Analytical and Geometrical Properties of Statistical Connections in Information Geometry

Paolo Gibilisco*
Dipartimento di Matematica
Politecnico di Torino
Corso Duca degli Abruzzi, 29
10129 Torino, Italy
gibilisco@polito.it

Giovanni Pistone
Dipartimento di Matematica
Politecnico di Torino
Corso Duca degli Abruzzi, 29
10129 Torino, Italy
pistone@calvino.polito.it

Keywords: Information Geometry, Exponential Manifold, Statistical Connections.

Abstract

Information Geometry is a field where one can measure the deep impact of geometry and analysis in statistics, information theory and related applied fields. The present contribution has the goal of showing also the impact that statistics and information theory can have in geometry and analysis. Indeed it is clear that the development of the non-parametric and non-commutative versions of Information Geometry need a massive use of mathematical instruments of infinite-dimensional analysis, geometry and operator theory. On the other side there is an increasing interest of mathematicians for the non trivial problems that are suggested by the applications of Information Geometry. Even in the elementary case of a finite state space, the geometrical approach adds considerable insight to the modeling of applied problems.

1 Introduction

Till recently the field of Information Geometry was severely restricted by three kinds of limitations: restriction to parametric case, restriction to dominated case, restriction to commutative (non quantum) case. We argue that the cause for this underdevelopment was the lack of interest of mathematicians for the difficult problems that statisticians (and applied mathematicians in general) had to face to remove the above quoted limitations. In the last ten years the situation is greatly changed. A number of researchers has started to apply a more and more sophisticated mathematics in the fields and a number of results and of new projects (of intrinsic mathematical interest) has been produced. In this note we would like to survey some of the results and to describe possible links between our works and the works of other researchers in the area.

*Second address: Centro Vito Volterra, Università di Roma Tor Vergata, Italy

2 Non parametric theory

2.1 The Exponential Statistical Manifold

Our non parametric approach is based on the result of [16]: on the set of non-negative probability densities of a given sample space one can give a manifold structure using the exponential Orlicz spaces (Zygmund spaces). This Banach manifold is called the *Exponential Statistical Manifold* (ESM). A subsequent paper [15] considers further developments, in particular connections with information theory concepts. On this manifolds it is possible to define the Amari-Centsov α-connections as is shown in [6]. This non parametric construction gives new light to the classical dual structure of this connections: indeed the classical duality is simply the L^a and Orlicz space duality. This approach is further developed in the non-commutative setting in [5]. This results are the main motivations for our point of view about Information Geometry: one should study the geometry of certain submanifolds of suitable Banach spaces and subsequently one should deduce geometrical properties of statistical models by the various embeddings of the model in the submanifolds.

2.2 Generalized Amari embeddings

The classical Amari α–embeddings [1] are written as $p \mapsto ap^{\frac{1}{a}}$ for $a \in [1, \infty)$ and $p \mapsto \log p$ for $a = +\infty$ where $\alpha = 1 - 2/a \in [-1, 1]$. One may generalize this functions in three directions. First case: allow $a \in (0, 1)$ so that one has to use the geometry of the L^a—spaces. These are non-locally convex topological vector spaces. Also for these space there is a geometry that can be used [10] despite the fact that we do not have at our disposal the much more powerful instruments of Banach space theory. Second case: consider an invertible Young function Φ. In this case $p \to \Phi^{-1}(p)$ take the densities into the unit

sphere of the Orlicz space L^Φ (see [6]). Third case: one may avoid positivity. Indeed if $f \in L^a$ define a map $\varphi_a^b : L^a \to L^b$ by

$$\varphi_a^b(f) = \|f\|_a^{1-\frac{a}{b}} \mathrm{sgn}(f)|f|^{a/b}$$

if $f \neq 0$ and $\varphi_a^b(0) = 0$ (where $a, b \in [1, \infty)$). This map is an isometry, coincide with the duality mapping if $b = \tilde{a}$ (where $1/a + 1/\tilde{a} = 1$) and $\varphi_a^c = \varphi_b^c \circ \varphi_a^b$. On the positive densities in L^1 this function reduces to the classical Amari embedding.

