes} T(JFM/G \underset{FM}{\times} FM)$$

$$:(x^\lambda,\gamma^\lambda_{\mu\nu},b^\mu_\nu) \to [(X^\lambda,Y^\lambda_{\mu\nu}) \to (X^\lambda,Y^\lambda_{\mu\nu},b^\mu_\nu\ \gamma^\lambda_{\mu\theta}\ X^\theta)]$$

briefly, $\Lambda = dx^\lambda \otimes \partial_\lambda + d\gamma^a \otimes \partial_a + b^\mu_\nu \gamma^\lambda_{\mu\theta}\ dx^\theta \otimes \tilde{\partial}^\lambda_\nu$.

A discussion of the smooth bijective correspondence between $\bar{\Lambda}$ and Λ is given in [8]. In some respects it is more convenient to work with Λ and FM rather than with $\bar{\Lambda}$ and TM when investigating a given system of linear connections. This is because every linear connection Γ on M induces a riemannian metric g_Γ on FM and the associated section Γ of JFM/G $\to\!\!\!\!\to$ M gives an isometric embedding of (FM,g_Γ) as a slice of JFM/G $\times_M$ FM. Then nearby connections induce nearby slices and it is in this sense that properties can be tested for stability under perturbation of the connection. The result in 2.7 above is an example of this kind; it originates from an investigation of singularities in spacetime manifolds, which have a linear connection but no related positive definite metric (cf. [5], [9]).

32. Embedding parametric models as riemannian slices of a system

It may be that some statistical models generate systems of connections without their being strongly linked to any particular riemannian structure. Indeed, for the α-connection system, only ${}^0\Gamma$ is necessarily the Levi-Civita connection of the expected information metric. So there may be merit in studying this system as the 1-parameter family of (diffeomorphic) embeddings of either

(i) FM into JFM/G $\times_M$FM by the maps α_Γ

or (ii) TM into $T^*M \otimes JTM \times_M TM$ by the maps ${}^\alpha\Gamma$

The advantage in case (i) is that each (FM,g_{α_Γ}) is a riemannian manifold so our family of statistical manifolds appears as a family of riemannian slices of the

manifold JFM/G $\times_M$FM. The slice for $\alpha = 0$ has the special property that it is complete if and only if (M,g) is complete, where g is the expected information metric. Barndorff-Nielsen, Cox and Reid [4] gave a brief survey of current usage of differential geometry in statistical theory and concluded with some open problems; the approach via systems that has been described above may yield new lines of attack on problems (b), (d), (e), (f), (g) and (i). For example, one approach to a comparison of expected and observed conditional geometries (open question (b) in [4]) would be to study their respective families of embeddings as riemannian submanifolds of JFM/G $\times_M$FM. Curvature also appears to have relevance to statistical models (cf.[1] and [4]). Now, a connection even in its most general formulation, as a section of a jet bundle, defines a curvature and this curvature is always the pullback of the curvature of the appropriate universal connection. Also, connections induce differential operators on spaces of projectable vector-valued tangent forms and distinguish Lie subalgebras of the graded Lie algebra of vector-valued tangent forms. Then the canonical connection on the appropriate system yields corresponding universal structure : a universal calculus, "overconnections", and associate Lia subalgebras (cf. [10], [16]. [8]). The vanishing of curvature controls global integrability conditions for analysis on a manifold so it will be important in studying the geometrical "field equations" of parametric models. Since curvature is a 2-form, it vanishes identically if dim M $\leq$ 1, in particular on $\mathbb{R}$ and S^1. Williams [18] showed that the Levi-Civita map taking a metric to a connection is not continuous in the Whitney topology on the space of connections but it is continuous in the difference-tensor-topology there and some geometrical consequences are discussed in [8]. These yield further stability properties that may be pertinent to statistical applications; for example, we can deduce the following.

3.3 Torus and cylinder parameter spaces

Let the parameter space of a parametric statistical model (M,g) be $S^1 \times S^1$ or $S^1 \times R$ and denote by ${}^0\Gamma_g$ the $\alpha = 0$ connection, induced by expected information metric g. Then, in the Whitney topology for the space of linear connections on M, we find

(i) if ${}^0\Gamma_g$ is incomplete then every open neighbourhood of ${}^0\Gamma_g$ contains a Γ which is complete

(ii) if ${}^0\Gamma_g$ is complete then every open neighbourhood of ${}^0\Gamma_g$ contains a Γ which is incomplete.

Neither geodesic completeness nor incompleteness is Whitney-stable under variation of a connection so we may expect this to be true also under variation of the expected information metric, that is, under variation of the pdf's in the model. Again it suggests a deficiency in the ability of the Whitney topology to generate neighbourhoods appropriate for connection spaces. Two other choices for stability studies on spaces of linear connections are available : the difference tensor topology and the connection metric topology induced on frame bundles. In the former, Williams [18] showed that incompleteness on S^1 is stable; on $\mathbb{R}$, incompleteness and completeness are both stable and here, unlike on S^1, all linear connections are Levi-Civita connections of some metric.

3.4 General stability of incompleteness of α-connections

Using the connection-metric g_Γ on FM to generate neighbourhoods of a connection, a more general version of 2.6 from [5] and using [9] implies the following stability for incompleteness with respect to α-connections, that is of (FM, g_{α_Γ}).

If a parametric statistical model M has incomplete horizontal curves in FM with respect to some α-connection, then this persists for all α'-connections having $|\alpha - \alpha'|$ sufficiently small.

If Γ_g is the Levi-Civita connection of a riemannian manifold (M,g) then, with respect to Γ_g, FM has complete horizontal curves if and only if (M,g) is geodesically complete, because horizontal curves in FM cover geodesics in M. Moreover, if and only if Γ is not the Levi-Civita connection of some riemannian metric on M then it is possible for M to be Γ-geodesically complete while yet (FM, g_Γ) is incomplete as a riemannian manifold and hence geodesically incomplete itself. So, for any given parametric statistical model, it would be of interest to know whether the inverse image by the Levi-Civita map of the system of α-connections contains more than just the original expected information metric.

In the geometrical study of physical field theory, stability is an important concept both for practical and theoretical reasons. Evidently a theory yielding predictions that are not stable under variation of the initial data will have little value in practice, because of the finite precision in real experiments. Equally, only stable features of classical field theories are likely to persist in quantization processes based on fluctuations of background geometries. Similar arguments apply to the role of stability for statistical theory. The constructions given in this paper immediately yield results in the context of completeness for parametric models. They should also adapt to investigate stability properties in other contexts, for example to study behaviour under incorrect models, with robust estimation and with tests of separate families of hypotheses and to a study of inference stability (cf. open questions (d), (f) in [4]).

Acknowledgements

The author is grateful to P.E. Jupp and W.S.Kendall for their helpful comments.

REFERENCES

1. S.I. Amari. Differential goemetry of curved exponential familites - curvatures and information loss. Ann. Statist. 10 (1982) 357-385.
2. S.I.Amari, **Differential geometric methods in statistics**. Lecture Notes in Statistics 28 Springer-Verlag, New York 1985.
3. O. E. Barndorff-Nielsen and P. E. Jupp. Differential geometry, profile likelihood and L-sufficiency. Research Report 118 Dept. Theor. Statistics Aarhus University November 1984.
4. O.E.Barndorff-Nielsen, D.R.Cox and N.Reid. Differential geometry in statistical theory. Internat. Statist. Rev. 54, 1 (1986) 83-96.
5. D.Canarutto and C.T.J.Dodson. On the bundle of principal connections and the stability of b-incompleteness of manifolds. Math. Proc. Camb. Phil. Soc. 98 (1985) 51-59.
6. N.N.Chentsov. **Statistical Decision Rules and Optimal Inference**. (Russian : Nauka Moscow 1972) English : Translations of Mathematical Monographs Vol 53 Amer. Math. Soc. Providence, Rhode Island 1982.
7. A.P.Dawid. Discussion of a paper by E.Efron. Ann. Statist. 5 (1977) 1249.
8. L. Del Riego and C.T.J.Dodson. Sprays, universality and stability, Math.Proc. Camb. Phil. Soc. 103 (1987) (in press).
9. C.T.J.Dodson. Space-time edge geometry. Internat. J.Theor.Phys. 17, 6(1978) 389-504.
10. C.T.J.Dodson and M.Modugno. Connections over conncetions and a universal calculus. **Proc.VI Convegno Nazionale di Relativita Generale e Fisica della Gravitazione**. Florence 10-13 October 1984 (in press).
11. C.T.J.Dodson and T.Poston, **Tensor Geometry**, Pitman, London 1979; cf. also C.T.J.Dodson, Connection geometry, in these Proceedings pp. 1-15.
12. P.L.García. Connections and 1-jet fiber bundles. Rend. Sem. Mat. Univ. Padova 47 (1972) 227-242.
13. S.Kobayashi and K.Nomizu, **Foundations of Differential Geometry**, Vol 1 (1963) Vol 2 (1969) Interscience, New York.
14. K.Mackenzie., **Lie Groupoids and Lie Algebroids in Differential Geometry**. LMS Lecture Notes 124, CUP, Cambridge 1987.
15. L.Mangiarotti and M.Modugno. Fibred spaces, jet spaces and connections for field theories. **Proc. Internat. Meeting on Geometry and Physics**. Florence 12-15 October 1982, Pitagora Editrice, Bologna 1983 pp 135-165.
16. M.Modugno. Sistemi di connessioni ed applicazioni alle teorie di gauge. **Proc.Vii Congresso del Gruppo di Matematici de Espressione Latina**. Coimbra 9-14 September 1985 (in press).
17. M.Modugno. An introduction to systems of connections Preprint, Ist.Mat.Appl. "G.Sansone" Florence 1986.
18. P.M.Williams, **Completeness and its stability on manifolds with connection**.

Brownian motion, computer algebra, and the statistics of shape.

Wilfrid S. Kendall

Department of Mathematics, Strathclyde University, Glasgow

G1 1XH, U.K.

Abstract:

This paper is divided into three sections.

Section 1 surveys various constructions of Brownian motion on a manifold, leading up to a method which depends primarily on specification of a connection on the manifold. This construction is of real use in applications of Brownian motion to differential geometry. In preparation for the next section it is noted how Brownian motion encodes the Riemannian geometry of a manifold.

Section 2 reviews the statistical theory of shape, which can be seen as an application of geometry not to the parameter space (as in statistical differential geometry) but to the state space of the statistics. In particular a recent result on the stochastic kinematics of shape is discussed. The correspondence noted above between Brownian motion and the Riemannian metric is relevant here.

The method of proof uses symbolic Itô calculus, which is to say stochastic calculus carried out by means of symbolic computation using a computer algebra package. In section 3 a

brief description is given of the components of symbolic Itô calculus. The section and the paper are concluded by some speculation on the future role of computer algebra in producing computational realizations of statistical theories.

1 Brownian motion on a manifold

There are several ways of defining ordinary (real-valued) Brownian motion, and most of these generalize to ways of defining Brownian motion on a manifold M. In the following M will be a Riemannian manifold of dimension m (without boundary). We shall abbreviate 'Brownian motion on a manifold M' as BM(M). All the following definitions are equivalent; for some of the links see Pinsky (1981) and Darling (1984).

The infinitesimal random walk approach

Consider the following algorithm for a simulation of an isotropic random walk X in an m-dimensional Euclidean space:

(1.1)

choose an initial point x:

REPEAT

pick a direction uniformly at random at the current point x;

proceed along this direction for a time δ^2 at speed $\sqrt{m}\,\delta^{-1}$

UNTIL simulation complete;

It is well-known that such simulations approximate Brownian motion for small δ. The algorithm is still meaningful when

X lives on M, since the Riemannian metric allows us to define (a) a uniformly random direction, (b) the notion of proceeding (by means of a geodesic) along a given direction. Hence it makes sense to think of BM(M) as the infinitesimal version of an isotropic random walk.

The diffusion approach

Euclidean Brownian motion can be defined as the diffusion with associated semigroup possessing the Laplacian as infinitesimal generator.

On M there is a particular elliptic differential operator, the Laplace-Beltrami operator Δ, which generalizes the Euclidean Laplacian. This arises as the sum of second derivatives taken along geodesics whose velocities form an orthonormal basis at the point in question. We can define BM(M) as the diffusion whose diffusion semigroup is minimal and has Δ as infinitesimal generator. (We omit discussion of the technicalities of definition of a domain for Δ.)

The martingale problem approach

According to a theorem of Paul Lévy, real-valued Brownian motion B is characterized among continuous random processes by the following properties:

(1.2)

(a) B is a martingale with $B_0 = 0$,

(b) $B_t^2 - t$ defines a martingale.

Stroock and Varadhan (1979) base their approach to diffusion theory on a development of this, centred around the martingale problem. Among continuous processes (killed on explosion to infinity), BM(M) is characterized as solving the martingale problem:

(1.3) $\phi(x_t) - \phi(X_o) - \frac{1}{2}\int_0^t \Delta\phi(X_s)ds$ is a martingale for all smooth ϕ of compact support.

The method of stochastic development

Consider the algorithm given at (1.1). This can be refined as follows:

(1.4)

choose an initial frame of reference ξ at the initial point x;

REPEAT

pick a direction uniformly at random working in the current frame of reference;

proceed along this direction for a time δ^2 at speed $\sqrt{m}\,\delta^{-1}$ (dragging the frame of reference along the path)

UNTIL simulation complete;

This appears to be merely an elaboration of (1.1) and indeed in the Euclidean case there is nothing new. However in the

case of a general nonlinear manifold M one needs the concept of parallel transport or (equivalently) connection, to make sense of 'dragging the frame of reference along the path'. Given this notion there is no longer any need to make explicit reference to any Riemannian metric, as geodesics are also defined via connections.

Moreover the random directions can now be picked once-for-all at the beginning of the simulation, and injected into the algorithm using the current frame of reference. This allows us to consider the stochastic development of a given simulation, obtained by using the same random directions to produce a Euclidean random walk.

Proceeding to the limit as δ tends to zero, we obtain a system of stochastic differential equations:

$$\begin{aligned} d_S X &= \Xi d_s B \\ d_S \Xi &= H_\Xi d_S X \end{aligned} \tag{1.5}$$

where the stochastic differentials are Stratonovich differentials. Here the stochastic development B is a Euclidean Brownian motion in m-space (where m is the dimension of the manifold), X is the candidate for BM(M), and Ξ is the frame of reference sitting at the point X. The frame Ξ is the frame of reference sitting at the point X. The frame Ξ is a map from m-space to the tangent space at X, and is parametrized by belonging to a manifold, namely the bundle of linear frames GL(M). The frame Ξ injects the

randomness of B into the appropriate tangent space on M, providing differential information for the evolution of X. Differential information for the evolution of Ξ itself is obtained via the horizontal lift H, which lifts the tangent space at X up to the tangent space at Ξ. Specification of H corresponds to fixing a particular connection for M. Recall the connection is said to be metric if the sub-bundle of orthonormal frames is invariant under the flow of (1.5). There is a unique torsion-free metric connection H (the Levi-Civita connection) and the process produced by (1.5) with this connection is once again BM(M) if the initial frame $\Xi(0)$ is an orthonormal frame.

