London Mathematical Society Durham Symposium
Mathematical Aspects of Graphical Models
Monday 30th June - Thursday 10th July 2008

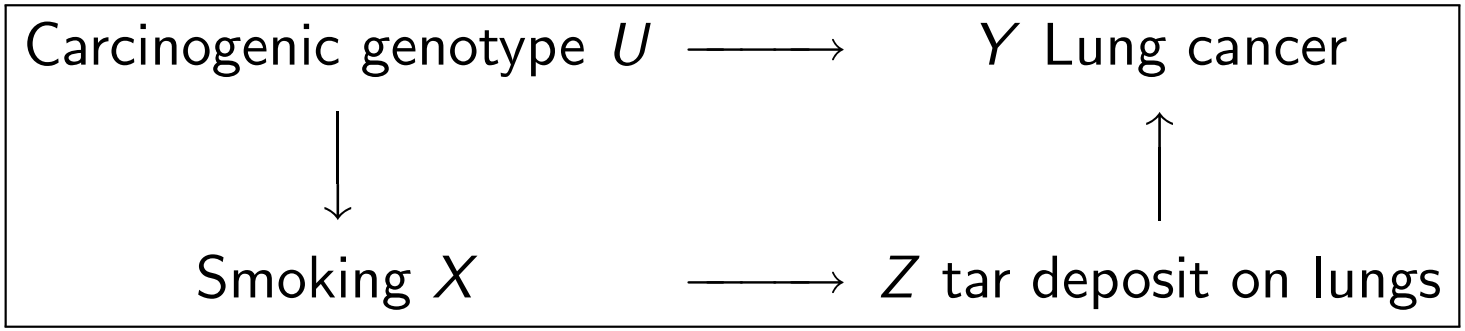
Information geometry of graphical models

Giovanni Pistone, Politecnico di Torino, Turin IT

Friday 4th July, 2008

An example of DAG

We consider the standard historical example:



There are 4 binary variables. We use the "harmonic" coding

$$\begin{array}{c} \text{TRUE} \mapsto 1 \mapsto (-1)^1 = -\mathbf{1} \\ \text{FALSE} \mapsto 0 \mapsto (-1)^0 = +\mathbf{1} \end{array}$$

We discuss the issue of parametrization from a polynomial algebra and geometrical perspective.

The previous DAG, together with the total *ordering*

$$U = X_1, X = X_2, Z = X_3, Y = X_4$$

encodes the factorization via the product formula

$$p(u, x, z, y) = p_1(u)p_{2|1}(x|u)p_{3|1}(z|x)p_{4|23}(y|u, z)$$

As the sample space is $\Omega = \{+1, -1\}^4$, a generic function is a square-free polynomial and densities have with respect to the uniform distribution have a special form:

$$\begin{aligned} p_1(x_1) &= 1 + a_1 x_1 \\ p_{2|1}(x_2|x_1) &= 1 + a_2 x_2 + a_{12} x_1 x_2 \\ p_{3|2}(x_3|x_2) &= 1 + a_3 x_3 + a_{23} x_2 x_3 \\ p_{4|13}(x_4|x_1, x_3) &= 1 + a_4 x_4 + a_{14} x_1 x_4 + a_{34} x_3 x_4 + a_{134} x_1 x_3 x_4 \end{aligned}$$

The product is computed with a symbolic computation software. It is a polynomial whose support is printed in boldface.

$$p(x_1, x_2, x_3, x_4; a_1, a_2, a_{12}, a_3, a_{23}, a_4, a_{14}, a_{34}, a_{124}) =$$

$$(a_1a_2a_3a_4 + a_1a_{12}a_3a_{14} + a_{12}a_3a_4 + a_1a_{23}a_4 + a_2a_3a_{14} + a_1a_2a_{34} + a_1a_{12}a_{134} + a_{12}a_{34} + a_2a_{134})\mathbf{x_1x_2x_3x_4} +$$

$$(a_1a_2a_3a_{34} + a_1a_{12}a_3a_{134} + a_1a_2a_4 + a_1a_{12}a_{14} + a_{12}a_3a_{34} + a_1a_{23}a_{34} + a_2a_3a_{134} + a_{12}a_4x_1 + a_2a_{14} +$$

$$a_{23}a_{134}x_1)\mathbf{x_1x_2x_4} + (a_1a_2a_{23}a_4 + a_1a_{12}a_{23}a_{14} + a_1a_3a_4 + a_{12}a_{23}a_4 + a_2a_{23}a_{14} + a_3a_{14} + a_1a_{34})\mathbf{x_1x_3x_4} +$$

