Second Italian Meeting on Probability and Mathematical Statistics

Information geometry of the Gaussian Space

Giovanni Pistone

DE CASTRO STATISTICS

Collegio Carlo Alberto

Salerno, June 17, 2019

web-page: www.giannidiorestino.it
e-mail: giovanni.pistone@carloalberto.org

Abstract I

This talk is based on the conference papers [1,2,3]. It presents an overview of the topic and some of the current developments.

The exponential manifold [4,5] on the finite-dimensional Gaussian space [1] has special features namely, the existence of a finite entropy and finite moments of all orders for all densities in the manifold. Moreover, it is possible to discuss the continuity of translations, Poincaré inequalities, and the generalized differentiability for densities. As a consequence, it is possible to define an exponential manifld for densities belonging to a given Orlich-Sobolev space with Gaussian weight.

A field of application is the study of the dimensionality reduction for of evolution equations in the sense of D. Brigo [2] i.e., the projection of the solutions onto a finite-dimensional exponential family.

The basic exponential representation of densities in the exponential manifold can be modified by the use of the so-called deformed exponentials for example, the Nigel Newton exponential [6]. The linear growth of the deformed exponential allows for a simplified treatment of the manifold of densities in a Sobolev space with Gaussian weight.

Abstract II

1. G. Pistone. Information geometry of the Gaussian space. In *Information geometry and its applications*, volume 252 of *Springer Proc. Math. Stat.*, pages 119–155. Springer, Cham, 2018

2. D. Brigo and G. Pistone. Projection based dimensionality reduction for measure valued evolution equations in statistical manifolds. In F. Nielsen, F. Critchley, and C. Dodson, editors, *Computational Information Geometry. For Image and Signal Processing*, Signals and Communication Technology, pages 217–265. Springer, 2017

3. L. Montrucchio and G. Pistone. Deformed exponential bundle: the linear growth case. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, number 10589 in LNCS, pages 239–246. Springer, 2017. Third International Conference, GSI 2017, Paris, France, November 7-9, 2017, Proceedings

4. G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, 23(5):1543–1561, October 1995

5. M. Santacroce, P. Siri, and B. Trivellato. New results on mixture and exponential models by Orlicz spaces. *Bernoulli*, 22(3):1431–1447, 2016

6. N. J. Newton. A class of non-parametric statistical manifolds modeled on Sobolev spaces. arXiv:1808.06451v5

7. G. Pistone. Examples of the application of nonparametric information geometry to statistical physics. *Entropy*, 15(10):4042–4065, 2013

8. G. Pistone. Nonparametric information geometry. In F. Nielsen and F. Barbaresco, editors, *Geometric science of information*, volume 8085 of *Lecture Notes in Comput. Sci.*, pages 5–36. Springer, Heidelberg, 2013. First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings

9. B. Lods and G. Pistone. Information geometry formalism for the spatially homogeneous Boltzmann equation. *Entropy*, 17(6):4323–4363, 2015

Scoring rules I

This section is an based on

- A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005
- M. Parry, A. P. Dawid, and S. Lauritzen. Proper local scoring rules. *Ann. Statist.*, 40(1):561–592, 2012

It is intended to show an an example my support of my case about the use of Sobolev-Orlicz exponential manifold as the functional structure for Information Geometry. Please compare with the critical discussion in

- N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information Geometry*. Springer, 2017

Let $\mathcal{M}$ be a statistical model of positive densities on a real space measure space, $(\mathbb{R}^n, \mathcal{B}, \nu)$, ν an absolutely continuous measure.

A local scoring rule is scoring rule i.e., a mapping from densities to real random variables,

$$S\colon \mathcal{M} \ni q \mapsto S(q) \in \mathcal{L}(\mathbb{R}^n)$$
$$S(q)\colon x \mapsto S(x,q) \in \mathbb{R}\ ,$$

which moreover is local.

Scoring rules II

"Local" means that the value of the scoring rule at the sample point x depends only on $(Dq(x) \colon D \in \mathcal{D})$, $\mathcal{D}$ a list of derivation operators. In this sense we might say that a scoring rule S is local of order 2.

Examples of local scoring rules are

log-score $S \colon q \mapsto -\log q$. Comments: point-wise definition and want of robustness suggest the model $q \in C(\mathbb{R}^n)$; it is local of order 0.

Hyvärinen $S \colon q \mapsto \Delta \log q + \frac{1}{2} |\nabla \log q|^2$. Comments: Point-wise definition and want of robustness suggest $q \in C^2(\mathbb{R}^n)$; it is local of order 2.