In coordinate form (1.5) can be written as

$$\begin{aligned} d_S X^i &= \Xi^i_a \, d_S B^a \\ d_S \Xi^i_j &= -\Gamma^i_{ak}(X) \Xi^k_j \, d_S X^a \end{aligned} \tag{1.6}$$

where we have replaced the horizontal lift H by $-\Gamma$ for Γ the classical Christoffel symbol. This coordinate form is given in Ikeda and Watanabe (1981, (4.29)). Several workers pioneered stochastic development, but Eells and Elworthy (1970) were among the first to give a systematic discussion.

At first blush the system (1.5) is a candidate for the most esoteric definition possible of BM(M), involving as it does both the notion of connection and the stochastic calculus of Stratonovich differentials. However it is the

limiting form of the algorithm (1.4), and is thus closely linked to the more elementary approach of (1.1).

The great advantage of the method of stochastic development is its flexibility. The author has found it of great utility in the exploitation of Brownian motion and stochastic calculus to prove results in differential geometry. The explicit construction given in (1.5) allows one to tie different BM(M) closely to Euclidean Brownian motions (as stochastic developments) and thence to each other. Moreover (1.5) is well-suited for definition of more general classes of random processes on Riemannian manifolds, such as the so-called Γ-martingales (processes whose stochastic developments B are Euclidean local martingales, but not necessarily Brownian motions). Darling (1982) and Emery (1985) review the theory of Γ-martingales, while their application in differential geometry is reviewed in W.S. Kendall (1987) and in more depth in W.S. Kendall (to appear). The monographs of Elworthy (1982) and Ikeda and Watanabe (1981) provide good expositions of the theory underlying stochastic development. See also Rogers and Williams (1987, chapter V).

The system (1.5) need not be based on the Levi-Civita connection. The underlying connection need not even be metric. One could in principle consider Γ-Brownian motion on M, where Γ refers to a completely general (non-metric) connection on M. However Ikeda and Watanabe (1981) note the process X will not be a Markov process for general Γ, as the infinitesimal characteristics of X will be affected by the (in general non-orthonormal) frame $\underline{e}$ sitting above X.

In the context of statistical manifolds connections come in dual pairs (apart from the self-dual connection which is actually the Levi-Civita connection for the statistical metric). This raises the (possibly idle) question, whether a new kind of random process, characteristic of the statistical manifold, might arise by combining these two connections. For example, if Γ and Γ^* form such a dual pair (being an α-connection and a $(-\alpha)$-connection) then consider the process $X^{(r)}$ formed by alternating between Γ-Brownian motion and Γ^*-Brownian motion at instants of a Poisson process of intensity r. In the limit as $r \to \infty$ does this yield a new random process on the statistical manifold, with distribution different from Brownian motion? Calculations suggest this is the case in general.

As noted by Ikeda and Watanabe (1981), to any elliptic diffusion (whose infinitesimal generator has smooth coefficients) one can associate a unique Riemannian metric for which the infinitesimal generator is of the form

Laplace-Beltrami operator plus drift.

For the second-order part of the generator has precisely the transformation properties required to produce a Riemannian

metric which in turn reproduces the Laplace-Beltrami operator. Thus the diffusion induces a 'correct' Riemannian metric with respect to which it is BM(M) plus drift.

Ikeda and Watanabe also note that by choice of a suitable metric connection Γ with torsion one can represent the diffusion as Γ-Brownian motion via system (1.5). I do not know of any applications of the Γ-Brownian motion representation; it does not seem to shed useful light on the behaviour of the diffusion. See however Hakim-Dowek and Lepingle (1986) who discuss Γ-Brownian motion on Lie groups, for Γ a connection arising from the group action.

The representation as Brownian motion plus drift has found use in discussion of various diffusions arising in biology and also in filtering theory (Antonelli, Elliott, Seymour, 1987, and references therein) and in the geometry of statistical shape (D.G. Kendall, 1977, W.S. Kendall, submitted) as discussed below.

9 The statistics of shape

The statistical analysis of shape is concerned with features of a dataset remaining invariant under translation, rotation, and dilatation. Geometry has a vital part to play in explaining what is the correct space on which to represent the shapes arising from the dataset. In contrast to the emphasis in statistical differential geometry, geometry is applied to the state space for the purposes of data analysis, rather than to the parameter space for the purposes of inference. Thus shape theory is similar to the theory of statistics of directional data (but is invariably concerned with constructed variables).

In the archaeological example discussed in Kendall and Kendall (1980) and further in D.G. Kendall (1984) the shape of a triangle of planar points is found as follows. The configuration of three points belongs to C^3 (but does not lie in the (real-)codimension two subspace corresponding to degenerate triples of coincident points). The orbit under the translation group is parametrized by $C^2 - (0,0)$. Taking a further quotient using dilatations yields parameter space S^1, the so-called space of 'pre-shapes'. Finally rotations are removed by taking orbits under an action by the rotation group $S^1 = SO(2)$. This step corresponds to the famous Hopf fibration

$$S^3 \to S^2$$

The last two steps can be collapsed into the projective operation; under the multiplicative action of the nonzero complex numbers we have

$$C^2 \to CP^1 = S^2$$

We are left with the shape space for three points in the plane,

$$\Sigma_2^3 = S^2$$

The natural geometry for Σ_2^3 turns out to be that of the sphere of radius 1/2. This makes the map $S^3 \to S^2$ into an isometric submersion. It is of great importance for the data analysis of shapes. The triple of three planar points has six relabelling symmetries under which the shape space can be reduced to a spherical triangle. Using an equi-areal projection of the 2-sphere onto the cylinder the spherical triangle is transformed into the spherical blackboard, a curvilinear triangle which provides a convenient representation space for the study of shape.

Chapter eight of Stoyan, Kendall, and Mecke (1987) gives an elementary introduction to this theory together with a simple example from geology.

In view of the correspondence between Brownian motion and Riemannian geometry one is led to consider the stochastic kinematics of shape. The projection of C^3 to Σ_2^3 sends a triple of three independent planar Brownian motions to a time-change of Brownian motion on $\Sigma_2^3 = S^2$ (D.G. Kendall, 1977). The time-change is precisely that to be found in the skew-product representation of Euclidean Brownian motion. This result follows immediately from properties of the Hopf fibration, which is a harmonic morphism (see Fuglede, 1978). In view of the relationship between Brownian motion and geometry discussed in section 2, this result echoes the geometry of Σ_2^3. It also implies the following. A shape formed by three independent identically distributed Gaussian points (with rotationally symmetric distribution) induces a random shape which is uniformly distributed on Σ_2^3. (Of course this follows more easily from symmetry arguments.) For further distributional results see D.G. Kendall (1986), D.G. Kendall and Le (1986, to appear), Le (to appear), and Small (1982).

Generalizations to higher dimensions and more points are discussed in D.G. Kendall (1984); for example the shape space of k points on the plane is a complex projective space

$$\Sigma_2^k = CP^{k-2} \qquad (2.1)$$

If the plane is replaced by a higher-dimensional Euclidean space then complications set in. For example the shape space of three points in 3-space, Σ_3^3, is the hemisphere S_+^2 of radius 1/2. In general a shape space need not even be a

manifold.

In an unpublished note D.G. Kendall also generalized the Brownian motion result; under a time-change a triple of three independent 3-space Brownian motions yields a shape diffusion which is Brownian motion on S^2_+ plus a certain drift directed to the pole of symmetry. The calculations involved are somewhat heavy, motivating the use of computer algebra in W.S. Kendall (submitted) to produce a further generalization:

The generalized D.G. Kendall theorem.

Consider the shape σ of the random triangle defined by three points diffusing independently in R^n with identical diffusion statistics. Suppose $n \geq 3$. Then the shape is naturally parametrized by a point on $\Sigma^3_2 = S^2_+$ as above.

If the common diffusion is either Brownian motion or the Ornstein-Uhlenbeck process then (subject to a random time-change depending only on the size of the triangle) the shape σ moves according to a process which is Brownian motion on S^2_+ modified by an added drift, directed towards the pole p and of strength

$$(2.2) \qquad g_n(\sigma) = ((n-2)/2)\ (2 \tan 2\ \mathrm{dist}(\sigma,p)).$$

Proof depends on recognizing that the normalized square side lengths of the Brownian triangle form a set of homogeneous coordinates for the shape σ. Computer algebra

is then used to determine the infinitesimal characteristics of the diffusion of σ, according to the method of symbolic Itô calculus described in the next section. The geometry of S^2_+ and the representation of the shape diffusion as Brownian motion plus drift follow.

As a useful spin-off one obtains a formula for the density of the random shape induced by three independent identically distributed Gaussian points in 3-space (with rotationally symmetric distribution). This may find application in multivariate analysis.

Here are two loose ends left by the above. The shape diffusion on Σ^3_n has a fairly simple form. It should therefore have a straightforward derivation, perhaps by means of stochastic differential geometry. Secondly, the application of symbolic Itô calculus to this problem was originally motivated by a desire to understand the so-called independent-reaction-time approximation of stochastic chemistry (Clifford, Green, Pilling, 1987). The above is only the first step to attaining this objective.

To conclude this section we note two other approaches to shape theory. Ambartzumian (1984; see also forthcoming monograph) uses factorization of differential forms and in particular has produced a fascinating variation in which the Euclidean geometry above is replaced by the geometry of the special linear group. Bookstein (1986) develops a different geometrical approach motivated by biological problems.

3 Computer Algebra

The impact of computer algebra on statistics to date appears to be limited, if "computer algebra" is interpreted in the strict sense of algebraic computation with symbolic expressions. To date I know of just four examples in the research: the work of Silverman and Young (1987) on asymptotics for density estimation; some unpublished work represented in K.P. Donnelly's thesis (Donnelly, 1981); as yet unpublished work of T.K. Carne, H-L. Le, D.G. Kendall and the work discussed below. However we can expect rapid development over the next few years following the recent developments of powerful and cheap desktop workstations which are capable of computer algebra. Microcomputers costing under £1000 can now support REDUCE (a computer algebra package written in the symbolic computation language LISP).

The first two examples mentioned above use computer algebra packages simply to perform algebraic computations which would otherwise require lengthy hand-calculation and verification. Carne, Le, and D.G. Kendall use REDUCE in computations of geometric invariants arising in shape calculations. The example reviewed here is somewhat different in that it implements the structure of a theory (namely Itô calculus) within a computer algebra package (namely REDUCE). I submit that this use of computer algebra in *implementation of a structural theory* has enormous potential for statistical theorists, allowing for clear *realization* of concepts either complex in themselves or leading to complicated algebraic expressions.

The example reviewed here is what I choose to call symbolic Itô calculus. For more details see W.S. Kendall (submitted), and my contribution to the discussion of Clifford, Green, and Pilling (1987).

Symbolic Itô calculus is based on procedures written in the REDUCE language which correspond to the following operations of stochastic calculus:

(1) introduction of random processes which are semimartingales, and their associated Ito differentials,

(2) definition of the second-order structure of the Itô differentials arising from these processes,

(3) definition of the first-order structure, namely the 'expected Itô differentials' or drifts of the processes,

(4) implementation of Itô's lemma, essentially as second-order truncation of power series expansions.

These procedures are relatively simple (they by no means exhaust the possibilities of computational representation of stochastic calculus) but combine with the extremely powerful algebraic manipulation facilities of REDUCE to provide an

effective tool for the probabilist. Thus the identification of the shape diffusion on Σ_n^3 for $n > 2$ can be achieved by manipulation of the Itô differentials of the squared side length processes arising from the Brownian (or Ornstein-Uhlenbeck) processes of the theorem in section 3. Indeed once the answer is known it can be checked by an entirely automatic operation! Finding the answer in a simple form was of course not an automatic process, and success arose from an iterative collaboration between computer and probabilist. (This contrast suggests a meta-theorem that refereeing a paper is intrinsically less difficult than writing it in the first place; not necessarily the case before the advent of computer algebra.)

Symbolic Itô calculus itself is at a very early stage and will develop considerably in response to the challenge of problems from complex stochastic systems. However as pointed out above it opens up an exciting vista of computational implementation of other statistical and probabilistic theories. A successful implementation will drastically reduce the need for involvement in detailed calculations, will allow much more scope for detailed examples, and will place a large premium on insight and general intuition. In particular I anticipate that future implementations of statistical computer algebra will complement and strengthen the program of geometrization of statistics. In the longer

term the theoretical statisticians of the early part of the next millennium may well consider courses in symbolic computation and LISP to be core components of a degree in mathematical statistics.

References

Ambartzumian, R.V. (1984) "Factorization in integral and stochastic geometry" In: *Stochastic Geometry, Geometric Statistics, Stereology* ed. R.V. Ambartzumian and W. Weil, 14-33, Teubner, Leipzig.

Antonelli, P.L., Elliott, R.J., Seymour, R.M. (1987) "Nonlinear filtering and Riemannian scalar curvature, II", *Adv. Appl. Maths.* 8, 237-253.

Bookstein, F. (1986) "Size and shape spaces for landmark data in two dimensions", *Statistical Science* 1, 181-242.

Clifford, P., Green, N.J.B., Pilling, M.J. (1987) "Statistical models of chemical kinetics in liquids" (with discussion), *J. Roy. Statist. Soc.* B49.

Darling, R.W.R. (1982) "Martingales in manifolds - definitions, examples, and behaviour under maps", in *Sem. Prob. XVI (supplement)*, 217-236, Lecture Notes in Maths 921, Springer-Verlag, Berlin.

Darling, R.W.R. (1984) "On the convergence of Gangolli processes to Brownian motion on a manifold", *Stochastics* 13, 277-302.

Donnelly, K.P. (1981) "Genetic linkage, detectable relationships, and other topics", Ph.D. thesis, University of Cambridge.

Eells, J., Elworthy, K.D. (1970) "Wiener integration on certain manifolds" In *Problems in Nonlinear Analysis*, CIME IV, 67-94.

Elworthy, K.D. (1982) *Stochastic Differential Equations on Manifolds.* Cambridge University Press, Cambridge.