$$(a_1a_{12}a_3a_4 + a_1a_2a_3a_{14} + a_2a_3a_4 + a_{12}a_3a_{14} + a_1a_{23}a_{14} + a_1a_{12}a_{34} + a_1a_2a_{134} + a_{23}a_{14})\mathbf{x_2x_3x_4} + (a_1a_2a_3 + a_{12}a_3 +$$

$$a_1a_{23})\mathbf{x_1x_2x_3} + (a_1a_2a_{23}a_{34} + a_1a_{12}a_{23}a_{134} + a_1a_3a_{34} + a_{12}a_{23}a_{34} + a_2a_{23}a_{134} + a_1a_4 + a_3a_{134} + a_{14})\mathbf{x_1x_4} +$$

$$(a_1a_{12}a_3a_{34} + a_1a_2a_3a_{134} + a_1a_{12}a_4 + a_1a_2a_{14} + a_2a_3a_{34} + a_{12}a_3a_{134} + a_1a_{23}a_{134} + a_2a_4 + a_{12}a_{14} + a_{23}a_{34})\mathbf{x_2x_4} +$$

$$(a_1a_{12}a_{23}a_4 + a_1a_2a_{23}a_{14} + a_2a_{23}a_4 + a_1a_3a_{14} + a_{12}a_{23}a_{14} + a_3a_4 + a_1a_{134} + a_{134}x_1 + a_{34})\mathbf{x_3x_4} + (a_1a_2)\mathbf{x_1x_2} +$$

$$(a_1a_2a_{23} + a_1a_{12}a_{23} + a_1a_3 + a_{12}a_{23})\mathbf{x_1x_3} + (a_1a_{12}a_3)\mathbf{x_2x_3} + a_1\mathbf{x_1} + (a_1a_{12} + a_{12}x_1 + a_2)\mathbf{x_2} + (a_2a_3x_2 + a_2a_{23} +$$

$$a_{23}x_2a_3)\mathbf{x_3} + (a_1a_{12}a_{23}a_{34} + a_1a_2a_{23}a_{134} + a_2a_{23}a_{34} + a_1a_3a_{134} + a_{12}a_{23}a_{134} + a_1a_{14} + a_3a_{34} + a_4)\mathbf{x_4} + \mathbf{1}$$

For each $\mathbf{x} = (x_1, x_2, x_3, x_4) \in \Omega = \{+1, -1\}^4$, the monomial map

$$\alpha \mapsto \mathbf{x}^\alpha = x_1^{\alpha(1)} x_2^{\alpha(2)} x_3^{\alpha(3)} x_4^{\alpha(4)}$$

is a representation of the additive group $\Omega = \mathbb{Z}_2^4$, so that each polynomial $f(x_1, x_2, x_3, x_4)$ is the discrete Fourier transform of its coefficients $A(\alpha)$, $\alpha : \{1, 2, 3, 4\} \to \{0, 1\}$, therefore

$$\begin{aligned} p_1(x_1) &= \hat{A}_1(\mathbf{x}) \\ p_{2|1}(x_2|x_1) &= \hat{A}_{2|1}(\mathbf{x}) \\ p_{3|2}(x_3|x_2) &= \hat{A}_{3|2}(\mathbf{x}) \\ p_{4|13}(x_4|x_1, x_3) &= A_{4|12}(\mathbf{x}) \end{aligned}$$

Therefore, the coefficients A of the joint density $p(\mathbf{x})$ in the previous disply are actually given by a convolution formula

$$\boxed{A = A_1 * A_{2|1} * A_{3|2} * A_{4|13}}$$

Manipulation, experiment, intervention?

Is that possible, under this model, to force X =FALSE, i.e. $x_2 = 1$? This is a sub-model, where $p(x_2|x_1) = 1 + x_2$, i.e. $a_2 = 1, a_{12} = 0$. Under this sub-model the cases where $x_2 = -1$ have probability 0, and the polynomial model is aliased with the model

$$p(x_1, x_3, x_4 \| x_2 = 1) =$$
$$(1+a_1x_1)(1+(a_3+a_{23})x_3)(1+a_4x_4+a_{14}x_1x_4+a_{34}x_3x_4+a_{134}x_1x_3x_4)$$

