Matematikos ir informatikos institutas Vinius

Information Geometry and Algebraic Statistics

Giovanni Pistone and Eva Riccomagno

POLITO, UNIGE

Wednesday 30th July, 2008

Information Geometry

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.

Algebraic Statistics

- The finite state space case does not require any special functional framework but exibits interesting algebraic features, e.g. polynomial invariants.
  - Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 1998. ISSN 0090-5364;
  - Giovanni Pistone, Eva Riccomagno, and Henry P. Wynn. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman&Hall, 2001;
  - Lior Pachter and Bert Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005;
  - Dan Geiger, Christopher Meek, and Bernd Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34:1463–1492, 2006;
- Look for the state of the art in IG and AS in a forthcoming book edited by P. Gibilisco, E. Riccomagno, M.-P. Rogantin, H. Wynn, to be published by OUP.

An example: the Gibbs model

- Ω a finite sample space with N points. $E : \Omega \to \mathbb{R}_{\geq 0}$, such that $E(x) = 0$ for some $x \in \Omega$, but not everywhere zero.
- $$p(x; \beta) = \frac{\mathrm{e}^{-\beta E(x)}}{Z(\beta)}, \quad Z(\beta) = \sum_{x \in \Omega} \mathrm{e}^{-\beta E(x)}, \quad \beta > 0. \tag{1}$$
- In Statistical Physics, E is the *energy*, β is the *inverse temperature*, Z is the *partition function*, $\mathrm{e}^{\beta E}$ is the *Boltzmann factor*, $p(\beta)$, $\beta > 0$, is the*Gibbs model* or *canonical ensemble*.
- This model is not weakly closed: for $\beta \to \infty$, then $Z(\beta) \to \#\{x : E(x) = 0\}$ and $\mathrm{e}^{-\beta E(x)} \to (x : E(x) = 0)$. I.e. the weak limit of $p(\beta)$ as $\beta \to \infty$ is the uniform distribution on the states $x \in \Omega$ with zero energy. The limit distribution is not part of the Gibbs model, because it has a reduced support with respect to (1).

- Changing $E \to \max E - E$ and $\beta \to \theta = -\beta$ we get a parametrisation of a model extending the Gibbs model to negative β's:

$$p(x;\theta) = \frac{\mathrm{e}^{\theta(\max E - E(x))}}{\mathrm{e}^{-\theta \max E} Z(-\theta)} \tag{2}$$

  Such an extended model is convergent the the uniform distribution on the set $\{E(x) = \max E\}$ as $\theta \to \infty$.
- A more canonical presentation is the exponential model

$$p(x;\theta) = \mathrm{e}^{\theta u(x) - K(\theta u)} \cdot p(x;0) \tag{3}$$

  where $p_0 = p(\cdot;0)$ is the uniform distribution on Ω, the random variable u is centered for p_0 and $K(\theta u)$ is the cumulant generating function.
- We shall derive both a geometric and an algebraic description of the Gibbs model. The geometric picture is useful to further clarify the way in which the limits are obtained. The algebraic description will be given by equations that are satisfied by the Gibbs model, the extended parameter model, and also by its two limits.

Partition function and Entropy

- The partition function Z function it is convex, together with its logarithm $\log Z(\beta)$. Moreover,

$$\frac{d}{d\beta} \log Z(\beta) = -\mathrm{E}_\beta [E] , \tag{4}$$

$$\frac{d^2}{d\beta^2} \log Z(\beta) = \mathrm{Var}_\beta (E) . \tag{5}$$

- From

$$-\log p(x;\beta) = \beta E(x) + \log Z(\beta), \tag{6}$$

we derive another well known formula for the entropy:

$$S(\beta) = -\mathrm{E}_\beta [\log p(x;\beta)] = \beta \mathrm{E}_\beta [E] + \log Z(\beta), \tag{7}$$

- Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4.

- Equations (6) and (4) give

$$-\frac{d}{d\beta}\log p(x;\beta) = E(x) - \mathrm{E}_\beta[E] \tag{8}$$

  where the right-end-side is an *estimating function*.
- Equations (4), (5) and (7) give important variational results:

$$\frac{d}{d\beta}\mathrm{E}_\beta[E] = -\mathrm{Var}_\beta(E) \tag{9}$$

$$\frac{d}{d\beta}S(\beta) = -\beta\,\mathrm{Var}_\beta(E) \tag{10}$$

- The derivative of the continuous function $\beta \mapsto \mathrm{E}_\beta[E]$ is negative, therefore the expected value of the energy E decrease monotonically to its minimum value 0 for $\beta \to +\infty$.
- The function $\beta \mapsto S(\beta)$ is decreasing, and $\lim_{\beta\to\infty}\beta^{-1}S(\beta) = 0$.

