Exponential Statistical Manifolds: Overview and Applications

Giovanni Pistone
Dipartimento di Matematica del Politecnico
Corso Duca degli Abruzzi, 24
10129 Torino, Italy

1. Introduction.—In the recent book by K. Murray and J. Rice (1993) a full account of the relations between general manifold theory and the special case arising in Statistics is considered. Such a special case, often called Information Geometry (IG), was previously presented in book form in (Čentsov 1972), (Amari 1985) and (Amari, Barndorff-Nielsen, Kass, Lauritzen and Rao 1987). Murray and Rice insist on a geometric and coordinate-free presentation at the conceptual level, switching to coordinates to deal with applications in actual parametric models. The need of a non-parametric presentation was already remarked by many authors, starting from P. Dawid (1977), S. Amari (1982). Also the need of a proper functional framework for the non-parametric case is frequently mentioned.

The formal construction of an atlas on the set of positive densities of a sample space was given in (Pistone and Sempi 1995). The development of the existing theory in this new functional framework has been carried over in (Pistone and Rogantin 1997) and (Gibilisco and Pistone 1997), where all basic results of IG are extended from parametric and semi-parametric settings to the non-parametric setting. Such results are summarized in the present paper, with the addition of a few unpublished results. The development of typically non-parametric applications is much less advanced, with the possible exception of applications to filtering, see (Brigo and Pistone 1996) and references therein.

During past seminar presentations of these results the question of the fruitfulness of such an approach has been questioned. In particular the following objection was frequently raised: the practical application of classical non-linear functional analysis—as used for example in the book by S. Lang (1995)—to function spaces of integrable functions is difficult, so that other approaches to infinite dimensional analysis should be preferred. This point is discussed in the last two Sections, together with the statement of our plans for future research work.

2. Functional Analysis.—The basic functional space on which our construction is based is the Orlicz space L^ϕ with exponential Young function, e.g. $\phi(x) = \cosh x - 1$. This class is naturally suggested by the theory of exponential models. In fact, if $p(t) = \exp(t \cdot u - \psi(t)) \cdot p$ is such a model, then the moment generating function of u is finite in a neighborhood of zero, so $u \in L^\phi(p)$. If we drop the natural parameters t, the 'parameter' is $u \in L^\phi(p)$ in the 'model' $p_u = \exp(u - K_p(u)) \cdot p$, where $K_p(u) = \log E_p(\exp u)$ is the 'cumulant functional'.

A basic property of the spaces $L^\phi(p)$ when p varies in the set $\mathcal{M}$ of positive probability densities of a given sample space $(X, \mathcal{X}, \mu)$ is the following: If p and q are two probability densities connected by an exponential model, i.e. such that they belong to and exponential model for parameters values

in the interior of the natural domain, then $L^\phi(p) = L^\phi(q)$. This property implies that such spaces can be used, at least locally, in place of the spaces $L^\infty(p)$ which have the property of being all equal for $p \in \mathcal{M}$. We can define the centered spaces $B_p = \left\{ u \in L^\phi(p) \mid E_p(u) = 0 \right\}$, and all B_q for q connected to p by an exponential model are closed subspaces of a fixed $L^\phi(p)$.

The non-parametric exponential density $\exp(u - K_p(u)) \cdot p$ defines a unique $u \in B_p$, defining a map e_p from a subset of $\mathcal{M}$ to B_p that will be the chart of the manifold. It follows from the general theory of Orlicz spaces that the space $L^\psi(p)$, $\psi(y) = (1+y)\log(1+y) - y$, is the pre-dual space of $L^\phi(p)$, and the corresponding centered space ${}^*B_p$ is the pre-dual of B_p.

The cumulant functional K_p is defined on an open domain of B_p and has a number of important properties: 1) Its proper domain contains $\mathcal{V}_p$, the open unit ball of B_p; 2) It is 0 at 0, otherwise is strictly positive; 3) It is convex and Fréchet C^∞ on $\mathcal{V}_p$; 4) $\forall u \in \mathcal{V}_p$, $q = e^{u - K_p(u)} \cdot p \in \mathcal{M}$; 5) The value of its n-th differential at u in the direction v is the n-th cumulant of v under q; 6) $\forall u \in \mathcal{V}_p$ and $q = e^{u-K_p(u)} \cdot p$, $\nabla K_p(u) = \frac{q}{p} - 1 \in {}^*B_p$ is its gradient and it is monotonic, in particular one-to-one; 7) The weak derivative of the gradient map ∇K_p at u applied to $w \in B_p$ is $\mathrm{D}(\nabla K_p(u))\, w = \left(\frac{q}{p}\right)(w - E_q(w))$ and it is one-to-one at each point.

3. Exponential Statistical Manifold (ESM).—Let us consider the following map: $e_p : \mathcal{V}_p \ni u \mapsto q = e^{u - K_p(u)} \cdot p \in \mathcal{M}$. This mapping is one–to–one because u is centered. We denote by $\mathcal{U}_p$ the image of $\mathcal{V}_p$ by the mapping e_p and by s_p the inverse of e_p on $\mathcal{U}_p$, $s_p : \mathcal{U}_p \ni q \mapsto \log\frac{q}{p} - E_p\left(\log\frac{q}{p}\right) \in \mathcal{V}_p$. The s_p maps are the centered log-likelihoods and they form an atlas on $\mathcal{M}$, defining the ESM.