A local scoring rule is not just a generic local operator.

The risk under a probability measure $\mu \in \mathcal{S} \supset \mathcal{M}$ is $d(\mu, q) = \mathbb{E}_\mu [S(q)]$. This is well defined if we assume $S(q) \in C(\mathbb{R}^n)$. $\mathcal{S}$ should include the model and the sample distributions. Moreover, we want $S(q) \in L^1(p)$ for all $p, q \in \mathcal{M}$.

Examples of risk are

Scoring rules III

log-likelihood $d(\mu, q) = -\mathbb{E}_\mu [\log q]$. It requires $\log q \in L^1(\mu)$, $q \in \mathcal{M}$, $\mu \in \mathcal{S}$.

Hyvärinen $d(\mu, q) = \mathbb{E}_\mu \left[\Delta \log q + \frac{1}{2} |\nabla \log q|^2 \right]$; it requires some Sobolev space assumption i.e., integrability of the derivatives up to the second order for all $\mu \in \mathcal{S}$.

The scoring rule is proper if $q \mapsto d(p, q)$ is minimized at $q = p$ only, that is, $d(p, q) \geq d(p, p)$ and $d(p, q) = d(p, p)$ implies $q = p$. In such a case, one defines the divergence associated to the proper and local scoring rule to be $\mathrm{D}(p \| q) = d(p, q) - d(p, p)$.

For example, from the log-score we obtain the KL divergence. $d(p, p) = -\mathbb{E}_p [\log p] = \mathcal{H}(p)$ is the entropy and

$$d(p, q) - d(p, p) = \mathbb{E}_p [-\log q + \log p] = \mathbb{E}_p \left[\log \frac{p}{q} \right] .$$

Hyvärinen divergence I

Let us assume now that the sample space is the n-dimensional real space and each density q in $\mathcal{M}$ is strictly positive C^2 and it is such that the partial derivatives $\partial_j \log q = \partial_j q/q$ are $L^2(p)$ all $p \in \mathcal{M}$.

The Hyvärinen divergence is

$$\mathrm{DH}\left(p|q\right) = \frac{1}{2}\int \left|\nabla \log p(x) - \nabla \log q(x)\right|^2 p(x)\ dx < \infty$$

By expanding the squared norm of the difference, we obtain

$$\frac{1}{2}\int \left|\nabla \log p(x)\right|^2 p(x)\ dx + \frac{1}{2}\int \left|\nabla \log q(x)\right|^2 p(x)\ dx - \int \nabla \log p(x) \cdot \nabla \log q(x)\ p(x)\ dx\ ,$$

where the first term does not depend on q.

Hyvärinen divergence II

If $\nabla \log p = \nabla p / p$ and the border terms in the integration by parts are zero

$$-\int \nabla \log p(x) \cdot \nabla \log q(x)\ p(x)\ dx =$$

$$-\int \nabla p(x) \cdot \nabla \log q(x)\ dx = \int \Delta \log q(x)\ p(x)\ dx$$

The Hyvärinen score is

$$S_{\mathrm{H}}(q) = \Delta \log q(x) + \frac{1}{2} \left| \nabla \log q(x) \right|^2 \ .$$

Minimization of the expected Hyvärinen score is the same as minimization of the Hyvärinen divergence.

All assumptions made are satisfied if $\mathcal{M}$ is the multivariate Gaussian model. This provides an example where a statistical method requires a detailed discussion of the properties of the spatial derivatives of the statistical model.

Variations on DH $(p|q)$

On the Gaussian space $(\mathbb{R}^n, \gamma)$ consider the densities of exponential form $p = \mathrm{e}^{u-K(u)} \cdot \gamma$. Then, at least formally,

$$\mathrm{DH}\,(p|q) = \frac{1}{2} \int |\nabla u - \nabla v|^2 \, \mathrm{e}^{u-K(u)} \gamma(x) \; dx$$

In this case, the ∇ operator could be taken in the sense of the analysis of the Gaussian space. Regularity of the operator should be discussed?

A variation where the integrability issue does not appear is based on the replacement the log function with the Nigel Newton balanced chart $\log_A(t) = \int_1^y ds/A(s)$, with $A(t) = s/(1+s)$. A possible new definition could be

$$\frac{1}{2} \int |\nabla \log_A p(x) - \nabla \log_A q(x)|^2 \, A(p(x)) \; dx$$

where the cancellation holds and $A \circ p$ is bounded.

