SIGMAPHI2008

Simposium FISHER INFORMATION AND GEOMETRY

Maximal Exponential Models on Gaussian Spaces

Giovanni Pistone

Politecnico di Torino

Thursday 17th July, 2008

Sets of densities

Definition

(Ω, μ) is a generic probability space, $\mathcal{M}^1$ is the set of real random variables f such that $\int f \, d\mu = 1$, $\mathcal{M}_{\geq}$ the convex set of probability densities, $\mathcal{M}_{>}$ the convex set of strictly positive probability densities:

$$\mathcal{M}_{>} \subset \mathcal{M}_{\geq} \subset \mathcal{M}^1$$

- We define the (differential) geometry of these spaces in a way which is meant to be a non-parametric generalization of the theory presented by Amari and Nagaoka (Jap. 1993, Eng. 2000).
- We try to avoid the use of explicit parametrisation of the statistical models and therefore we use a parameter free presentation of differential geometry.
- We construct a manifold modelled on an Orlicz space. In the N-state space case, it is a subspace of dimension $N - 1$ of the ordinary euclidean space

Vector bundles

The convex sets $\mathcal{M}^1$ and $\mathcal{M}_>$ are endowed with a structure of affine manifold as follows:

- At each $f \in \mathcal{M}^1$ we associate the linear fiber ${}^*T(f)$ which is a vector space of random variables whose expected value at p is zero. In general, it is an Orlicz space of $L \log L$-type; in the finite state space case, it is just the vector space of all random variables with zero expectation at p.
- At each $p \in \mathcal{M}_>$ we associate the fiber $T(f)$, which is an Orlicz space of exponential type; in the finite state space case, it is just the vector space of all random variables with zero expectation at p.
- $T(p)$ is the dual space of ${}^*T(p)$. The theory exploits the duality scheme:

$$T(p) = ({}^*T(p))^\star \subset L^2_0(p) \subset {}^*T(p)$$

Definition

For each $p \in \mathcal{M}_>$, consider the chart s_p defined on $\mathcal{M}_>$ by

$$q \mapsto s_p(q) = \log\left(\frac{q}{p}\right) + D(p\|q) = \log\left(\frac{q}{p}\right) - \mathrm{E}_p\left[\log\left(\frac{q}{p}\right)\right]$$

Theorem

The chart is defined for all $q = \mathrm{e}^{u-K_p(u)} \cdot p$ *such that* u *belongs to the interior* $\mathcal{S}_p$ *of the proper domain of* $K_p : u \mapsto \log\left(\mathrm{E}_p\left[\mathrm{e}^u\right]\right)$ *as a convex mapping from* $T(p)$ *to* $\mathbb{R}_{\geq 0} \cup \{+\infty\}$. *This domain is called maximal exponential model at* p, *and it is denoted by* $\mathcal{E}(p)$. *The atlas* $(s_p, \mathcal{S}_p)$, $p \in \mathcal{M}_>$ *defines a manifold on* $\mathcal{M}_>$, *called exponential manifold, briefly e-manifold. Its tangent bundle is* $T(p)$, $p \in \mathcal{M}_>$.

Remark One could replace exp, log with another couple of functions of interest, e.g. $\exp_\delta, \ln_\delta$. But see the following remark.

m-charts

Definition

For each $p \in \mathcal{M}_>$, consider a second type of chart on $\mathcal{M}^1$:

$$l_p : f \to l_p(f) = \frac{f}{p} - 1$$

Theorem

The chart is defined for all $f \in \mathcal{M}^1$ such that $f/p - 1$ belongs to ${}^*T(p)$. The atlas $(l_p, \mathcal{L}_p)$, $p \in \mathcal{M}_>$ defines a manifold on $\mathcal{M}^1$, called mixture manifold, briefly m-manifold. Its tangent bundle is ${}^*T(p)$, $p \in \mathcal{M}_>$.

Remark Other types of mappings are used in the literature to derive the Information Manifold. E.g. Amari uses $q \mapsto \sqrt{q} \in L^2(\mu)$. However, such a map does not define charts on $\mathcal{M}_>$, nor on $\mathcal{M}_\geq$. In fact, the set $L^2_\geq(\mu)$ has empty interior.