2.3 Family of dual connections, interpolation and the definition of statistical manifolds

As we have said before it has been shown that the α-connection can be defined by the pull-back of the natural connection on the unit sphere $S^a \subset L^a$ also in the non-parametric case (using the Amari embedding). This shows that the α–connections are the most important and natural examples in a much larger family of Orlicz connections. Indeed the non parametric construction is based on the fact that the L^a spaces are doubly uniformly convex (that is uniformly convex together with the dual). A Banach space X that is doubly uniformly convex is "almost" a Hilbert space. This means that there is a natural shortest projection given by the formula

$$\pi_{\ker(\tilde{x})} v = v - \langle \tilde{x}, v \rangle \frac{x}{||x||^2}$$

where $\tilde{x}$ is the duality mapping. Moreover the unit sphere S^X is a submanifold that inherits a natural connection from the trivial connection on X (see [6], [5]). Therefore, using the pull-back of the embedding $p \to \Phi^{-1}(p)$, we may speak of Φ–connections on a manifold of densities for any Φ such that L^Φ is a doubly uniformly convex Orlicz space.

Now consider, for example the unit sphere S^2 which is a Hilbert manifold. The pull-back of the previously defined maps φ_a^b gives a family of dual connections on this riemannian manifolds or, more precisely, a family of dual bundle-connection pairs. Let us call ∇^a (where $a \in (1, +\infty)$) this family of connections. Let $1/r = \lambda/a + (1-\lambda)/c$: in a suitable sense one may prove the following interpolation formula

$$\nabla^r = \lambda \nabla^a + (1 - \lambda)\nabla^c$$

Now let us take the limits $a \to 1, c \to +\infty$. We get

$$\nabla^{1/\lambda} = \lambda \nabla^1 + (1 - \lambda)\nabla^\infty$$

Let $(2/1 - \alpha) = a = 1/\lambda$ and substitute (formally) the mixture and exponential connection for the non-existing ∇^1 and ∇^∞ connections. We get

$$\nabla^\alpha = \frac{1-\alpha}{2}\nabla^m + \frac{1+\alpha}{2}\nabla^e$$

that is the classical relation between α, mixture and exponential connections: the Amari relation can be seen as a limiting case of the above interpolation formula. At the present moment this is a bit formal but there is big evidence: indeed the Zygmund spaces $L^{x \log x}$ and $L^{\exp}$ appear as natural 'smoother' substitutes for the spaces L^1 and L^∞. This is classical in interpolation theory and appears in the definition of non parametric exponential and mixture connections for non parametric statistical manifolds. So we suggest that somewhere it has to be found a link between the non parametric theory of statistical manifolds and the theory of interpolation of operators. This remark has been done in a complete independent way also by Zhu, see [20].

But even more important is the following: as we have described, the existence of family of dual connections it is not peculiar to manifolds of densities (see the S^2 example). Therefore some of the abstract notions of statistical manifolds that appear in the literature (Lauritzen, Kurose,...) maybe do not yet capture the essence of the concept of statistical manifold (see also [11]).

2.4 Geometry by embeddings

As we have seen, the needs of the non parametric theory shows that α–geometries are constructed by embeddings for $\alpha \in (-1, 1)$. But still the crucial construction of the ESM atlas is not made this way. Moreover the exponential and mixture connections do not appear as induced connections on submanifolds of an appropriate Banach space (that should be an Orlicz space). One way to solve this problem maybe is the following. Fix a reference density p. Consider the submanifolds $\mathcal{E}_p = \{f : E_p(e^f) = 1\} \subset L^{\exp}(p)$. Let q exponentially connected to p (see [16] [6]). We may embed q in $\mathcal{E}_p$ by $q \to \ln \frac{q}{p}$. We conjecture that the structure of the ESM can be derived in this way.

2.5 Dual manifolds and orthogonality

Another point is that of establishing a general theory of duality between manifolds in order to give rigorous meaning to some of the procedures of information theory, especially those concerning the use of α-geodesics. Indeed sometimes one want to conduct, for example, the mixture geodesics from a point (density) to a statistical model (manifold) and to discuss orthogonality. If one consider the ESM [16] it is difficult to see how this can be done 'inside' the manifold. So we suggest the following scheme. Consider two manifolds and a diffeomorphism $j : \mathcal{M} \to \mathcal{N}$ such that we have a duality pairing $\langle \cdot, \cdot \rangle$ between the tangent spaces $T_p\mathcal{M}, T_{j(p)}\mathcal{N}$. If we identify p and $j(p)$ we may say that two curves $\gamma \subset \mathcal{M}, \delta \subset \mathcal{N}$ are 'orthogonal' (with respect to j) in p if $\langle v, w \rangle = 0$ where v is the tangent vector of γ in $p \in \mathcal{M}$ and w is the tangent vector of δ in $j(p) \in \mathcal{N}$. As an example consider a doubly uniformly convex Banach space X and its dual $\tilde{X}$. The unit spheres

$S^X, S^{\tilde{X}}$ can be identified using the duality mapping $x \to \tilde{x}$ and the above discussion apply: if x is on the curve $\gamma \subset S^X$ and $\tilde{x}$ is on the curve $\delta \subset S^{\tilde{X}}$ we may calculate the 'angle' in x of γ and δ also if the two curves are on different manifolds. A full theory of this kind seems missing in differential geometry.