Emery, M. (1985) "Convergence des martingales dans les varietes". In *Proc. en L'honneure L.Schwartz*, volume 2, 47-63. Soc. Math. de France.

Fuglede, B. (1978) "Harmonic morphisms between Riemmanian manifolds", *Ann. Inst. Fourier (Grenoble)* 28, 107-144.

Hakim-Dowek, M., Lepingle, D. (1986) "L'exponentielle stochastique des groupes de Lie", In *Sem. Prob. XX, 352-374*, Lecture Notes in Maths 1204, Springer-Verlag, Berlin.

Ikeda, N., Watanabe, S. (1981) *Stochastic Differential Equations and Diffusion Processes.* North-Holland/Kodansha, Amsterdam/Tokyo.

Kendall, D.G. (1977) "The diffusion of shape (abstract)", *Adv. Appl. Prob.* 9, 428-430.

Kendall, D.G. (1984) "Shape manifolds, Procrustean metrics, and complex projective spaces", *Bull. London Math. Soc.* 16, 380-424.

Kendall, D.G. (1986) "Exact distributions for shapes of random triangle in convex sets" *Adv. Appl. Prob.* 17, 308-329.

Kendall, D.G., Kendall, W.S. (1980) "Alignments in two-dimensional random sets of points", *Adv. Appl. Prob.* 12, 380-424.

Kendall, D.G., Le, H-L. (1986) "Exact shape densities for random triangles in convex polygons" In *Analytic and Geometric Stochastics* ed. D.G. Kendall, 59-72, Special supplement to *Adv. Appl. Prob.*

Kendall, D.G., Le, H-L. (to appear) "The structure and explicit determination of convex-polygonally generated shape-densities on Σ_2^3". *Adv.Appl. Prob.*

Kendall, W.S (1987) "Stochastic differential geometry: an introduction", *Acta Applic. Math.* 9, 29-60.

Kendall, W.S. (to appear) "Martingales on manifolds and harmonic maps", Proceedings of the AMS conference *Geometry of Random Motion*, ed. R. Durrett and M. Pinsky.

Kendall, W.S. (submitted) "Symbolic computation and the diffusion of shapes of triads".

Le, H-L. (to appear) "Explicit formulae for polygonally generated shape-densities in the basic tile". *Math. Proc. Camb. Phil. Soc.*

Pinsky, M. (1981) "Homogenization and stochastic parallel displacement". In *Stochastic Integrals*, ed.D.Williams, 271-284. Lecture Notes in Maths 851, Springer-Verlag, Berlin.

Rogers, L.C.G., Williams, D. (1987) *Diffusions, Markov processes, and martingales. Volume 2, Ito calculus.* Wiley, Chichester.

Silverman, B.W., Young, G.A. (1987) "The Bootstrap: to smooth or not to smooth?" *Biometrika* 74, 469-479.

Small, C.G. (1982) "Random uniform triangles and the alignment problem" *Math. Proc. Camb. Phil. Soc.* 91, 315-322.

Stoyan, D., Kendall, W.S., Mecke, J. (1987). *Stochastic Geometry and its applications.* Wiley/Akademie-Verlag, Chichester/Berlin.

Stroock, D.W. and Varadhan, S.R.S. (1979) *Multidimensional Diffusion Processes.* Springer-Verlag, Berlin.

Yokes: Elemental properties with statistical applications

P. Blæsild

Department of Theoretical Statistics

Institute of Mathematics

University of Aarhus

Ny Munkegade

8000 Aarhus C

Denmark

Abstract

The definition of a yoke is reviewed and illustrated by two statistical examples which induce respectively the observed and expected geometries of a statistical model. Using some elemental properties of yokes, the analogy between higher order properties of a statistical model in terms of its observed and expected geometries is indicated. □

Let M be a d-dimensional differentiable manifold covered by a single chart and let $\tilde{M}$ be a copy of M. If the coordinates $\omega = (\omega^1,\ldots,\omega^d)$ of M vary in Ω we let $\tilde{\Omega}$ be a copy of Ω. For $g \in C^\infty(M \times \tilde{M})$ we set

$$g_{J_s;K_t}(\omega;\tilde{\omega}) = \partial_{J_s}\tilde{\partial}_{K_t} g(\omega;\tilde{\omega})$$

and

$$g_{J_s;K_t}(\omega) = g_{J_s;K_t}(\omega,\omega) \ .$$

Here J_s denotes the set of indices $j_1,\ldots,j_s$ and ∂_{J_s} denotes the operator $\partial^s/(\partial\omega^{j_1}\ldots\partial\omega^{j_s})$ and similarly for K_t and $\tilde{\partial}_{K_t}$.

A function $g \in C^\infty(M \times \tilde{M})$ is called a yoke if for every ω we have that

(i) $\quad g_{j;}(\omega) = 0 \ , \qquad j = 1,\ldots,d$

and

(ii) the matrix $(g_{jk;}(\omega))$ is positive definite

The concept of yokes was introduced in Barndorff-Nielsen (1987). Here some elemental properties

of yokes are discussed. For instance it is shown that the collection of arrays

$$g_{;} = (g_{J_s;K_t} | s,t=1,2,\ldots) \qquad (1)$$

is a double costring of infinite length (in the terminology of Barndorff-Nielsen and Blæsild, 1987) and furthermore that for $\alpha \in R$ the set of arrays

$$\overset{\alpha}{g} = \left\{ \frac{1+\alpha}{2} g_{K_t;j} + \frac{1-\alpha}{2} g_{j;K_t} | t=1,2,\ldots \right\}$$

constitutes a $(0,1)$ costring of infinite length. Together with (ii) this in turn implies that the yoke g for every $\alpha \in R$ induces an affine connection on M, which is called <u>the α-connection associated with g</u>.

Let $p(x;\omega)$ denote the model function of a statistical model with sample space $\mathfrak{X}$ and parameter space Ω and let $\hat{\omega}$ denote the maximum likelihood estimator of ω. Assuming the existence of an ancillary statistic a such that the transformation $x \to (\hat{\omega},a)$ is one-to-one the log likelihood function may be written as

$$l = l(\omega) = l(\omega;\hat{\omega},a) = l(\omega;x) = \log p(x;\omega) \ .$$

Two examples of yokes of particular statistical interest are

$$g(\omega;\tilde{\omega}) = \bar{l}(\omega;\tilde{\omega}) = l(\omega;\tilde{\omega},a) - l(\tilde{\omega};\tilde{\omega},a) \ ,$$

i.e. the conditional normalized log likelihood function given a, and

$$g(\omega;\tilde{\omega}) = -I(\tilde{\omega},\omega) = E_{\tilde{\omega}}(l(\omega) - l(\tilde{\omega})) \ ,$$

where I denotes the Kullback-Leibner information. The α-geometries associated with these yokes are called observed and expected geometries, respectively (cf. Barndorff-Nielsen 1986 and 1987). The elements of the corresponding double strings (1) are, respectively

$$l_{J_s;K_t} \ , \tag{2}$$

termed the mixed log model derivatives (cf. Barndorff-Nielsen 1986), and

$$\upsilon_{J_s;K_t} = \sum_{\tau=1}^{t} \sum_{K_t/\tau} \upsilon_{J_s,K_{t1},\ldots,K_{t\tau}} \tag{3}$$

where υ denotes the mixed moments of log likelihood derivatives, i.e.

$$\upsilon_{J_s,K_{t1},\ldots,K_{t\tau}} = E_\omega(\partial_{J_s} l \cdot \partial_{K_{t1}} l \cdots \partial_{K_{t\tau}} l) \ .$$

Let λ denote the joint cumulants of log likelihood derivatives, i.e.

$$\lambda_{J_s,K_{t1},\ldots,K_{t\tau}} = c_\omega(\partial_{J_s} l, \partial_{K_{t1}} l, \ldots, \partial_{K_{t\tau}} l) \ .$$

The double string $\upsilon_{;}$ in (3) was introduced by Blæsild (1987) who showed that the arrays obtained by replacing υ with λ in (3), i.e.

$$\lambda_{J_s;K_t} = \sum_{\tau=1}^{t} \sum_{K_t/\tau} \lambda_{J_s,K_{t1},\ldots,K_{t\tau}} \tag{4}$$

constitute another double string $\lambda_{;}$. Furthermore one has, rather surprisingly, that $\upsilon_{;} = \lambda_{;}$.

By considering the Bartlett adjustment Blæsild (1987) indicates that higher order properties of a statistical model which at first sight seem to be quite different in terms of the observed geometries and the expected geometries may be shown to be very similar in view of the analogy between (2) and (3).

References

Barndorff-Nielsen, O.E. (1986): Likelihood and observed geometries. Ann. Statist 14, 856-873.

Barndorff-Nielsen, O.E. (1987): Differential geometry and statistics: Some mathematical aspects. Research Report 156, Dept. Theor. Statist., Aarhus University. (To appear in Indian J. Math.)

Barndorff-Nielsen, O.E. and Blæsild, P. (1987): Derivative strings: Contravariant aspects. Proc. Roy. Soc. London A A 411, 421-444.

Blæsild, P. (1987): Further analogies between expected and observed geometries of statistical models. Research Report 160, Dept. Theor. Statist., Aarhus University.

Pre-geodesic Equations

B. L. FOSTER

Department of Mathematical Sciences, University of Montana, Missoula, MT 59812 U.S.A.

Abstract Finding Riemannian geodesic equations requires distance to be a geometric invariant. But for a natural extension of Riemann's metric, the giving on a manifold of a certain simple new-tensor, or derivative string, the resulting invariant is physical, depending on velocity and acceleration. For such invariants, general variational equations only are available; the "geodesic" equations obtained are complicated. An intermediate case applies these equations to Riemannian geometry itself. A novel form of geodesic equation emerges, biquadratic instead of quadratic, whose "connection" part is a new-tensor, but seemingly not a derivative string.

The first partial derivatives of a scalar field have a simple change law, $\partial_\alpha \phi = \partial_a \phi D^a_\alpha$, where the Greek letter denotes a new coordinate system and D^a_α is a Jacobian matrix. The second partials don't change so simply, but if they're taken together with the first, then simplicity is restored:

$$(\partial_\alpha \phi \;\; \partial_{\alpha\beta} \phi) = (\partial_a \phi, \; \partial_{ab} \phi) \begin{pmatrix} D^a_\alpha, D^a_{\alpha\beta} \\ 0 \;\; , D^a_\alpha D^b_\beta \end{pmatrix} .$$

The next, vital, step is to consider an arbitrary

(g_a, g_{ab}) having the same change law. The new, tensor-like, objects were studied in [1] and [2], and more recently in [3]. This enlargement of the idea of tensor has been independently rediscovered by Barndorff-Nielsen[4], in a statistical context. The study is continued by Barndorff-Nielsen and Blaesild in [5] and [6]. The definitions given in the purely mathematical part of their work are in some ways more general and the development from those definitions much more complete than was attempted in [1], [2], and [3].

The enlargement's reward is immediate. From the most basic new-tensor, or derivative string in Barndorff-Nielsen's words, (g_a, g_{ab}), and its dual set of mixed differentials can be formed an invariant, $g_a d^2x^a + g_{ab} dx^a dx^b$, whose square root generalizes Riemannian distance. This long-sought field unifies a covariant vector with a g_{ab} which is not a Riemann metric, because it changes like a second partial derivative, but which becomes one if g_a vanishes.

But there's a price. Along a curve $x^a(t)$, the new invariant $\int (g_a \ddot{x}^a + g_{ab} \dot{x}^a \dot{x}^b)^{1/2} dt$ depends on velocity and acceleration, unlike Riemannian distance whose velocity dependence is only apparent. Thus it is a physical invariant, rather than a geometric one. Hence only general variational equations can be used, without the simplifying assumption of parameter independence.

Let Riemannian distance have integrand $L = (r_{ab} \dot{x}^a \dot{x}^b)^{1/2}$. The variational equations are $\dot{L}_{\dot{x}^c} - L_{x^c} = 0$, taking partial derivatives first. Compute and collect to get

$$\dot{x}^a \dot{x}^b [(2r_{ab} r_{cd,e} - r_{cd} r_{ab,e} - r_{de} r_{ab,c}) \dot{x}^d \dot{x}^e + 2(r_{ab} r_{cd} - r_{ac} r_{bd}) \ddot{x}^d] = 0,$$

a quadratic form whose coefficients resemble, but are not, ordinary geodesic terms. These "pre-geodesic" equations are what must generalize.

In them, how do the pre-Christoffel encounters of the first kind transform? Setting $M_{abcde} = 2r_{ab} r_{cd,e} - r_{cd} r_{ab,e} - r_{de} r_{ab,c}$ and changing coordinates gives $M_{\alpha\beta\gamma\delta\varepsilon} = M_{abcde} D^a_\alpha D^b_\beta D^c_\gamma D^d_\delta D^e_\varepsilon + r_{ab} r_{cd} *$, where $*$ has six terms, each a product of three Jacobians and one Hessian, e.g., $D^a_\alpha D^b_\beta D^c_\gamma D^d_\delta$.

The six different second derivatives in $*$ rule out M or part of M being an affine connection, and, it is conjectured, also rule out M being a derivative string. But M is a new-tensor; this requires that its change matrix, whose main block is $*$, satisfy transitivity. In schematic form, this means checking that $K^{\text{Roman}}_{\text{Greek}} K^{\text{Greek}}_{\text{Cap. Roman}} = K^{\text{Roman}}_{\text{Cap. Roman}}$, which is straightforward.

REFERENCES

1. Foster, B. L., Differentiation on Manifolds Without a Connection, Mich. Math. Jour., vol. 5 (1958), 183-190.

2. Foster, B. L., Some Remarks on Tensor Differentiation, Annali di Matematica pura ed applicata (IV) Vol. LIV (1961), 143-146.

3. Foster, B. L., Would Leibniz Lie to You? (Three Aspects of the Affine Connection), Math. Intelligencer 8 (1986), 34-40, 57.

4. Barndorff-Nielsen, O. E., Strings, tensorial combinants and Bartlett adjustments. Proc. R. Soc. Lond. A 406 (1986), 127-137.

5. Barndorff-Nielsen, O. E. and Blaesild, P., Strings: mathematical theory and statistical examples. Proc. R. Soc. Lond. A 411 (1987), 155-176.

6. Barndorff-Nielsen, O. E. and Blaesild, P., Derivative strings: contravariant aspect. Proc. R. Soc. Lond. A 411 (1987), 421-444.