Connections

- At each point $p \in \mathcal{M}_>$ of the statistical manifold there is one reference system attached given by the e-chart and the m-chart at p.
- A change of reference system from p_1 to p_2 is just the change of reference measure.
- The change-of-reference formulæ are affine functions.
- The change-of-reference formulæ induce on the tangent spaces the **affine connections**:

$$\text{m-connection} \qquad {}^*T(p) \ni v \mapsto \frac{p}{q} v \in {}^*T(q)$$

$$\text{e-connection} \qquad T(p) \ni u \mapsto u - \mathrm{E}_q[u] \in T(q)$$

- The two connections are adjoint to each other.

Cumulant functional

Theorem

- *The divergence* $q \mapsto -D(p\|q)$ *is represented in the frame at* p *by* $K_p(u) = \log \mathrm{E}_p[\mathrm{e}^u]$, *where* $q = \mathrm{e}^{u - K_p(u)} \cdot p$.
- $K_p : T(p) \to \mathbb{R}_{\geq} \cup \{+\infty\}$ *is convex, infinitely Gâteaux-differentiable on the interior of the proper domain, analytic on the unit ball of* $T(p)$.
- *For all* v, v_1 *and* v_2 *in* $T(p)$ *the first two derivatives are:*

$$\mathrm{D}\, K_p(u)\, v = \mathrm{E}_q[v]$$

$$\mathrm{D}^2\, K_p(u)(v_1, v_2) = \mathrm{Cov}_q(v_1, v_2)$$

- *The divergence* $q \mapsto D(q\|p)$ *is represented in the frame at* p *by the convex conjugate* $H_p : {}^*T(p) \to \mathbb{R}$ *of* K_p.

Derivative

- Given a one dimensional statistical model $p_\theta \in \mathcal{M}_>$, $\theta \in I$, I open interval, $0 \in I$, the local representation in the e-manifold is u_θ with

$$p_\theta = \mathrm{e}^{u_\theta - K_p(u_\theta)} \cdot p.$$

- The local representation in the m-manifold is

$$l_\theta = \frac{p_\theta}{p} - 1$$

- To compute the velocity along a one-parameter statistical model in the s_p chart we use $\dot{u}_\theta$.
- To compute the velocity along a one-parameter statistical model in the l_p chart we use $\dot{p}_\theta / p$.

Relation between the two presentation

- We get in the first case

$$\dot{p}_\theta = p_\theta(\dot{u}_\theta - \mathrm{E}_\theta[\dot{u}_\theta])$$

so that

$$\frac{\dot{p}_\theta}{p_\theta} = \dot{u}_\theta - \mathrm{E}_\theta[\dot{u}_\theta] \quad \text{and} \quad \dot{u}_\theta = \frac{\dot{p}_\theta}{p_\theta} - \mathrm{E}_p\left[\frac{\dot{p}_\theta}{p_\theta}\right]$$

- In the second case we get

$$\dot{l}_\theta = \frac{\dot{p}_\theta}{p}$$

Example

For $p_\theta(x) = (2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(x-\theta)^2}$, in the coordinates at p_0, we have $p_\theta(x)/p_0(x) = \mathrm{e}^{\theta x - \frac{1}{2}\theta^2}$, therefore $u_\theta(x) = \theta x$, $\dot{u}_\theta(x) = x$, $\dot{p}_\theta(x)/p_0(x) = (x-\theta)\mathrm{e}^{\theta x - \frac{1}{2}\theta^2}$. Note: $\dot{p}_\theta(x)/p_\theta(x) = x - \theta$.

Moving frame

- Both in the e-manifold and the m-manifold there is one chart centered at each density. A chart of this special type will be called a *frame*. The two representations $\dot{u}_\theta$ and $\dot{l}_\theta$ are equal at $\theta = 0$ and are transported to the same random variable at θ:

$$\frac{\dot{p}_\theta}{p_\theta} = \dot{u}_\theta - \mathrm{E}_\theta\left[\dot{u}_\theta\right] = \dot{l}_\theta \frac{p}{p_\theta}.$$