This corresponds to the DAG

$$U \longrightarrow Y \longleftarrow Z$$

and two effects are confounded by $\tilde{a}_3 = a_3 + a_{23}$

Marginalization

Given a density of the form

$$f(x_1, \ldots, x_n) = \sum_{\alpha} A(\alpha)\mathbf{x}^{\alpha}$$

and a subset of indexes $I \subset \{1, \ldots, n\}$ the marginal density is obtained by adding over all the sample values of x_i's such that $i \notin I$. This kills all the monomials that contain such x_i's. In other words, $f_I = \widehat{A\mathbf{1}_I}$. For example, in the model after the intervention, if $U = X_1$ is not observable,

$$\begin{aligned} p_{34}(x_3, x_4 \| x_2 = 1) &= (1 + \tilde{a}_3 x_3)(1 + a_4 x_4 + a_{34} x_3 x_4) \\ &= 1 + \tilde{a}_3 x_3 + (a_4 + \tilde{a}_3 a_{34}) x_4 + (a_{34} + \tilde{a}_3 a_4) x_3 x_4 \end{aligned}$$

which is generic.
Invariants of the model could be derived this way.

The transitions of the DAG can be written in exponential-polynomial form:

$$\begin{aligned}
\log\left(p_1(x_1)\right) &= b_1 x_1 - \psi(b_1) \\
\log\left(p_{2|1}(x_2|x_1)\right) &= (b_2 + b_{12}x_1)x_2 - \psi(b_2 + b_{12}x_1) \\
\log\left(p_{3|2}(x_3|x_2)\right) &= (b_3 + b_{23}x_2)x_3 - \psi(b_3 + b_{23}x_2) \\
\log\left(p_{4|13}(x_4|x_1,x_3)\right) &= (b_4 + b_{14}x_1 + b_{34}x_3 + b_{134}x_1x_3)x_4 \\
&\quad - \psi(b_4 + b_{14}x_1 + b_{34}x_3 + b_{134}x_1x_3)
\end{aligned}$$

where $\mathrm{e}^{\psi(b)} = \cosh(b)$,

$\mathrm{e}^{\psi(b_2+b_{12}x_1)} = \frac{\cosh(b_2+b_{12})+\cosh(b_2-b_{12})}{2} + \frac{\cosh(b_2+b_{12})-\cosh(b_2-b_{12})}{2}x_1$,

...

All the conditional cumulant functions are polynomials in the x's.

Exponential model

The logarithm of the joint distribution has the additive form

$$\begin{aligned}\log\left(p(x_1,x_2,x_3,x_4)\right) = &\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 \\ &+ \theta_{12} x_1 x_2 + \theta_{23} x_2 x_3 + \theta_{34} x_3 x_4 + \theta_{14} x_1 x_4 \\ &+ \theta_{134} x_1 x_3 x_4 \\ &- \psi(\theta_1,\theta_2,\theta_3,\theta_4,\theta_{12},\theta_{23},\theta_{14},\theta_{134})\end{aligned}$$

Again, the special form of the basis leads to simple computations related with the model. In fact, the orthogonal space is generated by the monomials which are missing in the model. They are:

$$x_1x_3,\ x_2x_4,\ x_1x_2x_3,\ x_1x_2x_4,\ x_2x_3x_4,\ x_1,\ x_2x_3x_4$$

which leads to an easy computation of the binomial invariants of the model.

- Polynomial algebra computations could be used in some cases to replace argument based on graph theory. Dedicated software helps in doing symbolic computations on polynomials when the field of constants is $\mathbb{Q}$, $\mathbb{Z}_p$ ...
- A careful choice of the levels coding makes the term of the polynomial be orthogonal.
- On a finite state space, both the mixture form and the exponential form can be presented in polynomial form.
- In the mixture case, a density is of the form $p = 1 + u$. In the exponential case, a density is of the form $p = \mathrm{e}^{u-\psi}$. In both cases u is a random variable on the sample space whose mean with respect to the reference probability is zero.
- Both mappings $p \mapsto u = p - 1$ and $p \mapsto \log(p) - \psi$ look like a chart, i.e. both map a set of densities to an open set of a vector space.