INVARIANCE PROPERTIES OF METRICS AND CONNECTIONS IN REGULAR FAMILIES

Dominique B. Picard

CNRS Unite Associee 743 Statistique et Modeles, Université Paris-Sud Centre d'Orsay
91405 ORSAY cedex. FRANCE.

INTRODUCTION

The differential geometric framework provides a new view of statistics and, at the same time, a lot of theoretical questions: One basic question which arises among them is the following: how to summarize the information about a family of probabilities preserving the statistical and geometrical properties?

Of course, following Amari's framework, we would attempt to claim that, in any case, at second order, it will be sufficient to know the Fisher Information and the family of α-connections. But a more suspicious mind could certainly hesitate; since, as it is obivous that α-connections are basic quantities, it is not proven that they form the unique family of connections to consider.

A first partial answer to this question has been given by Chentsov [2]: Defining equivalence of experiments according to Le Cam's framework, as equivalence under Markov kernels, he proved that, in the case of experiments of a finite number of atoms,

Fisher-Rao metrics (resp. α-connections) were invariant and unique with this property.

Considering now the problem of generalizing this property in the case of general regular families of probabilities, it will be proven that : - The class of Markov kernels mapping a sub-experiment into a sub-experiment may be very sparse. We give a description of this class in the case of linear exponential families and prove that, in this case, Fisher metric and α-connections are invariant under Markov equivalents but they usually are not unique with this property. (We investigate, for counter example, the class of normal distributions with unkown mean and variance).

Thus, we propose to introduce a notion of invariance under approximation: Two sub-experiments are locally equivalent at a point θ_0, to order k, if they give the same expansions of the power of every likelihood ratio test $(\theta_0;\theta_0+ u/n^{1/2})$ calculated for n iid (independent identically distributed) observations, up to the power k/2.

We characterize this equivalence for k = 1,2 and prove that the Fisher metric is invariant of order 1 and 2 and unique with this property; α-connections are not usually invariant of order 1, they are invariant of order 2 and unique with this property.

Model - Markov Invariance

We consider a family $E = \{P_\theta, \theta \in \Theta\}$. It will be assumed to be regular enough to ensure the existence, non singularity and continuous differentiability of the Fisher information matrix :

$F(\theta) = (F_{ij}(\theta))$

$$F_{ij}(\theta) = -E_\theta\{\partial_i l(x,\theta)\ \partial_j l(x,\theta)\}$$

and furthermore the existence of the fundamental tensor

$$T_{ijk}(\theta) = E_\theta\{\partial_i l(x,\theta)\ \partial_j l(x,\theta)\ \partial_k l(x,\theta)\}.$$

It will be assumed, in addition, that if R(x,h) denotes the rest of the Taylor expansion of Log p(x,θ+h) - Log p(x,θ) up to order 3, then $|h|^3 R(x,h)$ is uniformly bounded by a function φ(x) which has moments up to the third, for each θ in Θ.

Definition of Markov equivalence:

Two sub-experiments of E, $E' = \{P_\theta, \theta \in T'\}$ and $E'' = \{P_\theta, \theta \in T''\}$ where T' and T'' are 2 regular submanifolds, are said to be Markov equivalent if there exists one diffeomorphism φ: T' → T'' and two Markov kernels π_{12} and π_{21} such that:

$$\forall\ \theta \in T',\ P_{\varphi(\theta)} = P_\theta \circ \pi_{12}\ ,\ P_\theta = P_{\varphi(\theta)} \circ \pi_{21}.$$

φ is then called a Markov morphism associated with the pair E' and E''.

Definition of Markov invariance:

A metric g (resp. a connection ∇) is said to be Markov invariant if it is invariant under every Markov morphism associated with any pair of equivalent sub-experiments E' and E''.

The usual meaning of this invariance is that Markov morphisms preserve curve length (resp. map geodesics into geodesics).

Proposition

If E is linear exponential family, let φ be a Markov morphism defined on an r-dimensional submanifold T' (r > 0), then

1. φ can be extended as a Markov morphism to $C(T')$, the convex hull of T' in Θ.

2. φ is an affine mapping :

   $\forall\ \theta \in C(T'), \varphi(\theta) = A\theta + b.$

3. $\forall\ \alpha \in [0,1],\ \forall\ \theta_1, \theta_2 \in C(T'),\ h(\alpha\theta_1+(1-\alpha)\theta_2) - \alpha h(\theta_1) - (1-\alpha)h(\theta_2) = h(\alpha\varphi(\theta_1)+(1-\alpha)\varphi(\theta_2)) - \alpha h(\varphi(\theta_1)) - (1-\alpha)h(\varphi(\theta_2))$

   (h is defined by : $l(x,\theta) = (\theta,S(x)) - h(\theta)$.)

Corollary 1:

In linear exponential families, the Fisher metric and α-connections are Markov invariant.

Proposition

In the case where $E = \{ N(\mu,\sigma^2),\ (\mu,\sigma) \in \mathbb{R} \times \mathbb{R}\}$, the set of Markov morphisms is restricted to the following maps:

(in natural coordinates)

$$\tau_1 = \frac{\mu}{\sigma^2}\ ,\quad \tau_2 = \frac{1}{\sigma^2}$$

$$\varphi_1(\tau_1,\tau_2) = c\tau_1 + d\tau_2\ ,\ \varphi_2(\tau_1,\tau_2) = c\,\tau_2 .$$

for c,d arbitrary, $c \neq 0$.

Consequently, the set of invariant metrics consists of the following: (in natural coordinates)

$$g|_{(\tau_1,\tau_2)}\left(\frac{\partial}{\partial\tau_i}, \frac{\partial}{\partial\tau_j}\right) \qquad (i,j = 1,2)$$

$$g(-)|_{(\tau_1,\tau_2)} = \begin{bmatrix} \gamma/\tau_2 & -\gamma\tau_1/\tau_2^2 \\ -\gamma\tau_1/\tau_2^2 & \gamma\tau_1^{\,2}/\tau_2^{\,3} + \delta/\tau_2^2 \end{bmatrix}$$

for arbitrary γ and δ different from 0.

It is noteworthy that these metrics differ from the Fisher metric as soon as we depart from (γ=-1, δ=-½). Moreover, it is not difficult to see also that the associated Riemannian connections differ from the α-connections of the family.

Invariance under approximation

Suppose, that we have now n iid observations of our model available. Let $\beta_n(\theta,\theta',\alpha)$ be the power of likelihood-ratio test of θ against θ', based on the n iid observations, at the level α

Definition

φ is a statistical morphism of order k (smk) of the experiment E iff there exists a point θ_0 such that:

1. φ is defined on T', an r-dimensional regular submanifold of Θ ($r > 0$), containing θ_0.

2. $\forall\ u \in T_{\theta_0}(T')$,

   $\lim_{n\to\infty} n^{k/2}[\beta_n(\theta_0,\theta_0+u/n^{1/2},\alpha) - \beta_n(\varphi(\theta_0),\varphi(\theta_0+u/n^{1/2}),\alpha)] = 0$

It is a consequence of Jorgensen [5] that Markov morphisms are smk for every k.

Characterization of sm1 and sm2

φ is an sm1 iff:

$$t_x\, t\, D\varphi|_{\theta_0} F(-)|_{\varphi(\theta_0)} D\varphi|_{\theta_0} x = t_x\, F(-)|_{\theta_0} x, \quad \forall x \in T_{\theta_0}(T')$$

φ is an sm2 iff:

1. φ is sm1

2. If $A = (a_{ij})$ denotes the jacobian matrix of φ in θ coordinates at the point θ_0, then $\forall\ (x_1,x_2,\ldots,x_k) \in T_{\theta_0}(T')$

$$\sum_{ijlrst} x_i\, x_j\, x_l\, a_{ri}\, a_{sj}\, a_{tl}\, T_{rst}\,(\varphi(\theta_0)) = \sum_{ijl} x_i\, x_j\, x_l\, T_{ijl}(\theta_0).$$

Theorem

A metric is invariant under the sml if it is up to a scaling factor the Fisher-Rao metric. A connection is invariant under the sml if it is the Riemannian connection associated to the Fisher-Rao metric. A metric is invariant under the sm2 if it is up to a scaling factor the Fisher-Rao metric. A connection is invariant under the sm2 if it is one of the α-connections. Details and proofs of these results can be found in [4].

BIBLIOGRAPHY

1. Akahira,M. and Takeuchi,K. (1981). Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency. Springer, Lecture Notes in Statistics, vol.7 Springer-Verlag, Berlin

2. Amari, S.I. (1985). Differential Geometrical Methods In Statistics. Lecture Notes in Statistics. New York: Springer-Verlag

3. Chentsov, N.N. (1972). Statistical Decision Rules and Optimal inference (in Russian). Moscow: Nauka. English translation (1982). Translation of Mathematical Monographs Vol.53, AMS

4. Picard, D.B. (1987). Invariance properties of Fisher-Rao metric and Chentsov-Amari connections in regular families. Republication Univ. Paris-Sud

5. Jorgensen, E.N. (1970). Comparison of experiments when the parameter space is finite. Z.Wahrsch. Verw. Geb. 16, 219-249

WHAT YOU CAN DO WITH n OBSERVATIONS

T.J. Lyons
Department of Mathematics
University of Edinburgh

ABSTRACT

We explain how, with only n likelihood functions available, it is possible to define a random metric (the empirical information metric) on a parameter space. This empirical metric is related to the Fisher information but, at least in some situations, seems preferable to it. One significant feature is that it depends only on the n likelihood functions. No expectations are taken and no prior is assumed to exist.

Using this metric we introduce an inferential measure on our parameter space. This measure can be characterised in several distinct ways and has some claim to be natural. However, to interpret it properly, we introduce the concept of the volatility of an inferential measure on the parameter space relative to a fixed choice of interest parameters. This seems to be a useful idea independent of the rest of our discussion and it gives a quantitative measure of how reliable or meaningful one's inference is - even in cases where good ancillary statistics are not available.

The paper has two parts. Firstly we introduce the basic definitions - then we examine particular examples.

I. The Model.

We take a restrictive view of what data is available to us to make our inferences. We suppose we have n (independent) observations ω_i from n sample spaces Ω_i. We have a parameter space M which is a smooth manifold, and for each ω_i we have a smooth strictly positive[1] likelihood function $p_i(m)$ indicating for each m, the relative frequence of ω_i occurring in repeated trials of the i^{th} experiment.

The $p_i(m)$ are the only data we use in our construction of metrics and inference probability measures. We do not assume the existence of a prior or knowledge of the probabilities of events that have not occurred.

This is a slightly restrictive type of model when compared to the usual statistical models; one should note that the notion of sufficient statistic appropriate here does not correspond to the usual one. The point is that there are many different ways of choosing the n full likelihood functions

$$p^i(m,\omega) \qquad m \in M \quad \omega \in \Omega$$

consistent with $p^i(m,\omega_i) = p_i(m)$.

II. The metric.

There are many equivalent ways one could describe the following metric. We outline three such ways. The likelihood functions p_i are only determined up to constant scalar multiples. But the exact 1-forms given by $\alpha_i = d \log p_i$ are intrinsic. We can recover all that is essential about p_i

[1] Strict positivity is not essential here.

by integrating α_i and so we identify each outcome ω_i with the 1-form α_i.

We define the Empirical metric on M simply to be the empirical covariance of the n 1-forms. That is if X, Y are tangent vectors over $m \in M$ and $\alpha = \sum_1^n \alpha_i$ we put

$$\langle X,Y\rangle_m = \sum_1^n \alpha_i(X)\alpha_i(Y) - \frac{1}{n}\alpha(X)\alpha(Y)$$

$$= \sum_1^n X \log p_i \; Y \log p_i - \frac{1}{n} X \log p \; Y \log p$$

where $p = \prod_1^n p_i$ is the total likelihood.

There might be a case for not subtracting off the mean - or at least modifying the multiple from $1/n$. However, it is clear that any symmetric polynomial homogeneous of degree two in the α_i must be of essentially the form given. Calculation of worked examples indicates that this metric has more canonical properties than the others.

The Fisher information at m is just

$$\frac{n}{n-1} \mathbb{E}_m(\langle,\rangle_m)$$

and so is essentially the expected value of this metric at m.

This metric measures the co-variance of the conflicting information given by the n observations and is standardising that noise at all points of the manifold.

There are two other constructions of the measure that

one might wish to consider.

(a) One can use the likelihood functions to embed M isometrically into R^{n+1} with the Lorentz metric

$$\begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ & & & \ddots \\ 0 & & & -1/n \end{bmatrix}$$

$$m \longrightarrow ((\log p_i(m))_{i=1}^{n},\ \log p(m))$$

This has the advantage of allowing a 2-parameter family of empirical connections to be defined as in the usual statistical theory.

Let $\tau(k)$ be i.i.d. uniform on $\{1, \ldots, n\}$ then the random walk in the space of 1-forms $\sum_{j=1}^{k} \alpha_{\tau(j)}$ is to second order $\tilde{=}$ αt + Brownian motion (where $\alpha = d \log p$). The co-variance of the Brownian motion is used to construct a metric on the tangent space at each point on M. This point of view is consistent with the perspective discussed later in III. In any case it makes it clear that this metric captures second order information about the observations, but has not captured more than that.

III. A measure for inference.

Suppose we have a metric g on M representing our uncertainty of choice of parameter. (The metric need not be that given in II and could for example be the Fisher metric) Then we no longer need one forms and we may introduce vector fields $\underline{\alpha}^i, \underline{\alpha}$ associated with our individual and total observation:

$$\underline{\alpha}^i = \nabla_g \log p_i$$

$$\underline{\alpha} = \nabla_g \log p .$$

We now introduce three reasonable choices of measure μ on M indicating our view of the location of the true parameter. It is a remarkable fact that for the empirical metrics described above and essentially only in this case the measure all coincide. We believe this is evidence that the empirical metric is natural and that for this choice of metric the inferential measure is also natural.

(a) Jeffries-Bayes.

Here we just introduce the Riemannian volume dV_g as a formal prior and condition on the observed total outcome $(\omega_1, \ldots, \omega_n)$. In this case one has a posterior probability distribution.

$$\mu_1(A) = \int_A \frac{p(m)dV_g(m)}{\int_M p(m)dV_g(m)} .$$

(b) The drunk following α.