Theorem

The random variable $\dot{p}_\theta / p_\theta$ is the Fisher score at θ of the one-parameter model p_θ. The Fisher information at θ is the L^2-norm of the score i.e. the velocity vector of the statistical model in the moving frame centered at θ. Moreover,

$$\mathrm{E}_\theta\left[\left(\frac{\dot{p}_\theta}{p_\theta}\right)^2\right] = \mathrm{E}_\theta\left[\left(\dot{u}_\theta - \mathrm{E}_\theta\left[\dot{u}_\theta\right]\right)\left(\dot{l}_\theta \frac{p}{p_\theta}\right)\right] = \mathrm{E}_p\left[\dot{u}_\theta \dot{l}_\theta\right].$$

Exponential models

- The Maximal Exponential Model $\mathcal{E}(p) = \left\{q = \mathrm{e}^{u-K_p(u)} \cdot p \colon u \in \mathcal{S}_p\right\}$ is the biggest possible statistical model in exponential form. Each smaller model has to be considered a sub-manifold of $\mathcal{E}(p)$.

Definition

Given a linear subspace V of $T(p)$, the exponential model on V is

$$\mathcal{E}_V(p) = \left\{q = \mathrm{e}^{u-K_p(u)} \cdot p \colon u \in V \cap \mathcal{S}_p\right\}$$

Example

When $V = \operatorname{Span}(u_i, \dots, u_n)$, the exponential model is

$$q(x;\theta) = \mathrm{e}^{\sum_{i=1}^n \theta_i u_i(x) - K_p(\sum_{i=1}^n \theta_i u_i)} p(x), \quad \sum_{i=1}^n \theta_i u_i \in \mathcal{S}_p$$

Exponential models in implicit form

- Let $V^\perp \subset {}^*T(p)$ be the orthogonal space of V. Then a positive density $q \in \mathcal{M}_>$ belongs to the exponential model on V if, and only if, $\mathrm{E}_p\left[\log\left(\frac{q}{p}\right) k\right] = 0$, for all $k \in V^\perp$.
- Assume $k \in V^\perp$ is of the form $k = l_p(r)$, i.e. $k = \frac{r}{p} - 1$. Then the orthogonality means $\mathrm{E}_r[u] = 0$ for $u \in V$ and implies

$$\mathrm{E}_p\left[\log\left(\frac{q}{p}\right)\left(\frac{r}{p} - 1\right)\right] = \mathrm{E}_r\left[\log\left(\frac{q}{p}\right)\right] + D(p\|q) = 0$$

  or

$$\mathrm{E}_r\left[\log\left(\frac{p}{q}\right)\right] = D(p\|q), \quad \mathrm{E}_r[u] = 0, u \in V$$

- In the finite state space case, with k integer-valued, the implicit form produces binomial invariants. (Toric case in Algebraic Statistics)

Optimization

- As an example, let us show how a classical optimization problem is spelled out within our formalism.
- Given a bounded real function F on Ω, we assume that it reaches its maximum on a measurable set $\Omega_{\max} \subset \Omega$. The mapping

$$\tilde{F} : \mathcal{M}_{>} \ni q \mapsto \mathrm{E}_q\left[F\right]$$

  is to be considered a regularization or relaxation of the original function F.
- If F is not constant, i.e. $\Omega \neq \Omega_{\max}$, we have $\tilde{F}(q) = \mathrm{E}_q\left[F\right] < \max F$, for all $q \in \mathcal{M}_{>}$. However, if ν is a probability measure such that $\nu(\Omega_{\max}) = 1$ we have $\mathrm{E}_\nu\left[F\right] = \max F$.
- This remark has suggested to look for $\max F$ by finding a suitable maximizing sequence $q_n \in \mathcal{M}_{>}$ for $\tilde{F}$.

Chart representation of the optimization problem

- The expectation of F is an affine function in the m-chart,

$$\widetilde{F}(q) = \mathrm{E}_p\left[F\left(\frac{q}{p} - 1\right)\right] + \mathrm{E}_p[F] = \mathrm{E}_p[F l_p(q)] + \mathrm{E}_p[F]$$

- Given any reference probability p, we can represent each positive density q in the maximal exponential model at p as $q = \mathrm{e}^{u - K_p(u)} \cdot p$. In the e-chart the expectation of F is a function of u, $\Phi(u) = \mathrm{E}_q[F]$.
- The equation for the derivative of the cumulant function K_p gives

$$\begin{aligned}\Phi(u) &= \mathrm{E}_q[F] \\ &= \mathrm{E}_q[(F - \mathrm{E}_p[F])] + \mathrm{E}_p[F] \\ &= \mathrm{D}\, K_p(u)\,(F - \mathrm{E}_p[F]) + \mathrm{E}_p[F]\end{aligned}$$