Definition

If $(\Omega, \mathcal{F}, \mu)$ 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 densities, $\mathcal{M}_{>}$ the convex set of strictly positive 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 parameterizations 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 od dimension $N - 1$ of the ordinary euclidean space

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 will be an Orlicz space; in the finite state space case, it is just the vecor space of all random variables with zero expectation at p.
- At each $p \in \mathcal{M}_>$ we associate the fiber $T(f)$. In the finite state space case the two types of fiber are equal as vector spaces.
- The theory will exploit the duality scheme: $T(p) \subset L^2_0(p) \subset {}^*T(p)$.

e-charts

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

$$q \mapsto s_p(q) = \log\left(\frac{q}{p}\right) + D(p\|q),$$

where

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

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 $T_0(p)$ to $\mathbb{R}_{>0} \cup \{+\infty\}$. 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 $T(p)$, $p \in \mathcal{M}_>$.

m-charts

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

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

The chart is defined for all $f \in \mathcal{M}^1$ such that $q/p - 1$ 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 ${}^*T(p)$, $p \in \mathcal{M}_>$.

Connections

At each point p in the statistical manifold there is one reference system attached given by the e-chart and the m-chart.

- A change of reference system from p_1 to p_2 is just the change of reference measure.
- The change of reference formulæare affine functions.
- The change of reference formulæinduce on the tangent spaces the **connections**

$$
\begin{aligned}
{}^{*}T(p) \ni v &\mapsto \frac{p}{q} v \in {}^{*}T(q) \\
T(p) \ni u &\mapsto u - \mathrm{E}_q[u] \in T(q)
\end{aligned}
$$

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 $\dot{u}_\theta$, while in the l_p chart we use $\dot{p}_\theta / p$. 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

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

The two cases are shown to represent the same geometric object by considering the the affine connections

$$T_p \ni u \mapsto u - \mathrm{E}_q[u] \in T_q \quad \text{and} \quad {}^*T_p \ni v \mapsto \frac{q}{p} v \in {}^*T_q$$

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:

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

This random variable is the *score* at θ of the one-parameter model. In other words, *the Fisher information at θ is the L^2-norm the velocity vector of the statistical model in the moving frame centered at θ*:

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

Let Ω be a finite sample space with N points and $E : \Omega \to \mathbb{R}_{\geq 0}$ a function, such that $E(x) = 0$ for some $x \in \Omega$, not everywhere zero. In Statistical Physics, E is called *energy* function. For each $\beta > 0$ consider the probability density function

$$p(x; \beta) = \frac{\mathrm{e}^{-\beta E(x)}}{\Lambda(\beta)}, \quad \text{where} \quad \Lambda(\beta) = \sum_{x \in \Omega} \mathrm{e}^{-\beta E(x)}.$$

The parameter β is called *inverse temperature*, the analytic function Λ *partition function*, and $p(\beta)$, $\beta > 0$, a *Boltzmann model* or *canonical ensemble*. This set of densities is not weakly closed. Indeed, if $\beta \to \infty$, then $\Lambda(\beta) \to \#\{x : E(x) = 0\}$ and $\mathrm{e}^{-\beta E} \to (x : E(x) = 0)$ point-wise, where for a set A, $\#(A)$ denotes its count and (A) its indicator function. The weak limit of $p(\beta)$ as $\beta \to \infty$ is the uniform distribution on the states $x \in \Omega$ with zero energy, namely on $\Omega_0 = \{E(x) = 0\}$. It is clear that this limit distribution is not part of the Boltzmann model.

Optimization

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 $\tilde{F} : \mathcal{M}_{\geq} \ni q \mapsto \mathrm{E}_q[F]$ 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 $\tilde{F}(q) = \mathrm{E}_q[F] < \max F$, for all $q \in \mathcal{M}_{>}$. However, if ν is a probability measure such that $\nu(\Omega_{\max}) = 1$ we have $\mathrm{E}_\nu[F] = \max F$. This remark has suggested to find $\max F$ by finding a suitable maximizing sequence q_n for $\tilde{F}$.

Given any reference probability p, we can represent each positive density q in the maximal exponential model at p as $q = \mathrm{e}^{u - K_p(u)} \cdot p$. The expectation of F is an affine function in the m-chart,

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

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

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

Stepest 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* $F - \mathrm{E}_q[F]$.