3 Non commutative theory.

The construction of the non parametric α–connection has been generalized to the non commutative case in [5]. One can generalize all the construction to manifolds of density operator respect to a normal semi finite trace τ on a semi finite von Neumann algebra M. Indeed the non commutative $L^a(M,\tau)$ spaces are doubly uniformly convex and this is enough to construct an α–bundle-connection pair using the non commutative Amari embedding $\rho \to \rho^{\frac{1}{a}}$. This construction still has to be confronted with other proposals by Hasegawa, Nagaoka, Petz. Observe that our proposal is non parametric also in this case and therefore much more general then the matrix case (or even of the case of von Neumann-Schatten classes $\mathcal{L}^a$). The major technical difficulty in the non commutative case was the non commutative version of the isomorphism between $L_0^a(p) = \{u \in L^a(p) | E_p(u) = 0\}$ and $T_{p^{\frac{1}{a}}} S^a = \ker(p^{\frac{1}{a}}) = \{v \in L^a(\mu) | \int v p^{\frac{1}{a}} = 0\}$ given in the commutative case by $u \to u p^{\frac{1}{a}}$. In this framework is still out of reach a result of ESM type for manifolds of density operators. But we have at our disposal the theory of non commutative Orlicz spaces and so also in this case there is some hope.

4 Non dominated models.

The approach based on the notion of ESM suffers from the severe limitation imposed by the assumption on the equivalence of probability measures in the model. The same limitation applies to the theory of connections as presented in the previous sections. An attempt to overcome this limitation, currently restricted to α-connections, has been developed by Zhu in the framework of a generalized Bayes approach, see [20]. Zhu has suggested that the generalized Lebesgue L^a spaces (without reference measure) should be used to understand how generalize Information Geometry for non dominated models. As we have seen, evidence in favor to this program is available form our work.

5 The algebraic approach

In [18] an algebraic approach to the theory of cumulants is considered, showing how to define (finite dimensional) *algebraic* statistical manifolds. Namely the cumulant generating funtion K of many classical distributions is shown to satisfy a *polynomial* differential equation $F(K', K'') = 0$. This property is called *finite generation* and it is related with the possibility to compute the infinite sequence of cumulants κ_n, $n = 1, 2, \ldots$, by polynomial recurrent relations. In this case the exponential model associated with the basic distribution is an algebraic variety in the following sense: Consider the chart in which each density in the model is represented by its expectation parameter and the variance of the exponential sufficient statistics. Then the model is described by a polinomial equation, i.e. it is an algebraic variety. This remark is computationally relevant because of the current availability of computer algebra tools for dealing with systems of polinomial equations both ordinary and differential.

6 New directions

6.1 Gromov distance and Gromov convergence.

We discuss now at the conceptual level the possibility of an asymptotic theory of statistical models similar to the Le Cam theory [12, 13] but based on the notion of Gromov convergence for Riemannian manifolds [8]. What is crucial for an asymptotic theory is the definition of a distance between statistical models. We suggest here that Information Geometry may give us a candidate for the distance. Indeed if one look at a statistical model as a riemannian manifold (via the Fisher information matrix) it is possible to define the Gromov–Hausdorff distance for two statistical models.

6.2 Motion by mean curvature

This is a completely wild speculation. After Efron's paper of 1975 we know that curvature may have a statistical meaning. In relatively recent times there has been a lot of activity in the so called motion by mean curvature theory, see [9] [7] (with important applications of the theory of viscosity solutions). Can we give a probabilistic or statistical meaning to dynamics of manifolds driven by mean curvature evolution (especially to the limit theorems)?

7 Applications

There are several aspects of these theories that are relevant for applications in estimation and filtering, namely: finite state space, statistical distances, models approximation, information theory, diffusion processes and filtering equations. Concerning the last applications, we refer to [2] and to the papers given in this conference, and discuss briefly the finite space issue.