Steepest ascent

- The derivative of Φ in the direction v is the Hessian of K_p applied to $(F - \mathrm{E}_p[F]) \otimes v$ and from the formula of the Hessian follows

$$\mathrm{D}\,\Phi(u)\,v = \mathrm{Cov}_q(v, F).$$

Theorem

- *The direction of steepest ascent of the expectation* $\mathrm{E}_q[F]$ *at* q *is*

$$F - \mathrm{E}_q[F] \in T(q).$$

- *The one dimensional statistical model of steepest ascent is the exponential BG model*

$$p(\theta) = \mathrm{e}^{\theta F} / \Lambda(\theta)$$

Vector field

Definition

A **vector field** F of the the m-bundle ${}^*T(p)$, $p \in \mathcal{M}_>$, is a mapping which is defined on some domain $D \subset \mathcal{M}_>$ and it is a section of the m-bundle, that is $F(p) \in {}^*T(p)$, for all $p \in D \subset \mathcal{M}_>$.

Example

1. For a given $u \in T_p$ and all $q \in \mathcal{E}(p)$
$$F : q \mapsto u - \mathrm{E}_q[u]$$
2. For all strictly positive density $q \in \mathcal{M}_>(\mathbb{R}) \cap C^1(\mathbb{R})$
$$F : q \mapsto -q'/q$$
3. For all strictly positive $q \in \mathcal{M}_>(\mathbb{R}) \cap C^2(\mathbb{R})$
$$F : q \mapsto q''/q$$

Differential equations

Definition

A one-parameter statistical model in $\mathcal{M}_>$, $p(\theta)$, $\theta \in I$, solves the differential equation associated to the vector field F if $p(\theta) = \mathrm{e}^{u(\theta) - K_p(u(\theta))} \cdot p$ and

1. the curve $\theta \mapsto u(\theta) \in T(p)$ is continuous in L^2;
2. the curve $\theta \mapsto p(\theta)/p - 1 \in {}^*T(p)$ is continuously differentiable;
3. for all $\theta \in I$ it holds

$$\boxed{\frac{\dot{p}(\theta)}{p(\theta)} = F(p(\theta))}$$

Theorem

Assume F is locally maximal monotone. Then the equation $\dot{p}/p + F(p) = 0$ has a solution which is unique.

Example

1. The exponential model $p_\theta = \mathrm{e}^{\theta F}/\Lambda(\theta)$ is a solution of the equation $\frac{\dot{p}_\theta}{p_\theta} = F - \mathrm{E}_{p_\theta}[F]$.
2. The second example follows by considering $\Omega = \mathbb{R}$ and taking for domain the class of C^2 positive densities q such that $F(q) = -q'/q \in {}^*T(f)$. We can therefore consider the differential equation $\dot{p}_\theta/p_\theta = -F(p_\theta)$.
   Given any f in the domain, the statistical model $p_\theta(x) = f(x-\theta)$ is such that the score is

$$\frac{\dot{p}_\theta(x)}{p_\theta(x)} = -\frac{f'(x-\theta)}{p(f-\theta)} = F(p(\cdot - \theta))(x)$$

   and therefore is a solution of the differential equation. The classical Pearson classes of distributions are related to this equation.
3. It is the simplest case of the equations studied by F. Otto in

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001. ISSN 0360-5302. URL ../publications/Riemann.ps.

Malliavin Calculus aka Stochastic Analysis

- Let $\nu(dx) = (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}x^2} dx$. The adjoint of the derivative operator d with respect to the scalar product of $L^2(\nu)$ is

$$\begin{aligned}
\langle d\phi, \psi \rangle_\nu &= \int \phi'(x)\psi(x)\nu(dx) \\
&= \int \phi(x) \left(-\psi'(x) + x\psi(x) \right) \nu(dx) \\
&= \langle \phi, \delta\psi \rangle_\nu
\end{aligned}$$

- The operator $\delta\psi(x) = \psi'(x) + x\psi(x)$ is called **divergence**. In finite dimension i.e. for random variables defined on $\mathbb{R}^n$, the calculus of divergence is useful for the computation of densities of non-linear functions of Gaussian random variables.
- It has been discovered in the 80's that there exist an extension of δ to a class of stochastic processes whose value is the Wiener - Ito - Stratonovich - Skorohod - Nualart-Pardoux -$\cdots$ stochastic integral.