- P. Malliavin. *Integration and probability*, volume 157 of *Graduate Texts in Mathematics*. Springer-Verlag, 1995. With the collaboration of Hlne Airault, Leslie Kay and Grard Letac, Edited and translated from the French by Kay, With a foreword by Mark Pinsky
- N. J. Newton. An infinite-dimensional statistical manifold modelled on Hilbert space. *J. Funct. Anal.*, 263(6):1661–1681, 2012

IG as the geometry of the statistical bundle

- In a typical set up, we have a set of positive densities $\mathcal{M}$ and a set of random variables B. We need the smoothness of a given map $\mathcal{M} \times B \ni (q, S) \mapsto F(q, S) \in \mathbb{R}$ i.e., $(q, S) \mapsto \mathbb{E}_q[S]$.
- A natural structure consists of endowing the model $\mathcal{M}$ with a differentiable atlas of charts and take as B a set of linear fibers on the manifold.
- The statistical bundle on $\mathcal{M}$ is

$$S\mathcal{M} = \{(p, u) \,|\, p \in \mathcal{M}, u \in B_p, \mathbb{E}_p[u] = 0\}$$

- Moreover, each fiber B_p is to be an expression in the atlas of the tangent space at p, $T_p\mathcal{M} \equiv B_p$. This last requirement is not trivial. For example, in general $L_0^2(p) \neq L_0^2(q)$.

- P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *IDAQP*, 1(2):325–347, 1998

Exponential bundle

$$\boxed{p \smile q} \implies \boxed{\begin{array}{ccccccc} \mathcal{E}(p) & \xrightarrow{s_p} & \mathcal{S}p & \xrightarrow{I} & B_p & \xrightarrow{I} & L^{(\cosh -1)}(p) \\ \| & & \downarrow{\scriptstyle s_q \circ s_p^{-1}} & & \downarrow{\scriptstyle d(s_q \circ s_p^{-1})} & & \| \\ \mathcal{E}(q) & \xrightarrow[s_q]{} & \mathcal{S}q & \xrightarrow[I]{} & B_q & \xrightarrow[I]{} & L^{(\cosh -1)}(q) \end{array}}$$

- If $p \smile q$, then $\mathcal{E}(p) = \mathcal{E}(q)$ and $L^{(\cosh -1)}(p) = L^{(\cosh -1)}(q)$.
- $B_p = \left\{ u \in L^{(\cosh -1)}(p) \,\middle|\, \mathbb{E}_p[u] = 0 \right\}$
- $\mathcal{S}p \neq \mathcal{S}q$ and $s_q \circ s_p^{-1} : \mathcal{S}p \to \mathcal{S}q$ is affine

$$s_q \circ s_p^{-1}(u) = u - \mathbb{E}_q[u] + \log\left(\frac{p}{q}\right) - \mathbb{E}_q\left[\log\left(\frac{p}{q}\right)\right]$$

- The tangent application is $d(s_q \circ s_p^{-1})(u)[v] = v - \mathbb{E}_{e_p(u)}[v] = {}^{\mathrm{e}}\mathbb{U}_p^{e_p(u)} v$ (does not depend on p).

Gaussian space

- The Gaussian maximal exponential manifold is $\mathcal{E}(\gamma)$ with

$$\gamma(x) = (2\pi)^{-n/2} \exp\left(-|x|^2/2\right), \qquad x \in \mathbb{R}^n$$

- The relevant Orlicz spaces are the exponential space $L^{(\cosh -1)}(\gamma)$ and the mixture space $L^{(\cosh -1)_*}(\gamma)$.
- The mixture space $L^{(\cosh -1)_*}(\gamma)$ is separable; its dual is the exponential space $L^{(\cosh -1)}(\gamma)$.
- A positive density $f \in \mathcal{P}_>$ has finite entropy if, and only if, f belongs to the mixture space

$$-\int f(x) \log f(x) \gamma(x)\, dx \; < +\infty \quad \Leftrightarrow \quad f \in L^{(\cosh -1)_*}(\gamma) \; .$$

- $L^\infty(\gamma) \hookrightarrow L^{(\cosh -1)}(\gamma) \hookrightarrow L^a(\gamma) \hookrightarrow L^{(\cosh -1)_*}(\gamma) \hookrightarrow L^1(\gamma), \; a > 1$
- Restriction to the ball Ω_R: $L^{(\cosh -1)}(\gamma) \to L^a(\Omega_R)$, $a \geq 1$, and $L^{(\cosh -1)_*}(\gamma) \to L^1(\Omega_R)$.