Consider an observer at $\underline{m} \in M$; he might take $\underline{\alpha}(m)$ as an indication of the direction in which the true choice of parameter lies. So he moves along α for unit time and repeats the process. Eventually he will get to some local maximum and stop. However suppose the observer is sceptical and say - O.K. "I want to move along $\underline{\alpha}$ for 1 unit of time - but the metric tells me that any of the points a fixed

distance from this position are just as reasonable so - I will choose one at random" then effectively one is considering the diffusion

$$dX_t = \underline{\alpha}(X_t)dt + \gamma dB_t$$

where B is the Brownian motion on M associated to g. The normalised stationary measure of X_t is our second candidate for an inference measure. In fact it is easily computed that the measure is given by

$$\mu_\gamma(A) = \frac{\int_A p^{1/\gamma}(m)dm}{\int_M p^{1/\gamma}(m)dm}.$$

So recovering the Bayesian approach of (a) if $\gamma = 1_r$ and more generally giving a 1-dimensinal family of inferences interpolating between point mass at the point of global maximal likelihood ($\gamma \to 0$) and the Riemannian volume as $\gamma \to \infty$.

(c) Stochastic Flows.

Our last construction is (at least from our perspective) more interesting as it preserves much more of the information concerning the observations than the others.

Consider the Random bootstrap observer who chooses one $\underline{\alpha}^{\tau(1)}$ of the $\underline{\alpha}^i$ at random - moves along it from unit time and then chooses a second moving $\alpha^{\tau(2)}$ and moves along that etc.

In stochastic language the continuous approximation is

$$dX_t = \underline{\alpha}\, dt + \sum_i (\alpha^i - \underline{\alpha}/n) \circ d\omega_t^i \qquad (\circ = \text{Stratonovich integral})$$

where ω^i are independent Brown motions on $\mathbb{R}$. What makes this approach rich is that the same choice of $\tau(i)$ (or ω_t^i) can be used simultaneously for all points $m \in M$ and one obtains a flow on the manifold. Thus one can consider the degree to which prechosen configurations of starting points (e.g. on a submanifold) get mixed up and chaotic - or are preserved in shape if not location.

We have not done this here - we only consider the stationary measure μ_3 of the trajectory of a single point m. For a general metric the determination of μ_3 is non-trivial. But in the case of the empirical metric this measure μ_3 is that already described in (a).

IV. Volatility or Stability.

We now introduce a concept of volatility (or in Geometric terms weighted energy) of a particular inference about an interest parameter. The basic idea is the following. Let M be a space of parameters, let $\pi : M \to N$ be projection onto some interest parameters. Suppose we have metrics g on M and h on N (the latter might be given externally) and μ is a probability measure representing reasonable choices for the parameter $m \in M$. Then it is natural to use $\pi(\mu) = \nu$ as a measure of the plausible locations for $n \in N$. There is a useful stability concept which can be applied here: Regard g as a metric of uncertainty on M, if we replace μ by any measure $\hat{\mu}$ which is within ε (in the g metric) from μ then $\hat{\nu} = \pi\hat{\mu}$

should not be too far from ν . A very natural way of assessing this stability question is to compute the volatility

$$V = \int_M \mathrm{Tr}((d\pi(m))^* d\pi(m))\, d\mu(m)$$

of the map. (If the measure μ here were the Riemannian volume then geometers are very familiar with the function V and refer to it as the energy of the map π ; the extremals mappings of this for fixed metrics are known as "harmonic mappings", the nonlinear Euler Lagrange equation gives the tension of π at each point.)

The volatility of statistical maps can be used in at least two ways (i) in the design of experiments - to choose least tension giving the required answers, (ii) in cases where π, h are both fixed - to assess the confidence one has about $\pi(\nu)$ as being in reasonable inference about the value of $\pi(m)$. Calculation of this number could easily be incorporated into computer packages analysing real data.

V. Examples.

In the case of standard simple models sensible things happen: if one considers the location model of the normal - the metric is just a scalar multiple of the usual one - the staionary measure in (iii) is just the Gaussian measure one would expect and the volatility between the empirical metric and the usual metric on $\mathbb{R}$ is $\sim \frac{\text{sample variance}}{n}$.

Three particularly interesting examples are

(i) A mixture of normals where at least one mean and one variance is unconstrained. Here the empirical metric g gives a finite integral

$$\int_M p \, d\,\mathrm{Vol}_g$$

where the Fisher information metric gives an infinite answer.

(ii) the location model for a uniform distribution on an interval in $\mathbb{R}$ (suitably smoothed out). This is interesting because the model has atypical asymptotic behaviour. In this case the metric contracts the interval of possible choices of location (given the data) to approximately a point. This is because one has no information to distinguish the point within the viable integral. If one considers the inferential measure for the location in $\mathbb{R}$ then it has most of its mass at the ends of the interval. A computation of the volatility of the identity map from the empirical metric on $\mathbb{R}$ to the usual one would be very high. On the other hand if we compressed the interval to a point in the interest parameter space it would be very low. On the data available one has no business trying to distinguish points in the relevant interval - but one has great confidence that the parameter is in there somewhere.

(iii) This approach could obviously be applied to the Fisher Barens model and other lsightly tricky situations such as the Cauchy distribution. At the moment we have numerical calculations in these cases and the Cauchy case.

It is hoped that an expanded version of this paper with more details will appear soon.

A Differential-Geometric Approach to Approximate Nonlinear Filtering

B. Hanzon
Faculty of Mathematics and Informatics
Delft University of Technology
P.O. Box 356
2600 AJ Delft, The Netherlands

Abstract

The nonlinear filtering problem leads to a complicated stochastic partial differential equation: the DMZ-equation. Here an approximation method is suggested that is based on the differential geometric approach to statistics (cf. e.g. [Ama 85]). Consider a differentiable manifold of densities. The solution of the DMZ-equation will in general not remain within this manifold. However, if one projects the r.h.s. of the DMZ-equation orthogonally, using the Hellinger metric, onto the tangent space of the manifold, the solution of the resulting stochastic partial differential equation will lie on the manifold, (if one uses the McShane-Fisk-Stratonovich form of s.d.e.'s).

§ 1. The DMZ-equation

Let us consider a dynamical system of the following form (cf. [May II], [Da-Ma])

$$(1\text{-}1)\qquad \begin{aligned} &(\text{Itô})\ dx(t) = f(x(t),t)dt + G(x(t),t)d\beta(t) \\ &(\text{Itô})\ dy(t) = h(x(t),t)dt + d\eta(t) \end{aligned}$$

(We will write "(Itô)" if an Itô-stochastic differential equation is meant, while without this prescript a stochastic differential equation is to be interpreted as being in McShane-Fisk-Stratonovich-form (MFS-form), cf. [E1], p. 114).

In (1-1) the symbols have the following meaning: $x(t) \in \mathbb{R}^n$ is the state at time t; f is an n-vectorfunction, G(x,t) is an n x s matrix and β with $\beta(t) \in \mathbb{R}^s$ a Brownian motion process with expectation zero and an s x s diffusion matrix Q(t);

$y(t) \in \mathbb{R}^k$ is the stochastic measurement process, h is a k-vectorfunction and $\eta(t) \in \mathbb{R}^k$ a standard Brownian motion process, independent of $\beta(t)$. Suppose f, G and h are twice continuously differentiable. Let $p_x(\xi,t|\rho,t')$ denote the transition probability density function. Then this function satisfies the forward Kolmogorov equation (cf. e.g. [May II])

$$(1\text{-}2)\quad \frac{\partial}{\partial t} p_x(\xi,t|\rho,t') = \mathcal{L}(p_x(\xi,t|\rho,t'))$$
$$:= -\sum_{i=1}^{n} \frac{\partial}{\partial \xi_i}\{p_x(\xi,t|\rho,t')f_i(\xi,t)\}$$
$$+\frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2}{\partial\xi_i\partial\xi_j}\{p_x(\xi,t|\rho,t')[G(\xi,t)Q(t)G^T(\xi,t)]_{ij}\}$$

One can use the so called DMZ (Duncan-Mortensen-Zakai)-equation for an unnormalized version $q(x,t)$ of the conditional density of $x(t)$ given $\mathcal{Y}_t := \sigma\{y(s)|t_0 \le s \le t\}$. In MSF-form this reads (cf. [Da-Ma])

$$(1\text{-}3)\quad dq(x,t) = \{\mathcal{L}(q(x,t)) - \frac{1}{2}h(x,t)^T h(x,t)q(x,t)\}dt + h^T(x,t)q(t,x)dy$$

§ 2. Hellinger distance and Fisher metric

A metric on the set of finite (non-negative) measures on a measurable space $(\Omega,\mathcal{F})$ is the socalled Hellinger distance (cf. [Hnz 86], [Ama 85]), which is given by

$$(2\text{-}1)\quad d(\mu_1,\mu_2) := \|r_1 - r_2\|_{L^2(\lambda)} = \left(\int_\Omega (r_1 - r_2)^2 d\lambda\right)^{\frac{1}{2}}$$

where λ is any probability measure on $(\Omega,\mathcal{F})$ such that $\mu_1, \mu_2 << \lambda$, with densities $p_1 = \frac{d\mu_1}{d\lambda}$ and $p_2 = \frac{d\mu_2}{d\lambda}$ and $r_1 = \sqrt{p_1}$, $r_2 = \sqrt{p_2}$. (Note: [Ama 85] calls $\frac{1}{2}d(\mu_1,\mu_2)$ the Hellinger distance). The metric space of all nonnegative measures on $(\Omega,\mathcal{F})$ with the Hellinger distance will be denoted by $\mathcal{H} = \mathcal{H}(\Omega,\mathcal{F})$.

Now let us consider a set of probability measures $\{\mu(\theta)|\theta \in \Theta,\ \mu << \lambda\}$ and let $p(.;\theta) = d\mu(\theta)/d\lambda$ and $r(.;\theta) = \sqrt{p(.;\theta)}$. Suppose Θ is a finite dimensional differentiable manifold and suppose that the mapping

$r : \Theta \to L^2(\lambda)(\subset \mathcal{H})$, $\theta \to r(.;\theta)$ is a smooth embedding. Then Θ inherits a Riemannian metric from $L^2(\lambda)$. The Riemannian metric tensor is $\frac{1}{4}$ times the Fisher information matrix (cf. [Ama 85], [Hnz 86]):

$$(2\text{-}2)\quad \int \frac{\partial r}{\partial\theta}\frac{\partial r}{\partial\theta^T}d\lambda = \frac{1}{4}E_\theta\left[\frac{\partial \ln(p)}{\partial\theta}\frac{\partial \ln(p)}{\partial\theta^T}\right].$$

§ 3. Orthogonal projection of a tangent vector on the tangent space of a submanifold.

Consider the space $\{\tau|\tau = \ln p; \sqrt{p} \in L^2(\lambda),\ \forall x \in X : p(x) > 0\}$. Its tangent vectors will be denoted by $T \in L^2(p\lambda)$, etc. From (2-2) it follows that the Riemannian metric on this space is given by the formula

$$(3\text{-}1)\quad \langle T,T\rangle = \frac{1}{4}\int T^2\, p d\lambda = \frac{1}{4}E_p T^2,\ \int p d\lambda \quad , \quad E_p(f) := \int f p d\lambda / \int p d\lambda .$$

Now consider the embedding of Θ in this space, via $\tau = \ln p$. Let $N = \dim \Theta$ and let $\{T_i\}_{i=1}^N$ be an orthogonal basis of the tangent space of this embedding at the "point" $\tau(.;\theta)$. If $T \in L^2(p\lambda)$ is as above, then its orthogonal projection P(T) on the tangent space of (the embedded submanifold) Θ is given by the formula

$$(3\text{-}2)\quad P(T) = \sum_{i=1}^{N} \frac{\langle T_i,T\rangle}{\langle T_i,T_i\rangle}T_i = \sum_{i=1}^{N}\left(\frac{E_\theta T_i T}{E_\theta T_i^2}\right)T_i$$

(3-3) Remarks (i) It easily follows that P(T) is the element in the tangent space of the submanifold which has the property

$$(3\text{-}4)\quad E_\theta\{P(T)\}\{T - P(T)\} = 0.$$

Because the Fisher information metric is infinitesimally equivalent to the Kullback-information, one can say that within the tangent space of the submanifold, P(T) is the vector which contains maximal information about the vector T.

§ 4. Application to the DMZ-equation

The method is now the following. Select a finite dimensional (sub-) manifold $\{p(.;\theta)|\theta \in \Theta\}$ of densities of which one knows (or believes or hopes) that they approximate

well the true conditional densities of the problem and which contains the prior density. Instead of the DMZ-equation (2-6) one considers the "projected version"

$$(4\text{-}1)\quad dq_p(x,t) = P\{\mathcal{L}(q_p(x,t)) - \frac{1}{2}\, h(x,t)^T h(x,t) q_p(x,t)\}dt + \\ + P\{h^T(x,t) q_p(t,x)\}dy.$$

Because (2-6) and (4-1) are in MFS-form the solution of (4-1) will remain in the manifold for all time. (cf. [El] p.123). Because $\otimes$ is N-dimensional the partial differential equation (4-1) reduces in fact (at least locally) to an N-vector stochastic "ordinary" differential equation, i.e. to a finite dimensional filter.

An important special case is the one in which $\{p(.:\theta) | \theta \in \otimes\}$ is the manifold of (nondegenerate) Gaussian densities. After transforming the Gaussian variables $x = (x_1,\ldots x_n)^T$ to the standard N(0,1) form $\xi = (\xi_1,\ldots,\xi_n)^T$, $\xi = \mathcal{R}^{-\frac{1}{2}}(x - Ex)$, $\mathcal{R}$ the covariance matrix, one obtains an orthogonal basis $\{T_1,\ldots,T_N\} = \{1\} \cup \{\xi_i\}_i \cup \{\xi_j\xi_k - \delta_{jk}\}_{j,k}$. Here the constant function 1 is included because in the DMZ-equation and in (4-1) one works with <u>unnormalized</u> densities. The r.h.s. of (1-3) can now be written as a linear combination of $T_1,\ldots,T_N$ and an orthogonal part, which is annihilated by P; P leaves the linear combination of $T_1,\ldots,T_N$ invariant. Calculations can be done with the Hermite-polynomials. If the r.h.s. of (4-1) is <u>polynomial</u> the problem can be solved without any (explicit) integrations.

An alternative method of computation shows that in this case the finite dimensional filter is really a special type of assumed <u>density filter</u>. More generally for exponential densities one can expect to arrive at some "assumed density filter". The properties of the approximate filter described here are currently under investigation.

<u>References</u>

[Ama 85] S.-I. Amari, 'Differential-Geometrical Methods in Statistics', LNS 28, Springer Verlag, Berlin, 1985.

[Da-Ma] M.H.A. Davis and S.I. Marcus 'An Introduction to Nonlinear Filtering' pp. 53-75 in: M. Hazewinkel and J.C. Willems (eds), 'Stochastic Systems: The Mathematics of Filtering and Identification and Applications', Reidel, Dordrecht 1981.