It has been shown in [4] and [17] that finite sample spaces are well described in an algebraic framework. In particular the description of the ring of random variables on a finite subset of $\mathbf{Q}^d$ can be given as follows: First find a system of polynomial equations whose solution set is the given sample space; then compute a special form

of such a system called *reduced Gröber basis*, see [3]; then derive by inspection a basis for the ring of random variables, together with the algebraic rules for computing. The algebraic approach to finite sample spaces can be extended to consider probability spaces, see [14], and statistical models, see [19]. In particular the ability to use computer algebra tools in the description of the geometry of the statistical model is quite interesting.

8 Conclusions.

In our century we have seen two major changes in probability. By one side we have seen a massive use of mathematics in probability. On the other side we have seen "probabilistic proofs" of theorems that are not probabilistic in content: the most striking example is still, maybe, the Bernstein proof of Weierstrass approximation theorem. A probabilistic proof often provides a shorter proof and moreover explains *why* a certain theorem holds. Also in statistics we have a massive use of mathematics. But we are still waiting for a "statistical" proof of a deep mathematical truth. To this respect we want to be optimistic: maybe the increasing link between Information Geometry and abstract mathematics shows that a "Bernstein proof" coming from statistics it is not too far ahead of us.

Detailed references and relevant preprints are available at `http://www.polito.it/~pistone`.

References

[1] S. Amari. *Differential-geometrical methods in statistics*. Number 28 in Lecture Notes in Statistics. Springer-Verlag, New York-Berlin, 2nd printing 1990 corrected edition, 1985.

[2] D. Brigo and G. Pistone. Projecting the Fokker-Planck equation onto a finite dimensional exponential family. Preprint 4, Dipartimento di Matematica Pura e Applicata dell'Università di Padova, March 1996.

[3] D. Cox, J. Little, and D. O'Shea. *Ideal, Varieties, and Algorithms*. Springer-Verlag, New York, 2nd edition, 1997. 1st ed. 1992.

[4] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, February 1998.

[5] P. Gibilisco and T. Isola. Connections on statistical manifolds of density operators by geometry of noncommutative L^p spaces. Rapporto Interno 21/98, Dipartimento di Matematica del Politecnico di Torino, 1998.

[6] P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 1(2), 1998. To appear.

[7] M. A. Grayson. The heat equation shrinks embedded curves to round points. *Journal of Differential Geometry*, 26:285–314, 1987.

[8] M. Gromov. *Structures métriques pour les variétes riemanniennes*. Textes Mathématiques. Cedic/Fernand Nathan, Paris, 1981. Rédigé par J. Lafontaine et P. Pansu.

[9] T. Ilmanen. *Elliptic Regularization and Partial Regularity for Motion by Mean Curvature*, volume 520 of *Memoirs*. American Mathematical Society, 1994.

[10] N. J. Kalton, N. T. Peck, and J. W. Roberts. *An F-space sampler*. Cambridge University Press, Cambridge, 1984.

[11] R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. John Wiley, New York, 1997.

[12] L. M. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag, New York—Berlin, 1986.

[13] L. M. Le Cam and G. L. Yang. *Asymptotics in Statistics. Some Basic Concepts*. Springer Series in Statistics. Springer-Verlag, Berlin–New York, 1990.

[14] G. Pistone, E. Riccomagno, and H. P. Wynn. Gröbner bases and factorization in discrete probability and Bayes. revised version of the paper presented at the AMS Montreal meeting 1997 by E. Riccomagno and H. Wynn. Submitted, 1998.

[15] G. Pistone and M.-P. Rogantin. The exponential statistical manifold: Mean parameters, orthogonality, and space transformation. *Bernoulli*, 1998. To appear.

[16] G. Pistone and C. Sempi. An infinite dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 33(5):1543–1561, October 1995.

[17] G. Pistone and H. P. Wynn. Generalised confounding with Gröbner bases. *Biometrika*, 83(1):653–666, Mar. 1996.

[18] G. Pistone and H. P. Wynn. Finitely generated cumulants. Technical Report 3/97, Dipartimento di Matematica del Politecnico di Torino, 1997. Revised 1998. Submitted.

[19] G. Pistone and H. P. Wynn. Moment's aliasing in finite sample spaces. Menton, July, 1998.

[20] H. Zhu. Generalised Lebesgue spaces and application to statistics. Submitted, June 4 1998.