- G. Pistone. Information geometry of the Gaussian space. arXiv:1803.08135, 2018

Notable elements in $L^{(\cosh -1)}(\gamma)$ and $L^{(\cosh -1)_*}(\gamma)$ I

- The exponential space $L^{(\cosh -1)}(\gamma)$ contains all polynomials with degree up to 2 and all functions which are bounded by such a polynomial.
- The mixture space $L^{(\cosh -1)_*}(\gamma)$ contains all $f : \mathbb{R}^d \to \mathbb{R}$ which are bounded by a polynomial, in particular, all polynomials.
- Poincaré inequality If $u \in \operatorname{dom}(\nabla)$ in the sense of the Gaussian space i.e., $u, \partial_j u \in L^2(\gamma)$ then

$$\int \left| u(x) - \int u(y)\gamma(y)\, dy \right|^2 \gamma(x)\, dx \;\leq\; \int \|\nabla u(x)\|^2 \gamma(x)\, dx$$

- $f \in C_{\mathrm{p}}^1(\mathbb{R}^n)$

$$\left\| f - \int f(y)\gamma(y)\, dy \right\|_{L^{(\cosh -1)_*}(\gamma)} \leq \text{const} \, \| |\nabla f| \|_{L^{(\cosh -1)_*}(\gamma)}$$

In particular, if f is a density of the Gaussian space,

Notable elements in $L^{(\cosh -1)}(\gamma)$ and $L^{(\cosh -1)_*}(\gamma)$ II

- $f \in C_{\mathrm{p}}^1(\mathbb{R}^n)$

$$\|f - 1\|_{L^{(\cosh -1)_*}(\gamma)} \leq \text{const} \, \| |\nabla f| \|_{L^{(\cosh -1)_*}(\gamma)}$$

  This inequality is similar to an bound on the entropy.

- If $f, g \in C_{\mathrm{p}}^2(\mathbb{R}^n)$ and $|x \cdot y| \leq |x|_1 |y|_2$ then

$$|\mathrm{Cov}_\gamma(f, g)| \leq \left| \|\nabla f\|_{L^{(\cosh -1)_*}(\gamma)} \right|_1 \left| \|\nabla g\|_{(L^{(\cosh -1)_*}(\gamma))^*} \right|_2 .$$

Orlicz-Sobolev with Gaussian weight (GOS)

- The GOS spaces with weight M are the vector spaces

$$W^{1,(\cosh -1)}(\gamma) = \left\{ f \in L^{(\cosh -1)}(\gamma) \,\middle|\, \partial_j f \in L^{(\cosh -1)}(\gamma), j = 1, \ldots, n \right\}$$

$$W^{1,(\cosh -1)_*}(\gamma) = \left\{ f \in L^{(\cosh -1)_*}(\gamma) \,\middle|\, \partial_j f \in L^{(\cosh -1)_*}(\gamma), j = 1, \ldots, n \right\}$$

  where ∂_j is the derivative in the sense of distributions.

- Both are Banach spaces with the norm of the graph

$$\|f\|_{W^{1,(\cosh -1)}(\gamma)} = \|f\|_{L^{(\cosh -1)}(\gamma)} + \sum_{j=1}^{n} \|\partial_j f\|_{L^{(\cosh -1)}(\gamma)}$$

$$\|f\|_{W^{1,(\cosh -1)_*}(\gamma)} = \|f\|_{L^{(\cosh -1)}(\gamma)} + \sum_{j=1}^{n} \|\partial_j f\|_{L^{(\cosh -1)}(\gamma)}$$

- J. Musielak. *Orlicz spaces and modular spaces*, volume 1034 of *Lecture Notes in Mathematics*. Springer-Verlag, 1983
- B. Lods and G. Pistone. Information geometry formalism for the spatially homogeneous Boltzmann equation. *Entropy*, 17(6):4323–4363, 2015

Smoothness of GOS spaces I

- Every $u \in W^{1,(\cosh -1)}(\gamma)$ when restricted to an open ball of radius $R > 0$ belongs to the Sobolev space $W^{1,a}(\Omega_R)$ for all $a \geq 1$ i.e. $u_R \in \cap_{a\geq 1} W^{1,a}(\Omega_R)$.
- Every $f \in W^{1,(\cosh -1)_*}(\gamma)$ when restricted to an open ball of radius $R > 0$ belongs to the dual of the space $\cap_{a\geq 1} W^{1,a}(\Omega_R)$, in particular to $W^{1,1}(\Omega_R)$.
- Sobolev Each $u \in W^{1,(\cosh -1)}(\gamma)$ is a.s. continuous and Hölder of all orders on each $\overline{\Omega}_R$.
- If $u \in W^{1,(\cosh -1)}(\gamma)$, then $u, \partial_j u \in L^a(\gamma)$ for all $a \geq 1$ i.e.,

$$e^{-\frac{1}{2a}|x|^2} u, e^{-\frac{1}{2a}|x|^2} \partial_j u \in L^a(\mathbb{R}^n)$$