[El] K.D. Elworthy 'Stochastic Differential Equations on Manifolds', Cambridge University Press, Cambridge 1982.

[Hnz 86] B. Hanzon 'Identifiability, Recursive Identification and Spaces of Linear Dynamical Systems', Doct. diss., Rotterdam 1986.

[May II] P.S. Maybeck, 'Stochastic Models, Estimation, and Control', vol. II, Academic Press, New York, 1982.

GEODESICS CONNECTED WITH THE FISHER METRIC ON THE MULTIVARIATE NORMAL MANIFOLD

P.S. Eriksen

Institute of Electronic Systems
Department of Mathematics and Computer Science
Aalborg University Centre, Strandvejen 19
DK 9000 Aalborg, Denmark

INTRODUCTION

The Riemannian manifold of multivariate normal distributions equipped with the Fisher information metric has been studied by several people, initiating with Rao (1945). The main emphasis has been put on determining the geodesic curves and the distance between two distributions. For the univariate model, -which is the Poincaré half-plane, - this was solved by Yoshizawa (1971) and Atkinson & Mitchell (1981), and for the p-variate normal with mean zero by James (1973) and Atkinson & Mitchell (1981). In this contribution we give a representation of the geodesic in the general case.

GEODESICS

Let $M = \{(\Sigma,\mu) | \Sigma \in PD(p), \mu \in \mathbb{R}^p\}$ represent the class of p-variate normal distributions, where μ is the mean vector and Σ is the covariance matrix, which is positive definite (PD). The Fisher information defines a Riemannian metric on M. In general, the geodesic curves are solutions of the differential equation

$$\ddot{\Sigma} + \dot{\mu}\dot{\mu}^* - \dot{\Sigma}\,\Sigma^{-1}\,\dot{\Sigma} = 0$$

$$\ddot{\mu} - \dot{\Sigma}\,\Sigma^{-1}\,\dot{\mu} = 0 \qquad (1)$$

Since M is isometric to $GA^+(p)/SO(p) - GA^+(p)$ denoting positive affine transformations - we can by a translation argument restrict ourselves to describe the geodesic through $(I_p,0)$, - and in the direction (B,x), say. By considering the parameterization $\Delta = \Sigma^{-1}$ and $\delta = \Sigma^{-1}\mu$, it is fairly easy to show that the geodesic equation (1), then can be reformulated as

$$\begin{aligned}
\dot{\Delta} &= -B\Delta + x\delta^* \\
\dot{\delta} &= -B\delta + (1+\delta^*\Delta^{-1}\delta)x \qquad (2)\\
\Delta(0) &= I, \delta(0) = 0
\end{aligned}$$

Now let

$$A = \begin{Bmatrix} -B & x & 0 \\ x^* & 0 & -x^* \\ 0 & -x & B \end{Bmatrix}$$

and

$$\Lambda = \exp(At) = \sum_{n=0}^{\infty} (At)^n/n! \qquad (3)$$

$$= \begin{Bmatrix} \Delta & \delta & \Phi \\ \delta^* & \varepsilon & \gamma^* \\ \Phi^* & \gamma & \Gamma \end{Bmatrix} \qquad t \in \mathbb{R}$$

Then we have

Theorem The geodesic curve through $(I,0)$ with tangent $(-B,x)$ is given by $(\Delta(t), \delta(t))$, where $(\Delta(t), \delta(t))$ is determined by (3).

Proof It is clear that $\dot{\Lambda} = A\Lambda$, i.e.

$$\begin{Bmatrix} \dot{\Delta} & \dot{\delta} & \dot{\Phi} \\ \dot{\delta}^* & \dot{\varepsilon} & \dot{\gamma}^* \\ \dot{\Phi}^* & \dot{\gamma} & \dot{\Gamma} \end{Bmatrix} = \begin{Bmatrix} -B & x & 0 \\ x^* & 0 & -x^* \\ 0 & -x & B \end{Bmatrix} \begin{Bmatrix} \Delta & \delta & \Phi \\ \delta^* & \varepsilon & \gamma^* \\ \Phi^* & \gamma & \Gamma \end{Bmatrix} \qquad (4)$$

It follows that

$$\begin{aligned}
\dot{\Delta} &= -B\Delta + x\delta^* \\
\dot{\delta} &= -B\delta + \varepsilon x
\end{aligned}$$

If $\varepsilon = 1+\delta^*\Delta^{-1}\delta$ then (Δ,δ) is seen to be a solution to (2). Since $\varepsilon(0)=1$ it clearly suffices to show that $\dot{\varepsilon}=\dot{\sigma}$, where $\sigma=\delta^*\Delta^{-1}\delta$.

By (4) we have that

$$\begin{aligned}
\dot{\varepsilon} &= x^*\delta - x^*\gamma \\
\dot{\sigma} &= x^*\delta - x^*([\sigma-\varepsilon]\Delta^{-1}\delta + \Phi^*\Delta^{-1}\delta)
\end{aligned}$$

where it has been utilized that $\dot{\delta}^* = x^*\Delta - x^*\Phi^*$. It follows that $\dot{\varepsilon}=\dot{\sigma}$ if

$$\gamma = [\sigma-\varepsilon]\Delta^{-1}\delta + \Phi^*\Delta^{-1}\delta$$

This is established by the relation $\Lambda\,\Lambda^{-1} = I$, where we observe that $\Lambda^{-1} = \exp(-At)$, i.e.

$$\Lambda^{-1} = \Lambda(-t) = \begin{Bmatrix} \Gamma & \gamma & \Phi^* \\ \gamma^* & \varepsilon & \delta^* \\ \Phi & \delta & \Delta \end{Bmatrix} \qquad (5)$$

The relation $\Lambda\,\Lambda^{-1} = I$ implies that

(i) $\Delta\gamma + \varepsilon\delta + \Phi\delta = 0$

(ii) $\Delta\Phi^* + \delta\delta^* + \Phi\Delta = 0$

Clearly, (ii) is equivalent to $\Phi^*\Delta^{-1} + \Delta^{-1}\delta\delta^*\Delta^{-1} + \Delta^{-1}\Phi = 0$, so that

(iii) $\Phi^*\Delta^{-1}\delta + \sigma\Delta^{-1}\delta + \Delta^{-1}\Phi\delta = 0$

Furthermore, (i) is equivalent to

(iv) $\gamma = -\varepsilon\Delta^{-1}\delta - \Delta^{-1}\Phi\delta$

and it is now obvious from (iii) and (iv) that

$$\gamma = [\sigma-\varepsilon]\ \Delta^{-1}\delta + \Phi^*\Delta^{-1}\delta.$$

This completes the proof.

DISCUSSION

It is interesting to note that (5) means that $(\Delta(-t),\delta(-t)) = (\Gamma(t),\gamma(t))$, i.e. (Γ,γ) is the "point opposite" to (Δ,δ). Besides we have shown that $\varepsilon = 1+\delta^*\Delta^{-1}\delta = 1+\gamma^*\Gamma^{-1}\gamma$. Considering the representation (3) of the solution, there has been no success in seeking for an interpretation of the p×p matrix Φ.

Further problems:

i) Uniqueness of geodesic connecting two points?

ii) Determination of distance.

iii) Any relation to SO(p,p+1), which is the least group containing the Λ's defined by (3)?

iv) Characterization and statistical interpretation - if any - of totally geodesic submanifolds.

REFERENCES

Atkinson, C. and Mitchell A.F.S. (1981): Rao's distance measure. *Sankhyá, 43*, 345-365.

James, A.T. (1973): The variance information manifold and the functions on it. *Proc. 3rd Int. Symp. on Mult. Anal.*, Krishnaiah (ed.). Academic Press.

Rao, C.R. (1945): Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Math. Soc., 37*, 81-91.

Skovgaard, L.T. (1984): A Riemannian geometry of the multivariate normal model. *Scand. J. Statist., 11*, 211-223.

Yoshizawa, T. (1971): A geometry of parameter space and its statistical interpretation. Memo TYH-2, Harvard University.

Connections and Expectation values in Field Theories

R W Tucker
Department of Physics
University of Lancaster, Lancaster LA1 4YB

Abstract

A brief pedagogical discussion is given of the role played by certain connection forms in classical gauge field theory and expectations values in the interpretation of quantum field theory.

Classical Field Theory

Many modern developments in theoretical physics have benefited greatly from geometrical formulations. The basic interactions in nature, classified in terms of Lie groups, are mediated by gauge fields described mathematically as sections of a principal fibre bundle over a space-time manifold. Such gauge fields are coupled to "charged" matter in a manner ensuring that the dynamical field equations are "covariant" under an arbitrary choice of gauge section. Such a covariance may be established by encoding "charged" matter into sections of vector bundles associated to the corresponding princpal bundle. Maxwell's theory of electromagnetism has been successfully incorporated into the electroweak model in which the local gauge group is $SU(2) \times U(1)$. The force that binds quarks together in nuclei is thought to be described in terms of an $SU(3)/Z_3$ gauge group. Grand unified theories (GUT) that incorporate both the strong and electroweak physics are believed to have their origin in some larger gauge group G that possesses $SU(3)/Z_3 \times SU(2) \times U(1)$ as a subgroup. Gravitation is currently described in terms of the curvature of the base space-time manifold. Since the light-cone structure of space-time is regarded to be of fundamental importance and special relativitiy a good symmetry in the absence of gravitation such interactions are developed in terms of a metric compatible connection on the bundle of ortho-normal tangent frames and a Lorentzian space-time metric (which is not necessarily torsion free). Under certain mild topologocial constraints on the space-time manifold the Lorentz structure group $SO(3,1)$ can be lifted to its covering and a connection established on a principal $Spin(3,1)$ bundle. Sections of the associated bundle to this are called spinor fields and describe the quark and lepton matter fields that constitute some of the primary fields in all gauge models. The following table summarises the basic interactions and their associated symmetry groups.

Theory	Gauge group	Lie algebra-valued 1-form connection on space time	Lie algebra-valued 2-form field strength
Maxwell Electromagnetism	U(1)	A	$F = dA$
Electroweak	SU(2)×U(1)	$\mathbf{A}_{SU(2)} + \mathbf{A}_{U(1)}$	$\mathbf{F} = (d\mathbf{A}+[\mathbf{A},\mathbf{A}])_{SU(2)} + d\mathbf{A}_{U(1)}$
Strong (colour)	SU(3)	$\mathbf{A}_{SU(3)}$	$\mathbf{F} = (d\mathbf{A}+[\mathbf{A}, \mathbf{A}])_{SU(3)}$
GUT	G	$\mathbf{A}_G$	$\mathbf{F} = (d\mathbf{A}+[\mathbf{A},\mathbf{A}])_G$
Gravitation	Spin(3,1)	$\omega^a{}_b$	$\mathbf{R} = d\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b$

In this table the bracket between a Lie algebra-valued p-form **H** and a Lie algebra-

valued q-form $\mathbf{B}$ is $[\mathbf{H},\mathbf{B}] \equiv (\mathbf{H}\wedge\mathbf{B} - (-\mathbf{1})^{pq}\mathbf{B}\wedge\mathbf{H})/2$.

For gravitation the connection is Spin(3,1) valued and the connection form defines ∇ in terms of any frame $\{X_a\}$ by $\nabla_{X_a}X_b = -\omega_b{}^c(X_a)X_c$. Since all physical fields on space-time are either sections of a tensor or spinor bundle, Lorentzian covariant combinations may be constructed in terms of the space-time metric and induced spinor inner-product. (An efficient calculus that embodies these ideas is best approached in the language of Clifford bundles [1].) In the (idealised) limit in which the electroweak symmetry is regarded as exact the interaction between matter is introduced by constructing covariant (exterior) derivatives of matter fields. Thus for a quark spinor field ξ ,

$$D\xi = d\xi + \omega\circ\xi + (\mathbf{A}_{SU(2)}\times A_{U(1)})\circ\xi + \mathbf{A}_{SU(3)}\circ\xi$$

Here $\circ$ denotes the action of the preceeding Lie algebra-valued connection form on a basic representation of the associated group. Thus the second term on the right of the above equation incorporates the gravitational coupling (quarks carry intrinsic spin), the third and fourth terms the electric and weak charge couplings while the last term signifies the "colour" or gluonic interaction. The basic lepton spinors λ form doublets of chiral (lefthanded) electrons and neutrinos or righthanded electrons singlets. The covariant derivative is

$$D\lambda = d\lambda + \omega\circ\lambda + (\mathbf{A}_{SU(2)}\times A_{U(1)})\circ\lambda$$

Thus leptons carry spin and electroweak charge but no overt colour interaction. In flat space-time , experiment has motivated a pattern of couplings for the basic fields $\xi, \lambda, \omega, A_{U(1)}, \mathbf{A}_{SU(2)}, \mathbf{A}_{SU(3)}$ which provides a viable phenomenology [2] for a considerable body of high energy physics when couched in the language of quantum field theory. It is the purpose of current research to seek a viable unifying Group that also incorporates gravitation in a way that renders the quantum field theory meaningful.

Quantum Field Theory

In general, field quantisation is a scheme in which the quantum dynamics of space-time fields (motivated by a system of tensor or spinor equations on space-time) is constructed in terms of an algebra of *bounded observables*. Positive normalisable linear functionals on such observables constitute a space of *states*, the action of the state on any such observable being called the *expectation* of the observable in that state. If any observable $\mathcal{O}$ can be represented by a bounded Hermitian operator in a Hilbert space $\mathcal{H}$ of *state vectors* $\{f_i\}$ with inner product $(\ ,\ )_{\mathcal{H}}$, then the state functional may be defined by the expectation $(f,\mathcal{O}f)_{\mathcal{H}}$ where $(f,f)_{\mathcal{H}} = 1$. In general there is no preferred Hilbert space on which to represent the field algebra and consequently the choice of Hilbert space is often based on physical requirements. It is this choice that often dictates the subsequent physical interpretation of the theory and permits extraneous notions (such as observers and detectors) to enter the picture.

Field theories based on linear field equations present the most primitive systems to analyse. The space of classical solutions is mapped into an algebra of observables in which commutators or anticommutators are in the centre. If the space-time M admits time-like Killing vectors it may be possible to represent the observables on a Hilbert space of mode-functions that separate the field equations in some domain $U \subset M$. Such a representation may be labelled by a time-like Killing vector field V and a space-like 3-chain $\Sigma : [0,1]^3 \mapsto U$.