By the use of both the m- and e-geometry, we have obtained a quite precise description of the setting of this problem:

1. The problem is a convex problem in the m-geometry as the utility function $q \mapsto \mathrm{E}_q[F]$ is linear and the admissible set $\mathcal{M}^1$ is convex and closed in $L^1(\mu)$. The level sets are affine subspaces in the m-charts.
2. In the e-geometry, given any starting point $q \in \mathcal{M}_>$, the exponential model $\mathrm{e}^{\theta F}/\mathrm{E}_q\left[\mathrm{e}^{\theta F}\right]$ gives the steepest strict ascent. In fact, on such a statistical model the second derivative of the expected value of F is maximal at each point.
3. Let us assume that F is continuous. If the exponential model of steepest ascent has a weak limit point whose support belongs to $\Omega_{\max}$, $\lim_{\theta\to\infty} \int F\mathrm{e}^{\theta F}/\mathrm{E}_p\left[\mathrm{e}^{\theta F}\right]\ d\mu = \max F$.
4. Practical computational implementations of these scheme look for maximizing sequences in $\mathcal{M}_>$ that belong to a restricted subclass of densities, usually an exponential model.

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 $q \in \mathcal{E}(p)$ we can define the vector field

$$F : q \mapsto u - \mathrm{E}_q [u]$$

2. On the real sample space, for all strictly positive density $f \in C^1(\mathbb{R})$, we define the vector field

$$F : f \mapsto \frac{f'}{f}$$

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

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

- The differential equation above is written with respect to the moving frame at p_θ because $\dot{p}(\theta)/p(\theta)$ is the representation of the velocity vector in ${}^*T(p(\theta))$.
- However, with respect to a fixed frame at p, we should have written

$$\begin{cases} \dot{u}_\theta = F(p(\theta)) - \mathrm{E}_p\left[F(p(\theta))\right] & \text{e-connection, assuming } \dot{u}_\theta \in T_{p_\theta} \\ \dot{l}_\theta = \frac{p}{p(\theta)} F(p(\theta)) & \text{m-connection} \end{cases}$$

Let us consider the exponential model

$$p_\theta = \mathrm{e}^{\theta F} / \mathrm{E}_p \left[\mathrm{e}^{\theta F} \right] \quad \theta \in \mathbb{R}$$

In this case the velocity in the moving frame is

$$\frac{\dot{p}_\theta}{p_\theta} = F - \mathrm{E}_{p_\theta} [F]$$

And the vector field is $p \mapsto F - \mathrm{E}_p [F]$. In general, exponential models are solution of the differential equation for a constant vector field, that is to say a vector field whose unique dependence on p is the centering operation.

A second example follows by considering $\Omega = \mathbb{R}$ and taking for D the class of positive densities f with logaritmic derivative $f'/f \in {}^*T(f)$. For such densities, the mapping $F : f \mapsto F(f) = -f'/f$ is a vector field. We can therefore consider the differential equation $\dot{p}_\theta/p_\theta =$.
If $f \in D$, the translation model $p_\theta(x) = f(x - \theta)$ is such that the score is

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

and the translation model is a solution of the differential equation. The classical Pearson classes of distributions, such as the Cauchy distribution, are special cases of this construction.

More generally, any semigroup τ_t on the space of positive densities, with infinitesimal generator A, i.e. $(d/dt)\tau_t f = A\tau_t f$, on some domain D will produce the same situation. The model $p_\theta = \tau_\theta f$ has score

$$\frac{\dot{p}_\theta}{p_\theta} = \frac{A\tau_\theta f}{\tau_\theta f} = F(p_\theta)$$

where the vector field is defined by $F(q) = A(q)/q$, $q \in D$.

The heat equation

$$\frac{\partial}{\partial t}p(t,x) - \frac{\partial^2}{\partial x^2}p(t,x) = 0$$

is an interesting example of differential equation in $\mathcal{M}_>$. In fact, we can consider the vector field

$$F(p)(x) = \frac{\frac{\partial^2}{\partial x^2}p(x)}{p(x)}$$

Upon division of both sides of of the heat equation by $p(t,x)$, we obtain an equation of our type, whose solution is the solution of the heat equation, i.e. the model obtained by the action of the heat kernel on the initial density. Moreover, the heat equation has a variational form. For each $v \in D$

$$\mathrm{E}_p\left[F(p)v\right] = \int p''(x)v(x)\,dx = -\int p'(x)v'(x)\,dx = -\mathrm{E}_p\left[\frac{p'}{p}v'\right]$$

from which the weak form of the differential equation follows.

Conclusions

- The IG framework suggest a natural language to talk of statistical models from the Fisher point of view.
- In the finite state space case, the IG framework fits the algebraic framework when random variables are described with polynomials.
- Optimization problems are clarified and possibly standard variational theory is usable in prectice.
- There are many results available from the theory of variational differential equation in infinite dimension.
- The classical filtering problem (e.g. Zakai stochastic differential equations) has a precise formulation as a stocastic differential equation on the statistical manifold.