Abstract Wiener Space

Definition

Abstract Wiener Space $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $\mathcal{H}$ a Gaussian sub-space of $L^2(\Omega, \mathcal{F}, \mathbb{P}) = L^2$ such that $\sigma(\mathcal{H}) = \mathcal{F}$, H a separable Hilbert space, $\delta : H \to \mathcal{H}$ a mapping such that $\langle \delta(h_1), \delta(h_2) \rangle_{\mathcal{H}} = \langle h_1, h_2 \rangle_H$. The mapping δ is a linear and surjective isometry of H unto $\mathcal{H}$ called *divergence* or *abstract Wiener integral*.

- The exponential manifold does not use at all the structure of the underlying sample space. However, by using features of the underlying space, we give rise to a much richer theory.
- In the case of a finite state space consisting of a finite set of points of an affine space, random variables and density functions can be represente as polynomials and statistical models as algebraic varieties.
- Maximal exponential models with Gaussian reference measure have special algebraic and analytical features that can be discusses in the framework of Malliavin calculus.

Less abstract Wiener spaces

- In the two basic examples, H is the space of trajectories, see e.g. Nualart [2006].

Example

Let $X_1, X_2, \dots$ be a Gaussian White Noise (GWN) on the canonical space $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}, \nu^{\otimes \mathbb{N}})$, $\nu(dx) = (2\pi)^{-1/2} \exp\left(-x^2/2\right) dx$. The Hilbert space $H = \ell^2$ is the domain of a divergence as the mapping $\delta \colon a \mapsto \sum_{i=1}^{\infty} a(i) X_i$, $a \in H$ is a linear isometry between H and the closure $\mathcal{H}$ of $\operatorname{Span}\left(X_i : i = 1, 2, \dots\right)$.

Example

Let μ be the Wiener probability measure on the space of continuous trajectories $(C[0,1], \mathcal{B})$, W_t, $t \in [0,1]$, the canonical process. The divergence is defined on $H = L^2[0,1]$ by the Wiener integral $h \colon \int_0^1 h(s) dW_s$, because $\left\langle \int_0^1 h_1(s) dW_s, \int_0^1 h_2(s) dW_s \right\rangle_{\mathcal{H}} = \langle h_1, h_2 \rangle_H$.

Stocastic Analysis: derivative

Definition

The derivative operator ∇ is defined as a closed operator whose domain is the Gauss-Sobolev space $\mathbb{D}_1^2$. For $F \in \text{Poly}(\delta)$, $F = f(\delta(h_i)\colon i = 1, \ldots, n)$,

$$\nabla F = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\delta(h_i)\colon i = 1, \ldots, n) h_i$$

The ∇ of such an F is a polynomial stochastic process.

- The linear operator ∇ is a derivation of the $\mathbb{R}$-algebra $\text{Poly}(\delta)$:

$$\nabla(FG) = G\nabla F + F\nabla G$$

- Moreover, ∇ can be considered a gradient, because for $F = f(\delta(e_i)\colon i = 1, \ldots, n)$ and $h \in H$, we have

$$\left.\frac{d}{dt} f\left(\delta(e_i) + t \left\langle e_i, h\right\rangle_H\right)\right|_{t=0} = \left\langle \nabla F, h\right\rangle_H$$

Example

Let F be a monomial with respect to an orthonormal sequence $e_1, \ldots, e_n \in H$, $F = \delta(e_1)^{\alpha_1} \cdots \delta(e_n)^{\alpha_n}$. The set of such random variables is a linear basis of $\mathrm{Poly}(\delta)$. It follows that

$$\langle F, \delta(h) \rangle_{L^2} = \langle \nabla F, h \rangle_{L^2 \otimes H}$$

Therefore, the value at h of the adjoint of ∇ is $\nabla^*(h) = \delta(h)$.

Definition

The adjoint of ∇ is defined on $\mathrm{Poly}(\delta) \otimes_{\mathbb{R}} H$ and for $F = \delta(e_1)^{\beta_1} \cdots \delta(e_n)^{\beta_n}$ and $G = Fh$, one has

$$\nabla^* G = - \langle \nabla F, h \rangle_H + \delta(h) F$$

As ∇^* extends δ, it is denoted by $\delta = \nabla^*$ and it is called the *divergence*.