To illustrate their role consider for simplicitly the real scalar field ϕ satisfying

$$d{*}d\phi = 0 \qquad .1$$

in Minkowski space. Denote by $(\ ,\ )_\Sigma$ the Hilbert space inner product defined by

$$(\phi_1,\phi_2)_\Sigma = \int_\Sigma j[\phi_1,\phi_2]. \qquad .2$$

The 3-form

$$j = i(\phi_1{*}d\phi_2 - \phi_2{*}d\phi_1) \qquad .3$$

and is closed for solutions to .1 on U. Let .1 be separated in terms of complex eigenfuctions of V and denote by $\{\phi^\pm_{\mathbf{n}}\}$ an ortho-normalised basis for $\mathcal{H}$. (The index n is a triplet mode separation label and $V\phi^\pm_{\mathbf{n}} = \pm E_{\mathbf{n}}\phi^\pm_{\mathbf{n}}$ in terms of the n th V-mode energy.)

$$(\phi^+_{\mathbf{n}},\phi^-_{\mathbf{m}})_\Sigma = -(\phi^-_{\mathbf{n}},\phi^+_{\mathbf{m}})_\Sigma = \delta_{\mathbf{nm}}$$
$$(\phi^+_{\mathbf{n}},\phi^+_{\mathbf{m}})_\Sigma = (\phi^-_{\mathbf{n}},\phi^-_{\mathbf{m}})_\Sigma = 0. \qquad .4$$

If the mode-functions are chosen appropriately on Σ we have a mapping from classical solutions to self-adjoint operators in $\mathcal{H}$. In the traditional notation for a quantum-field one introduces the local symbol:

$$\hat\phi = \sum_{\mathbf{m}} \hat A_{\mathbf{m}}(V)\phi^+_{\mathbf{m}} + \hat A^*_{\mathbf{m}}(V)\phi^-_{\mathbf{m}} \qquad .5$$

from which relations involving the operator $\Phi(f)$ can be deduced using the symbolic relation:

$$\Phi(f) = \int_\Sigma \hat\phi f \hat{*} 1. \qquad .6$$

In this expresion f is a suitably smooth 0-form on U and $\hat{*}$ is defined with respect to the induced metric on Σ. The choice of field algebra is motivated by the mode decomposition of the classical free Hamiltonian H_V defined by the pair (V,Σ). For any classical solution ϕ:

$$H_V = \int_\Sigma j(\mathcal{L}_V\phi,\phi) \equiv \frac{1}{i}(\mathcal{L}_V\phi,\phi)_\Sigma. \qquad .7$$

Since

$$H_V = \sum_{\mathbf{m}}(A_{\mathbf{m}}(V)A^*_{\mathbf{m}}(V) + A^*_{\mathbf{m}}(V)A_{\mathbf{m}}(V)) \qquad .8$$

the field-algebra is put into correspondence with the oscillator-algebra defined by

$$[\hat A_{\mathbf{m}}(V),\hat A^*_{\mathbf{n}}(V)]_- = \delta_{\mathbf{mn}} \qquad .9$$

and represented on a Fock space built on an equilibrium or vacuum-state vector $|0:V,\Sigma>$ defined by

$$A_{\mathbf{n}}|0:V,\Sigma>=0 \qquad .10$$

for all modes. $\hat H_V$ is chosen to have the ground-state eigenvalue for this state vector. The vectors $|\mathbf{n_j}:V,\Sigma> \equiv \Pi_{j_r}\{A^*_{\mathbf{n}_r}\}^{j_r}|0:V,\Sigma>$ are usually referred to as many-particle (or quantum) excitations of the equilibrium state. There are many representations (and consequently many equilibrium state vectors) on Σ. Certainly

one should not expect in general that quantum interpretations based on different vector fields V be physically indistinguishable. One might advance the hypothesis that a V-*quantum* of the field can be identified if a suitable detector can be constructed that responds to individual modes that are stationary with respect to the time-like vector field V. Such quanta populate the Fock space built from the equilibrium state vector $|0 : V, \Sigma>$ which should describe the quiescent state of such a field-detector system. For an idealised (world-line) detector the vector-field V can be identified with its world-velocity although for a realistic (spatially extended) detector in general motion the above hypothesis would need refining.

If W is another time-like Killing vector and $\{\rho_n^{\pm}\}$ an associated complete orthobasis of mode-functions resulting from a separation of .1 on Σ with respect to a different co-ordinate choice on U, one may symbolically relate the different representations by writing :

$$\hat{\phi} = \sum_n \hat{a}_n(W)\rho_n^+ + \hat{a}_n^*(W)\rho_n^- \qquad .11$$

and defining the field-algebra analogously from H_W. Consequently the mode operators are related by:

$$(\hat{\phi}, \rho_n^-)_\Sigma = \hat{a}_n(W) = \sum_m [\hat{A}_m(V)(\phi_m^+, \rho_n^-)_\Sigma + \hat{A}_m^*(V)(\phi_m^-, \rho_n^-)_\Sigma]$$

$$(\hat{\phi}, \rho_n^+)_\Sigma = -\hat{a}_n^*(W) = \sum_m [\hat{A}_m(V)(\phi_m^+, \rho_n^+)_\Sigma + \hat{A}_m^*(V)(\phi_m^-, \rho_n^+)_\Sigma] \qquad .12$$

A much studied example is $W = \partial/\partial t$ in the global chart (t, x^i) for U with Minkowski metric $g = -dt \otimes dt + \sum_i dx^i \otimes dx^i$ and $t = t_0$, the image of Σ. A non-inertial quantisation is based on $V = \partial/\partial\eta$ in the local chart (η, ξ', y, z) with Minkowski metric $g = \exp(2\xi')\{-d\eta \otimes d\eta + d\xi' \otimes d\xi'\} + dy \otimes dy + dz \otimes dz$ on a Rindler wedge with $-\pi < \xi' < \pi$. The local field $\{\partial/\partial\eta\}$ has integral curves that can model a field of ideal uniformly accelerating detectors [**3**].

REFERENCES

1. I M Benn, R W Tucker, An Introduction to Spinors and Geometry with Applications in Physics. Adam Hilger (In Press).
2. D Bailin, A Love, Introduction to Gauge Field Theory. Adam Hilger 1986.
3. W Unruh, R M Wald, Phys. Rev. **D29** (1984), p. 1047.

A SUMMARY OF POINTS RAISED IN DISCUSSION SESSIONS

C.T.J.Dodson, P.E.Jupp, W.S.Kendall and S.L.Lauritzen

GENERAL STATISTICAL MANIFOLDS

Following the paper of Lauritzen, two avenues of further enquiry were suggested by Amari:

(i) study the extent to which fundamental properties such as flatness and conjugate symmetry were reflected by properties of the two embeddings $\lambda:\Omega \to V$ and $\pi:\Omega \to W$;

(ii) for Riemannian (M,g) investigate the existence of a skewness tensor D such that it 'flattens' the Levi-Civita connection ∇^g in the sense that there is a flat ∇ with

$$g(\nabla^g_X Y - \nabla_X Y\ ,\ Z) = \tfrac{1}{2}D(X,Y,Z).$$

We note the fact that every trivial principal G-bundle MxG admits a canonical flat connection. Moreover, a connection in a principal G-bundle (P,G,M) (e.g. a frame bundle) is flat if and only if every point in the base M admits an open neighbourhood U such that the induced trivialization yields bundle morphisms which project horizontal subspaces onto the canonical horizontal distribution of the flat connection for UxG; this is equivalent to the connection having identically zero curvature. Conversely, if M is paracompact and simply connected, with a flat connection ∇ in a principal G-bundle (P,G,M) over M, then $P \cong MxG$ and this

isomorphism maps horizontal subspaces of ∇ onto those of the canonical flat connection. Geometrically, flatness of a connection corresponds to local freedom of parallel transport from dependence on the curve along which the transport is effected. (Cf. Kobayashi and Nomizu [13] p.92 et seq.)

It was pointed out by Dodson that geometers have amassed a large body of results on locally symmetric Riemannian manifolds; these have parallel Riemannian curvature i.e. $\nabla^g R = 0$. Some new theorems of Dodson, Vanhecke and Vazquez-Abal characterise locally symmetric manifolds by harmonicity of local geodesic symmetries [8], isolate geodesics as curves in which reflections are harmonic and characterise constant Riemannian curvature as precisely the situation when reflections in all geodesics are harmonic [24]. Wolf [27] gives the classification of homogeneous Riemannian manifolds of constant curvature; up to isometry the ones with positive constant curvature are controlled by their fundamental groups. Wolf discusses also locally symmetric Riemannian manifolds. Attention was drawn by Lauritzen to the interest in statistical manifolds with constant curvature, but he and Ross emphasised the different interpretation that must be employed for curvatures of other connections than the Levi-Civita one.

Relating to the projection theorem of Lauritzen and Picard in §5 of Lauritzen's paper, Kendall suggested an approach via Jacobi fields to generalise the result. He observed also that in §3 was

in implicit seach for a generalization of the Nash C^k isometric embedding theorem. (Briefly, $\mathbb{R}^N$ with $N = \frac{1}{2}n(3n+11)$ is sufficient for compact and $\mathbb{R}^N$ with $N=\frac{1}{2}n(n+1)(3n+11)$ is sufficient for noncompact C^k Riemannian n-manifolds with $3 \leq k \leq \infty$; in particular;

$\mathbb{R}^{17}$ is sufficient to embed isometrically any compact 2-manifold. For further details see Chen [4] and Griffiths and Jensen [9]. Concerning global geometric procedures in statistical inference Lyons noted that the Lauritzen-Picard criterion on uniqueness of projections is one of only a few examples of global geometric considerations. Why are there no more? A test case for thinking globally would be to start with a finite parameter space. Since indefinite metric tensors have appeared in the context of statistical manifolds, Dodson drew attention to the fact that isometric embeddings in those cases may need many more dimensions. An important embedding theorem for pseudo-Riemannian manifolds was proved by Clarke [5]. Any extension of the Nash theorem would probably be difficult but the statistical problem may be easier because finite-dimensionality was not necessarily a restriction. Lauritzen mentioned that the projection theorem generalized under certain conditions to non flat spaces, but then the curvature became involved. At present, these 'certain conditions' did not appear to have a geometric interpretation but the idea of using Jacobi fields might lead to one. We note that Jacobi fields quantify the divergence of

nearby geodesics, two useful references are Postnikov [19] and Hicks [11].

In the general context of manifolds with a connection not necessarily metric, Dodson referred to the authoritative work of Weyl [25] which has an annotated bibliography to early research on differential geometry and its application to physics. Meanwhile, Foster advanced the opinion that connections could be avoided altogether if his acceleration-sensitive distance-like function was employed. He agreed that it was parameter dependent but unlike Tucker and Dodson did not view this as a serious disadvantage. Some recent work of Dodson and Radivoiovici [7] showed that the fibre bundle of curves equivalent up to acceleration formed a vector bundle if and only if there was a linear connection and then the acceleration bundle splits into two isomorphs of the tangent bundle. In that case accelerations have invariant definition as $\nabla_{\dot{c}}\,\dot{c}$.

Picard remarked that the set of Markov-invariant metrics for the experiment $\{N(\mu,\sigma^2) = (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}^+\}$ is precisely the set of Fisher information metrics of symmetric location-scale families, those with p-d-f's $\sigma^{-1}f(x-\mu)/\sigma$ with f even. (See also Mitchell [16]). It is unclear what is the significance of this. Amari pointed out that the $N(\mu,\sigma^2)$ family mentioned by Picard is a transformation model. In such a model, a Lie group G acts on both the sample space and the parameter space Θ so that the p-d-f's satisfy $f(gx;g\Theta) = f(x;\Theta)$. Then the Fisher information metric and

α-connections are invariant under the G-action on Θ. If G acts transitively on Θ then the metric and skewness tensors are determined by their values at any given point. If G acts freely and transitively on Θ then the tangent map of left translation on G yields the parallel translation operator of a flat affine connection on Θ (which has non-zero torsion unless G is abelian). What is its statistical relevance? The new work of Mackenzie on Lie algebroids, referred to in Dodson's paper, should yield applications to transformation models.

Given an incomplete Riemannian manifold then a sufficiently large conformal transformation will make it complete. For statistical manifolds Lauritzen raised the question of classifying the ones which can be made flat by a conformal transformation. Conformal geometry was related to sequential statistical inference by Amari. This corresponds to the case when the number of observations $n(\Theta)$ depends on the point $\Theta\in M$. Then the statistical manifold (M,g,D) is transformed conformally to $(M,\sigma g,\sigma D)$ where $n(\Theta)=\sigma(\Theta)N$. By using the sequential stepping rule, where $n(\hat{\Theta})$ is a random variable depending on $\hat{\Theta}$ or $x_1,x_2,\ldots$, we can realize the conformal structure. Another viewpoint of conformal transformations is given by an encoder of random processes, which changes the length of a sequence. In Dodson's paper, the system of α-connections is enlarged to systems involving both variable α and conformal changes of metric. Are there characterizations of these connections in terms of almost Markov

morphisms of degree k?

We draw attention to Note 11 on p.309 of Kobayashi and Nomizu [13], which discusses conformal transformation, summarising important theorems with references to the literature. The α-connections also give rise to embeddings of parametric models as Riemannian slices of a system. Can these be expolited here? There are similar characterizations for other geometrics because every connection determines an embedding and the universal connection for the system determines a unique Riemannian structure on the embedded copy.

METHODOLOGY

The problem encountered when the Fisher information diverges was raised by Amari. It is not clear what geometric structure is appropriate in such cases. For example, consider the location problem, with likelihood $f(x-\theta)$, when f has compact support. Reasonable estimators $\hat{\theta}$ seem to lead to asymoptotic stable laws for renormalizations of $(\hat{\theta}-\theta)$. He suggested, since Finsler geometry, stable distributions and irregular models have common features, that a Finsler metric may be appropriate for the parameter space $-\infty<\theta<\infty$. In the case of the 'empirical metric' discussed by Lyons, the (smoothed) metric $g^{(n)}$ depends on the n observations. Analysis will be sensitive to the rate at which ϵ tends to 0. However it seems likely that the 1-forms of this analysis will lead to random variables with divergent variance.