Algebra

- Let $V^{\perp}$ be the orthogonal space of the space $V = \mathrm{Span}\,(1, E)$:

$$\sum_{x \in \Omega} k(x) = 0, \qquad \sum_{x \in \Omega} k(x)E(x) = 0. \tag{11}$$

- For each probability density $p = p(\beta)$ in the Gibbs model, that

$$\sum_{x \in \Omega} k(x) \log p(x) = 0, \qquad k \in V^{\perp} \tag{12}$$

- Vice versa, if a positive probability density function p satisfies Equation (12), then

$$\log p = \theta E + C \tag{13}$$

for suitable $\theta, C \in \mathbb{R}$. It follows that p belong to the larger model in equation (2).

Polynomial invariants

- For each k in the orthogonal space, we can take its positive part k^+ and its negative part k^-, so that $k = k^+ - k^-$ with $k^+ k^- = 0$. Equation (12) becames

$$\prod_{x \in \Omega} p(x)^{k^+(x)} = \prod_{x \in \Omega} p(x)^{k^-(x)} \tag{14}$$

- Equation (14) does not require the the strict positivity of each $p(x)$.
- When k has integer values, Equation (14) is a *polynomial invariant* for the probabilities in the Gibbs model. This algebraic invariant has the form of a binomial with unit coefficients.
- Equation (14) does not require the strict positivity of the density p and, in fact, the limit densitiy $p(\infty) = \lim_{\beta \to \infty} p(\beta)$ satisfies Equation (14) by continuity.

A numerical example

- When the energy function E takes its values on a regular grid, we can assume integer valued random variables $k_1, \ldots, k_{N-2}$ to be a basis of the orthogonal space of E and the constants.
- Consider e.g. a 5-points sample space $\Omega = \{1, 2, 3, 4, 5\}$ and the energy function $E(1) = E(2) = 0$, $E(3) = 1$, $E(4) = E(5) = 2$.
- Integer valued k_j, $j = 1, 2, 3$ are

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & E & k_1 & k_2 & k_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 & 1 \\ 1 & 1 & 0 & 0 & -4 \\ 1 & 2 & 0 & 1 & 1 \\ 1 & 2 & 0 & -1 & 1 \end{bmatrix} \end{array} \qquad \begin{array}{c} \begin{array}{cccccc} k_1^+ & k_1^- & k_2^+ & k_2^- & k_3^+ & k_3^- \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

- Equation (14) in this case is the system of polynomial (binomial) equations:

$$\begin{cases} p(1) = p(2) \\ p(4) = p(5) \\ p(1)p(2)p(4)p(5) = p(3)^4 \end{cases} \tag{15}$$

- The set of all polynomial invariants is an *ideal* of the polynomial ring $\mathbb{Q}[p(1), p(2), p(3), p(4), p(5)]$ and Equation (15) gives a *set of generators* of the ideal.
- If a non strictly positive density is a solution, then it is either $p(1) = p(2) = p(3) = 0$, $p(4) = p(5) = 1/2$, or $p(1) = p(2) = 1/2$, $p(3) = p(4) = p(5) = 0$. These two solutions are the uniform distributions of the sets of values that respectively maximize or minimize the energy function.

Toric models

- In the integer valued case, a further algebraic presentation is possible. In the equation $p(x:\beta) = \mathrm{e}^{-\beta E(x)}/Z(\beta)$ we could introduce the parameters $\zeta_0 = Z(\beta)^{-1}$ and $\zeta_1 = \mathrm{e}^{-\beta}$, so that $p(x;\zeta_0,\zeta_1) = \zeta_0\zeta_1^{E(x)}$.
- The probabilities are monomials in the parameters:

$$p(1) = p(2) = \zeta_0, \quad p(3) = \zeta_0\zeta_1, \quad p(4) = p(5) = \zeta_0\zeta_1^2 \tag{16}$$

- In algebraic terms, such a model is called a *toric model*. In (16) the parameter ζ_0 is required to be stricly positive, while the parameter ζ_1 could be zero. In such a case, Equation (16) gives the uniform distribution on $\{1,2\} = \{x\colon E(x) = 0\}$. The other limit solution is not obtained by (16).
- The *algebraic elimination* of the indeterminates ζ_0, ζ_1 in (16) will produce back polynomial invariants.

- David Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1997. ISBN 0-387-94680-2. An introduction to computational algebraic geometry and commutative algebra
- Martin Kreuzer and Lorenzo Robbiano. *Computational Commutative Algebra 1*. Springer, Berlin-Heidelberg, 2000.

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:

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

e-charts

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* $K_p(u) = \log \mathrm{E}_p\left[\mathrm{e}^u\right]$, *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,

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