As

$$\partial_j e^{-\frac{1}{2a}|x|^2} u = -\frac{1}{a} x_j e^{-\frac{1}{2a}|x|^2} u + e^{-\frac{1}{2a}|x|^2} \partial_j u$$

it follows

$$\left(e^{-\frac{1}{2a}|x|^2} u\right) \in W^{1,a}(\mathbb{R}^n) \quad a \geq 1$$

Smoothness of GOS spaces II

- Morrey Because of

$$W^{1,(\cosh -1)}(\gamma) \ni u \mapsto \left(\mathrm{e}^{-\frac{1}{2a}|x|^2} u\right) \in W^{1,a}(\mathbb{R}^n) \quad a \geq 1$$

it holds for each $a > n$ the uniform bound

$$u \in W^{1,(\cosh -1)}(\gamma) \quad \Rightarrow$$
$$\mathrm{e}^{-\frac{1}{2a}|x|^2} |u(x)| \leq C(n,a) \left\| \mathrm{e}^{-\frac{1}{2a}|x|^2} u \right\|_{W^{1,a}(\mathbb{R}^n)} \quad \text{a.s.}$$

and the RHS is dominated by $\|u\|_{W^{1,(\cosh -1)}(\gamma)}$.

- The same assumption implies the global Hölder inequality

$$\mathrm{e}^{-\frac{1}{2a}|x|^2} u(x) - \mathrm{e}^{-\frac{1}{2a}|y|^2} u(y) \leq C(n.a) |x-y|^{1-n/a} \left\| \mathrm{e}^{\frac{1}{2a}|x|^2} u \right\|_{L^a(\mathbb{R}^n)} \leq$$
$$C(n.a) |x-y|^{1-n/a} \|u\|_{W^{1,(\cosh -1)}(\gamma)}$$

- R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003
- H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011

Exponential family modeled on $W^{1,(\cosh -1)}(\gamma)$

- If we restrict the exponential family $\mathcal{E}(\gamma)$ to $W^{1,(\cosh -1)}(\gamma)$,

$$W_\gamma = W^{1,(\cosh -1)}(\gamma) \cap B_\gamma = \left\{ u \in W^{1,(\cosh -1)}(\gamma) \,\middle|\, \mathbb{E}_\gamma[u] = 0 \right\}$$

  we obtain the non-parametric exponential family

$$\mathcal{E}_1(\gamma) = \left\{ \mathrm{e}^{u - K(u)} \cdot \gamma \,\middle|\, u \in W^{1,(\cosh -1)}(\gamma) \cap \mathcal{S}_\gamma \right\}$$

- Because of $W^{1,(\cosh -1)}(\gamma) \hookrightarrow L^{(\cosh -1)}(\gamma)$ the set $W^{1,(\cosh -1)}(\gamma) \cap \mathcal{S}\gamma$ is open in W_γ and the cumulant functional $K : W^{1,(\cosh -1)}(\gamma) \cap \mathcal{S}\gamma \to \mathbb{R}$ is convex and differentiable.
- Many features of the exponential manifold carry over to this case. In particular, we can define for each $f \in \mathcal{E}_1(\gamma)$ the space

$$W_f = W^{1,(\cosh -1)}(\gamma) \cap B_\gamma = \left\{ u \in W^{1,(\cosh -1)}(\gamma) \,\middle|\, \mathbb{E}_f[u] = 0 \right\}$$

  to be models for the tangent spaces of $\mathcal{E}_1(\gamma)$. The e-transport acts on these spaces, ${}^{\mathrm{e}}\mathbb{U}_f^g : W_f \ni u \mapsto u - \mathbb{E}_g[u] \in W_g$, so that we can define the statistical bundle to be

$$S\mathcal{E}_1(\gamma) = \{(g, v) \,|\, g \in \mathcal{E}_1(\gamma), v \in W_f\}$$

  and take as charts the restrictions of the charts defined on $S\mathcal{E}(\gamma)$.