Attention was drawn by Barndorff-Nielsen to work in the literature on semi-parametric models. Dodson mentioned that the next Finsler space conference would be organized by Professor Atanasiu at the University of Braşov in Transylvania, Romania, 10-15 February 1988. On the presence of singularities and general connection incompleteness, he pointed out that there is a method of handling the problem, even in the absence of a metric. The technique uses any given linear connection to provide a natural Riemannian metric on the frame bundle and was devised to circumvent the indefiniteness of Lorentz metrics on spacetime. If the base manifold does have a Riemannian metric and its Levi-Civita connection is used to induce the metric on the frame bundle, then the usual results are regained for geodesic incompleteness. In the statistical situation, Kendall observed that geodesic incompleteness corresponds to breakdown of estimation; it was agreed that a good example would be useful. Geodesic incompleteness is stable and has been studied recently (cf. references given in Dodson's paper). In his thesis, P.M.Williams [26] showed that, with the right topology, the Levi-Civita map from metrics to connections and the symmetrization map on connections are both continuous. Moreover, for compact manifolds the space of symmetric connections is simply connected. It follows [6] that the symmetrization map on connections is a retraction so symmetric connections form a closed set, and hence having non zero torsion is stable.

Following the paper of Jupp, the question was raised: Which pseudolikelihood functions on the manifold of parameters of interest should we use? Considerations of maximum likelihood support the use of profile likelihood. However, there is evidence (e.g. Smith and Naylor [23]) that integrated likelihood has better sampling properties than profile likelihood. Also, it should be more stable numerically. Instead of insisting that the transfer of a tensor is well-defined, one could integrate over the fibres of π.

A number of points arose from the paper of Lyons:

How does his metric compare with

(i) the observed information metric $-\sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \theta^2}(\hat{\theta}; x_i)$,

(ii) the metric $i(\theta)^2 \{\sum_{i=1}^{n} (\frac{\partial l_i}{\partial \theta}(\theta; x_i))^2\}^{-1}$

introduced by Royall [21] to obtain robust confidence intervals?

His methods may be useful also for finite parameter spaces, or condensation onto them. Lyons' metric is used in certain established optimization procedures. Are there other metrics in computational use which would be of statistical interest? What can be said about the influence of extra data points on the metric? What is the effect of further structure (e.g. a 2-way layout) on the data points? Lauritzen suggested the following approach to this question. Consider the case of a two-way layout (X_{ij}), in which the first set and second set of indexes are both subject to symmetry conditions. If $(d\ln_{ij})$ are the corresponding

1-forms then the general invariant quadratic form in the 1-forms is

$$\alpha \Sigma_{ij} (d\ln_{ij} - d\ln_{i\cdot} - d\ln_{\cdot j} + d\ln_{\cdot\cdot})^2$$
$$+ \Sigma_i (d\ln_{i\cdot} - d\ln_{\cdot\cdot})^2 + \gamma \Sigma_j (d\ln_{\cdot j} - d\ln_{\cdot\cdot})^2$$
$$+ \delta (d\ln_{\cdot\cdot})^2$$

and one needs to know what principles to adopt in choosing particular values for the constants $\alpha, \beta, \gamma, \delta$. Otherwise there appears to be arbitrariness in the choice of 'empirical metric'. It was noted that the energy function used by Lyons for a submersion $\pi : M \to K$ measures departure of the map π from being harmonic, on which subject there is an extensive literature. Perhaps experiments should be designed to vary the metric on K in order to minimise energy. It seems sensible to choose the interest parameters as *stable* parameters (those which can be estimated most precisely). How is this related to choosing π to be (nearly) harmonic?

Picard showed that Fisher information and the α-connections together capture power properties of likelihood ratio tests up to the second order. What geometrical objects are needed to capture higher order properties? Picard's results are related to the expansion in $n^{-\frac{1}{2}}$ of the power of the likelihood ratio test of $\theta=\theta_0$ against the local alternative $\theta=\theta_0 + n^{-\frac{1}{2}}u$. Similarly, one could expand the Le Cam [3] deficiency between the repeated experiment and the corresponding set of limiting Normal distributions. How do these compare?

Markov-equivalence of experiments is closely related to a decision-theoretic optimality property. Do other forms of optimality lead to similar characterizations?

The definition of statistical morphism of degree k involves a concept of k^{th} order local equivalence of experiments. Can we then usefully regard experiments as sections of a fibred space of equivalence classes? How could we glue local results together? Dodson's system of conformal α-connections should allow generalization of the results of Picard in selection of metrics by invariance, through extension from scaling to conformal equivalence. In the context of yet more geometrical structure, Barndorff-Nielsen and Lauritzen exchanged views on the following. The dual connection structure allows one to view not just $0(n^{-\frac{1}{2}})$ but $0(n^{-1})$ approximations geometrically. For both practical and theoretical reasons (taking account of kurtosis, and the so-called p*-formula) it would be useful to go further, to $0(n^{-3/2})$. What geometry is induced by this? Picard's local approximations should be helpful here. On another tack, the following tensor occurs in Bartlett adjustments, for example. Let $g=g(w;\tilde{w})$ be a yoke. The tensor is given by

$$h_{i_1i_2j_1j_2} = g_{i_1i_2;g_1j_2} - g_{i_1i_2i}\, g_{j;j_1j_2}\, g^{i;j}$$

when the lower-suffix terms in the second summand relate to $\alpha=1$ and $\alpha=-1$ connections respectively. There are relations to connection curvature. Karmakhar's interior method [12] for linear programming was mentioned by Amari. K.Tanabe suggested the

construction of a convex function using the sum of the logs of the constraint deficits. By this means one obtains a statistical or dual differential geometry producing inferior trajectories in a continuous variant of Karmakhar's solution.

A stastistical geometry (g,D) gives a diffusion, Brownian motion with drift, obtained by using the g-Laplace-Beltrami operator and adding a drift related to contraction of D by g. That this might be connected to some continuous version of the maximum likelihood process of estimation was suggested by Jupp.

In connection with the paper of Hanzon, Kendall drew attention to the work of Goldstein [10]. On the analysis of shape discussed by Kendall, Dodson mentioned the work of Selig [22]. He used spectral sequence methods to evaluate the cohomology of a shape space modulo symmetries in various cases. Other unpublished work of D.G.Kendall and Hui-Ling Le discusses the differential geometry of several general cases.

Amari and Lyons discussed parametrization by location along the following lines. Consider distributions of compact support on $\mathbb{R}$ and parametrized by location. Other estimators $\hat{\theta}$ of θ can be expressed

$$E((\hat{\theta}-\theta)^2) = g^{-1}\epsilon/k_1(n) + g^{-1}\epsilon^2/k_2(n) + \ldots$$

where ϵ is a smoothing parameter depending on n. The asymptotics of k_1 and k_2 are sensitive to how sharply the original and smoothed densities converge to zero; that is, on how fast $\epsilon(n) \to 0$ as $n \to \infty$. Indeed, the law of $(\hat{\theta}-\theta)$ appropriately

renormalized should converge to some stable distribution. For the case of the empirical metric, the metric $g_\epsilon^{(n)}$ depends on the n observations and Lyons conjectures that in most cases we have

$$E((\hat{\theta}-\theta)^2) = g_\epsilon^{(n)-1} + g_\epsilon^{-1}/k_2(n).$$

So presumably, $\nabla \log p_i$ is a random variable without finite variance as $\epsilon \to 0$.

References to recent work on statistics and information theory were requested, and supplied by Amari as [1],[2],[28].

Barndorff-Nielsen suggested the following as a possible framework for a coordinate-free definition of strings as new-tensors.

Let E and F be sets and FxF → F be an associative binary operation.

An *(E,F)-tool* is a function K: ExE → F satisfying

$$K(e_1,e_3) = K(e_1,e_2)\ K(e_2,e_3).$$

A *representation* of F is a composition-preserving map $\varphi : F \to GL(n)$. The corresponding *representation* of K is the (E,GL(n))-tool $L = \varphi.K$. *e.g.* For M a smooth manifold and $P \in M$, $[M]_p$ denotes the set of germs at p of local coordinate systems (embeddings $f:(V,p) \to (\mathbb{R}^d,0)$ for V a neighbourhood of p). Let Φ denote the set of germs at 0 of local diffeomorphisms $(\mathbb{R}^d,0) \to (\mathbb{R}^d,0)$. Composition is a binary operation on Φ. The *basic germ tool at p* is the $([M]_p,\Phi)$-tool

$H_p : [M]_p \times [M]_p \to \Phi$ defined by

$H_p([\bar{\omega}],[\bar{\psi}]) = [\bar{\omega} o\ \bar{\psi}^{-1}]$.

An object $\bar{M}$ which assigns to each local coordinate system $\bar{\omega}$ a function $\bar{M}(\bar{\omega}) : \mathrm{dom}(\bar{\omega}) \to \mathbb{R}^n$ will be regarded as defining a *geometric object* if for all p in M there is a $([M]_p,GL(n))$-tool L_p such that

(i) L_p is a representation of the basic germ tool H_p at p,

(ii) $\bar{M}(\bar{\omega})_p = \bar{M}(\bar{\psi})_p\ L_p([\bar{\psi}],[\bar{\omega}])$.

Could we use ∞-jets instead of germs? What is the common ground between jets and strings? Do quotients of jet spaces play a role in strings? Reference was made to recent new developments in jet calculus (cf. Mangiarotti and Modugno [14]) which may provide the appropriate vehicle for derivative strings.

The following point was emphasied by Barndorff-Nielsen : It is not always desirable to make procedures invariant, if that obscures real distinctions between various parameters! The emphasis should be on *what* to geometrize in statistics, as well as *how*.

The possibility of general results on splitting invariant terms was raised as a problem by Barndorff-Nielsen in the context of making asymptotic expansions. A kind of 'maximal partition' into 'minimal terms' would be nice. Similarly, a systematic fashion for doing geometrically invariant Taylor expansions, was desirable. Barndorff-Nielsen finally asked whether anyone knew about computer programs for manipulating the algebraic aspects of geometric calculations. Dodson referred to a program called STENSOR made by R. d'Inverno at Southampton University that did calculate curvature tensors etc.

It was noted that the concept of a yoke, given in Blaesild's paper, essentially was the same as the maximising functional mentioned in Lauritzen's paper, and there was a remarkable similarity between the two "copies" of Ω and two embeddings. Foster mentioned that there was some similarity between the notion of a 'yoke' and that of a Hamiltonian.

REFERENCES

1. R.Ahlswede and I.Csiszar. **Hypothesis testing with communication constraint.** IEEE Trans. Inf. Theory, vol.IT-32, pp533-542 1986

2. S-I.Amari and T.S.Han. **Statistical inference under multiterminal rate restrictions - a differential geometrical approach.** METR 87-2, Univ. Tokyo 1987

3. L.Le Cam. **Asymptotic Methods in Statistical Decision Theory.** Springer-Verlag, New York 1986

4. B.Y.Chen. **Total mean curvature and submanifolds of finite type.** World Scientific Press, Singapore 1984

5. C.J.S.Clarke. **On the global isometric embedding of pseudo-Riemannian manifolds.** Proc. Roy. Soc. A314 (1970) 417-428. (cf. also [9] below)

6. L.Del Riego and C.T.J.Dodson. **Sprays, universality and stability.** Math. Proc. Camb. Phil. Soc. 103 (1988) in press

7. C.T.J.Dodson and M.S.Radivoiovici. **Tangent and frame bundles of order two.** An. St. Univ. Iasi XVIII I, 1(1982) 63-71

8. C.T.J.Dodson, L.Vanhecke and M.E.Vazquez-Abal. **Harmonic geodesic symmetries.** Comp. Rend. Ac Sc. Canada IX 5, (1987) in press

9. P.A.Griffiths and G.R.Jensen. **Differential Systems and Isometric Embeddings.** Princeton University Press Princeton 1987. (cf. also : R.Greene. Isometric embeddings of Riemannian and pseudo-Riemannian manifolds. Memoirs Amer. Math. Soc 97 (1970),and : C-S Lin. The local isometric embedding problem in $\mathbb{R}^3$ of two-dimensional Riemannian manifolds with Gaussian curvature changing sign nicely. Thesis, NYU, New York 1983)

10. M.Goldstein. **Revising previsions: a geometric interpretation.** JRSS B43 (1981) 105-130

11. N.Hicks. **Notes on Differential Geometry.** Van Nostrand, Princeton 1965

12. N.Karmarkar. **A new polynomial time algorithm for linear programming.** Combinatorica 4 (1984) 373-395. (cf. Math. Intell. 9,2 (1987) 4-10).

13. S.Kobayashi and K.Nomizu. **Foundations of Differential Geometry.** Volume 1, Interscience, New York 1963

14. L.Mangiarotti and M.Modugno. **Fibred spaces, jet spaces and connections for field theories.** in Proc. Int. Meeting Geometry and Physics, Florence 12-15 October 1982. Pitagora Editrice, Bologna 1983

15. P.McCullagh. **Tensor Methods in Statistics.** Chapman and Hall, London 1987

16. A.F.S.Mitchell. **Statistical manifolds of univariate elliptical distributions:** International Statistical Review, 56 (1988) in press

17. M.K.Murray. **Coordinate systems and Taylor series in statistics.** (1987) Research Report: School of Math. Sci., Flinders Univ. of South Australia.

18. M.K.Murray and J.W.Rice. **On differential geometry in statistics**(1987) Research Report: School of Math Sci., Flinders Univ. of South Australia

19. M.M.Postnikov. **The Variational Theory of Geodesics.** Saunders, London 1967

20. L.C.G.Rogers and D.Williams. **Diffusions, Markov Processes and Martingales.** Vol.2 Itô calculus. Wiley, London 1987

21. R.M.Royall. **Model robust confidence intervals using maximum likelihood estimators.** International Statistical Review 54, (1986) 221-226

22. J.M.Selig. **Topology of configuration spaces of three identical particles.** Ph.D. Thesis, Dept. App. Math. Theor. Physics, Liverpool University 1984 (cf. also F.J.Bloore, I.Bratley and J.M.Selig SU(n)-bundles over the configuration space of three identical particles moving in $\mathbb{R}^3$. J.Phys. A 16 (1984) 729-736)

23. R.L.Smith and Naylor. **A comparison of maximum likelihood and Bayesian estimators for the 3-parameter Weibull distribution.** Applied Statistics (1987) 36, 358-369

24. L.Vanhecke and M.E.Vazquez-Abal. **Harmonic reflections** Preprint, Katholieke Univ. Leuven 1987

25. H.Weyl. **Space Time Matter.** Dover, New York 1922

26. P.M.Williams. **Completeness and its stability on manifolds with connection.** Ph.D.Thesis Dept. Math Univ. Lancaster 1984

27. J.A.Wolf. **Spaces of constant curvature.** Publish or Perish Inc. 5th edition Wilmington Delaware 1984

28. Z.Zhang and T.Berger. **Estimation via encoded information.** IEEE Trans. on Inf. Theory to appear