As each $\mathcal{U}$ is modeled on B_p, then B_p in a model of the tangent space $T_p\mathcal{M}$. As at least locally all this spaces are subspaces of the same Orlicz, there is a natural trivialization of the tangent bundle $T\mathcal{M}$. An other interesting representation of the tangent space at p is the set of all one-dimensional exponential models trough p.

The manifold structure defined is this way presents a number of technical problems, depending on the unusual structure of the model space B_p. Namely B_p is not a reflexive space, and its structure is not compatible with the product of measure spaces. This gives rise to a non-trivial theory of sub-manifolds, which on the other side is essential, because we would like to present statistical models, both parametric and non parametric as submanifolds of the ESM. The other candidates to be charts of a manifold structure on $\mathcal{M}$, e.g. $q \mapsto \frac{q}{p} - 1 \in {}^*B_p$ (the 'expectation parameters') and $q \mapsto q^{1/a} \in L^a$ or $q \mapsto \left(\frac{q}{p}\right)^{1/a} \in L^a(p)$ (the 'Amari embeddings') cannot be charts of an atlas because of the positivity constrain. However they are smooth parameterizations on $\mathcal{M}$

4. Connections.—The introduction of typically statistical connections on the ESM has a key role in the geometrization of Statistics, and has an important role in applications to asymptotic theory, approximation, stochastic differential equations for filtering. The tangent bundle $T\mathcal{M}$ end its pre-dual ${}^*(T\mathcal{M})$ support respectively the exponential and mixture connections of via the flat parallel transports $(T\mathcal{M})_p \ni u \mapsto u - E_q(u) \in (T\mathcal{M})_q$ and ${}^*(T\mathcal{M})_p \ni u \mapsto \left(\frac{p}{q}\right) u \in {}^*(T\mathcal{M})_q$. The ESM $\mathcal{M}$ is naturally embeddable into the sphere of Lebesgue spaces $L^a(\mu)$ through the Amari embedding A^α in Eq. (1). A^α is a smooth map with differential at $p \in \mathcal{M}$ given by $d_p A^\alpha : (T\mathcal{M})_p = B_p \ni u \mapsto p^{1/a} v \in L^a$. If we consider the vector bundle $\mathcal{F}^\alpha$ with fiber $\mathcal{F}^\alpha_p = L^a_0(p)$ and the tangent bundle TS_a of the sphere S_a in $L^a(\mu)$, they are connected via the mapping I^α, see Eq. (1).

$$
\begin{array}{ccccccccc}
T\mathcal{M} & \longrightarrow & \mathcal{F}^\alpha & \xrightarrow{I^\alpha} & \mathcal{G} & \longrightarrow & TS_a & \longrightarrow & L^a(\mu) \\
\pi\downarrow & & \pi\downarrow & & \pi\downarrow & & \downarrow\pi & & \\
\mathcal{M} & = & \mathcal{M} & = & \mathcal{M} & \xrightarrow[A^\alpha]{} & S_a & &
\end{array}
\qquad \text{where:} \quad
\begin{cases}
\alpha & \in &]-1,1[\\
a & = & 2/(1-\alpha) \\
I^\alpha_p(u) & = & p^{1/a} \\
A^\alpha(p) & = & a p^{1/a}
\end{cases}
\tag{1}
$$

Because of the embedding of the tangent bundle TS_a into $L^a(\mu)$, and the existence of a natural splitting based on the projection along the vectors on the sphere, a natural connection is defined

on the tangent bundle TS_α. The covariant derivative of such a connection $\tilde{\nabla}$ is transferred to the covariant derivative on the bundle $\mathcal{F}^\alpha$, ∇^α, by the equation $\nabla^\alpha = I^{-1}\tilde{\nabla}(I \circ S)$: it is then derived from the pull-back of a connection. This construction is rigorous in our framework and it is equivalent to the classical definition. It has to be noted that the actual derivation of the explicit global equations given by Amari has to be performed in a different framework, for example by taking derivative of products in the sense of convergence in μ-measure. Everything we have discussed so far concerning α-connections could be generalized to bundles based on general Orlicz spaces.

5. Applications.—The main application of the non parametric theory is the treatment of semi-parametric inference due to Amari. In such an application the rigorous construction of a manifold and its connections should be of some help. In our opinion another important application of a non-parametric theory should be in the direction of the study of Fokker-Planck equations, with possibly an extension to filtering theory. In fact elliptic equations give rise (because of the maximum principle) to evolutions in an ESM. Geometric arguments based on the Amary embedding have been used to find finite dimensional approximations of the filtering equations.