Exponential models in the AWS (In progress.)

- In the context of an abstract Wiener space $(\Omega, \mathcal{F}, \mathbb{P}, H, \delta)$, we want to discuss the densities in $\mathcal{E}(1)$, i.e. the densities of the form $F = \exp(U - K(U))$, $\operatorname{E}(U) = 0$.
- Because of the density in L^2 of the polynomial random variables, it has been suggested in various context the approximation of general exponential models with polynomial exponential models. Moreover, polynomial models could be of interest by themselves.
- We consider two cases: polynomial form for U or polynomial form for F.
- In the first case the main issue is the exponential integrability of U.
- In the second case the main issue is the strict positivity of the polynomial random variable.

Example

- A random variable u of the AWS belongs to the Orlicz space $T(1)$ (constant reference measure) if, and only if, $\mathrm{E}(u) = 0$ and the Laplace transform $\mathrm{E}(\mathrm{e}^{tu})$ is finite on an open interval containing 0.
- **Assume that the distribution of u has a density p_u wrt dx.** We can always write $p_u(x) = \tilde{p}_u(u)\left((2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}x^2}\right)$.

$$\begin{aligned}
\mathrm{E}\left(\mathrm{e}^{tu}\right) &= \int \mathrm{e}^{tx}\tilde{p}_u(u)\left((2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}x^2}\right) dx \\
&= \mathrm{e}^{\frac{1}{2}t^2}\int (2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(x-t)^2}\tilde{p}_u(x)dx \\
&= \mathrm{e}^{\frac{1}{2}t^2}\int (2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}x^2}\tilde{p}_u(x+t)dx
\end{aligned}$$

- The Malliavin calculus provides a number of conditions that imply the existence of a density p_u. E.g. $\frac{\nabla u}{\|\nabla u\|_H^2}$ is in the domain of the divergence.

Example

The exponential model whose canonical statistics are $\delta(e_1), \delta(e_2), \delta(e_1)\delta(e_2)$ has the form

$$F_{\theta_1,\theta_2,\theta_{12}} = \exp\left(\theta_1\delta(e_1) + \theta_2\delta(e_2) + \theta_{12}\delta(e_1)\delta(e_2) - \psi(\theta_1,\theta_2,\theta_{12})\right)$$

$$\psi(\theta_1,\theta_2,\theta_{12}) = \frac{1}{2}\frac{\theta_1^2+\theta_2^2+2\theta_1\theta_2\theta_{12}}{1-\theta_{12}^2} - \frac{1}{2}\log\left(1-\theta_{12}^2\right), \quad \theta_{12}^2 < 1$$

The expectation parameters are rational functions:

$$\eta_1 = \frac{\theta_1+\theta_2\theta_{12}}{1-\theta_{12}^2}, \quad \eta_2 = \frac{\theta_2+\theta_1\theta_{12}}{1-\theta_{12}^2}$$

$$\eta_{12} = \frac{\theta_1\theta_2(1+\theta_{12}^2)+\theta_{12}(1-\theta_{12}^2)}{\left(1-\theta_{12}^2\right)^2}$$

The orthogonal space of the model space is generated by all square-free monomials on the basis $e_1, e_2, \ldots$ other then those in the model.

Bibliography on MEM

The Information Geometry structure as it is defined in

- Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. American Mathematical Society, Providence, RI, 2000. ISBN 0-8218-0531-2. Translated from the 1993 Japanese original by Daishi Harada

has been extended to the non parametric case, see e.g

- Giovanni Pistone and Carlo 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. ISSN 0090-5364;
- Giovanni Pistone and Maria Piera Rogantin. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5(4):721–760, August 1999. ISSN 1350-7265;
- Paolo Gibilisco and Giovanni Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *IDAQP*, 1(2):325–347, 1998. ISSN 0219-0257;
- Alberto Cena. *Geometric structures on the non-parametric statistical manifold*. PhD thesis, Dottorato in Matematica, Università di Milano, 2002;
- Alberto Cena and Giovanni Pistone. Exponential statistical manifold. *AISM*, 59:27–56, 2007. ISSN 0020-3157. doi10.1007/s10463-006-0096-y. On line since December 16, 2006.