6. Discussion.—The idea of taking as a model theory for IG the theory of manifolds modeled on a Banach space has been questioned. Especially our choice of the Orlicz exponential space puts us in a difficult framework, because of the peculiarity of such a Banach space. Other frameworks have been suggested, that we simply mention without giving the relevant references because of the lack of space: 1) One could consider only a weaker structure which is not globally a manifold, but induces on each "finite-dimensional sub-manifold" the right statistical structure; 2) Proper differentiability could be dropped, asking only differentiability in a dense set of preferred directions; 3) The geometry of spheres via embedding could be used, ignoring the fact that the image of the set of probability densities has empty interior. All these choices are actually related, but much work has to be done to fully clarify the matter.

7. New Research.—One of the main limitations of the geometric theory as discussed above is the restriction to a set of equivalent densities. Of course most basic asymptotic situations (for example the convergence of a binomial model to a Gaussian limit) do not fit at all into this framework. L. Le Cam has developed in the sixties an asymptotic theory to deal with this situation, see (Le Cam 1986). Le Cam theory is based on Hellinger distance and convergence in distribution of likelihood ratios. As each individual model in the sequence is actually a Riemannian manifold, we could think of a notion of convergence for sequences of manifold, the difficulty being the fact the manifolds are not sub-manifold of the same bigger structure. M. Gromov (1981) has developed tools for dealing with this problem. He gives a definition of distance between generic Riemannian manifolds and a theorem of relative compactness for sequences of manifolds based on the boundness from below of the Ricci curvature and other more technical conditions. Exploration of the connection between the two theories is just at the beginning.

The characterization problem of statistical manifolds in the class of generic manifolds was raised by some authors (Lauritzen 1987, Kurose 1990). What is needed in connection to the previous convergence problem is the possibility to characterize the statistical manifolds that arise as manifold limits of statistical manifolds. As a statistical manifold is a collection of densities, what is needed is some structure associated to the manifold and characteristic of the distribution. Natural candidates are: cumulant functionals and relative information, α-connections. We would like to mention also the characterization of distributions based on variance functions, see (Letac 1992), and a recent generalization in the direction of computer algebra in (Pistone and Wynn 1997).

BIBLIOGRAPHY

Amari, S. (1982), 'Differential geometry of curved exponential families. curvature and information loss', *The Annals of Statistics* **10**, 357–387.

Amari, S. (1985), *Differential-geometrical methods in statistics*, number 28 *in* 'Lecture Notes in Statistics', 2nd printing 1990 corrected edn, Springer-Verlag, New York-Berlin.

Amari, S., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L. and Rao, C. R. (1987), *Differential geometry in statistical inference*, number 10 *in* 'Institute of Mathematical Statistics Lecture Notes—Monograph Series', Institute of Mathematical Statistics, Hayward, CA.

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

Čentsov, N. N. (1972), *Statistical Decision Rules and Optimal Inference*, number 53 *in* 'Translations of Mathematical Monographs', American Mathematical Society, Providence, Rhode Island. Translation 1982.

Dawid, A. P. (1977), 'Further comments on a paper by Bradley Efron', *The Annals of Statistics* **5**, 1249.

Gibilisco, P. and Pistone, G. (1997), Connections on non-parametric statistical manifolds by Orlicz space geometry, Rapporto Interno 8/97, Politecnico di Torino, Dipartimento di Matematica. Submitted to *Infinite Dimensional Analysis, Quantum Probability and Related Topics.*

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

Kurose, T. (1990), 'Dual connections and affine geometry', *Matematishe Zeitschrift* **203**, 111–121.

Lang, S. (1995), *Differential and Riemannian Manifolds*, Springer-Verlag, New York.

Lauritzen, S. L. (1987), Statistical manifolds, in *Differential geometry in statistical inference* (Amari et al. 1987), pp. 163–216.

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

Letac, G. (1992), *Lectures on Natural Exponential Families and their Variance Functions*, number 50 *in* 'Monografias de Matemática', Istituto de Matemática Pura e Aplicada, Rio de Janeiro.

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

Pistone, G. and Rogantin, M.-P. (1997), The exponential statistical manifold: Mean parameters, orthogonality, and space transformation. 2nd revision. Submitted to *Bernoulli.*

Pistone, G. and Sempi, C. (1995), '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.

Pistone, G. and Wynn, H. P. (1997), Finitely generated cumulants, Technical Report 3/97, Dipartimento di Matematica del Politecnico di Torino. Submitted to *Statistics & Computing.*

SUMMARY

A fully non parametric theory of statistical manifolds can be developed from the idea of modeling the infinite dimensional manifold of positive probability densities of a given sample space over the Orlicz space of the exponential Young function. Functional analysis methods help in developing this construction: we show how to derive properties of the basic functionals, expectation parameters, orthogonality, connections by exploiting the duality between Orlicz spaces. Other approaches and possible direction for future research are briefly discussed.

RESUMÉ

La théorie non parametrique des variétés statistiques peut être développée dans le cadre des varietés de dimension infinie sur l'espace de Orlicz avec fontion de Young exponentielle. Des méthodes d'analyse fonctionelle permettent de realiser ce objectif: l'exposé momtrera la contruction des fonctionelles de base, des paramètres moyen, de l' orthogonalite et des connections, basée sur la dualité entre espaces de Orlicz. Différentes démarches et des possibilites des recherches ulteriéures sont presentés brievment.