WOGAS3 - Workshop on Geometric and Algebraic Statistics 3
CRiSM - Centre for Research in Statistical Methodology
University of Warwick Apr 5-7 2011

Algebraic Statistics and Information Geometry Examples

Giovanni Pistone giovanni.pistone@carloalberto.org

Collegio Carlo Alberto

April 5th, 2011

Abstract (old)

Statistical models on a finite state space fit in the framework of both Algebraic Statistics and Information Geometry. We discuss the general principles of this fruitful interaction such as the notion of tangent space of a statistical model and toric models. Two examples taken from our recent research work are used for illustration.

1. The study of the critical points of the expected value $\theta \mapsto \mathbb{E}_\theta(f)$ of a generic function f, where θ is the natural parameter of an exponential family with binary sufficient statistics is of interest in relaxed discrete optimization. The study uses both the geometry of exponential families and the algebra of polynomials on a binary space.

2. Reversible Markov chains are defined via algebraic assumptions, i.e. the detailed balance condition, which turn implies that such models are actually toric models. At the same time the information geometry language is used to describe reversible Markov chains as a sub variety of general Markov process.

This is joint work with Luigi Malagò and Maria Piera Rogantin.

Plan

1. Information Geometry
   1.1 Non-parametric Information Geometry
   1.2 Examples of computation
   1.3 Combinatorial optimization
   1.4 Evolution equation
2. Algebraic Statistics
   2.1 Toric statistical models
   2.2 Example: reversible Markov chain

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

Model space

- Let Φ be any Young function equivalent to exp, e.g. $\Phi(x) = \cosh(x) - 1$, with convex conjugate Ψ, e.g. $\Psi(y) = (1 + |y|) \log (1 + |y|) - |y|$. The relevant Orlicz spaces are denoted by L^Φ and L^Ψ, respectively.
- We denote by L_0^Φ, L_0^Ψ the sub-spaces of centered random variables. If the sample space is not finite, then the exponential Orlicz space is not separable and the closure M^Φ of the space of bounded functions is different from L^Φ.
- There is a natural separating duality between L_0^Φ and L_0^Ψ, which is given by the bi-linear form

$$(u, v) \mapsto \int uv \, d\mu$$

- For each $p \in \mathcal{M}_>$ we use the triple of spaces

$$L_0^\Phi(p \cdot \mu) \cong L_0^\Psi(p \cdot \mu)^* \hookrightarrow L_0^2(p \cdot \mu) \hookrightarrow L_0^\Psi(p \cdot \mu).$$

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

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

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 the couple $\exp, \log$ with another couple of functions of interest, e.g $q \mapsto 2\sqrt{q}$. There are many reasons to support our choice. For example, it is possible to derive such a choice from the natural affine structure of positive functions.

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:*

$$\mathrm{D}\, K_p\,(u)\, v = \mathrm{E}_q\left[v\right]$$
$$\mathrm{D}^2\, K_p\,(u)\,(v_1, v_2) = \mathrm{Cov}_q\,(v_1, v_2)$$

- *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.

Example: Exponential family

The exponential family

$$q_\theta = \exp\left(\sum_{j=1}^{d} \theta_j T_j - \psi(\theta)\right) \cdot p$$

is parameterized at p by the random variables

$$u_\theta = \log\left(\frac{q_\theta}{p}\right) - \mathrm{E}_p\left[\log\left(\frac{q_\theta}{p}\right)\right] = \sum_{j=1}^{d} \theta_j \left(T_j - \mathrm{E}_p\left[T_j\right]\right)$$

Example: Decision geometry (Dawid&Lauritzen) I

If the sample space is $\mathbb{R}$ and $p, q \in \mathcal{M}_>$, write $q = \mathrm{e}^{u - K_p(u)} \cdot p$, so that

$$\log q - \log p = u - K_p(u).$$

Assume u belongs to the Sobolev space

$$W^{\Phi,1} = \left\{ u \in L_0^{\Phi}(p) \colon \nabla u \in L_0^{\Phi}(p) \right\}.$$

It follows

$$\begin{aligned} d(p,q) &= \frac{1}{4} \mathrm{E}_p \left[\left\| \nabla \log q - \nabla \log p \right\|^2 \right] \\ &= \frac{1}{4} \mathrm{E}_p \left[\left\| \nabla u \right\|^2 \right]. \end{aligned}$$

Example: Decision geometry (Dawid&Lauritzen) II

For $u, v \in W^{\Phi,1}$ we have a bilinear form

$$\begin{aligned}
\langle u, v \rangle_p = \mathrm{E}_p [\nabla u \nabla v] &= \int u_x(x) v_x(x) p(x) dx \\
&= -\int \nabla (u_x(x) p(x)) v(x) dx \\
&= -\int (\Delta u(x) p(x) + \nabla u(x) \nabla p(x)) v(x) dx \\
&= \mathrm{E}_p [(-\Delta u - \nabla \log p \nabla u) v]
\end{aligned}$$

We have

$$\mathrm{E}_p [\nabla u \nabla v] = \mathrm{E}_p [F_p uv], \qquad F_p u \in {}^*W^{\Phi,1}$$

i.e a classical setting for evolution equations $\partial_t u_t = F_p(u_t)$.

m-charts

Definition

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

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

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

If the sample space is not finite, such a map does not define charts on $\mathcal{M}_>$, nor on $\mathcal{M}_\geq$.

Example: Optimization I

- 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 $\Omega_{\text{max}} \subset \Omega$. 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. $\Omega \neq \Omega_{\text{max}}$, 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 $\nu(\Omega_{\text{max}}) = 1$ 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 $\tilde{F}$.
- The expectation of F is an affine function in the m-chart,

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

Example: Optimization II

- 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

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

- 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

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

Example: binary case (L. Malagò)

Consider the optimization of $\theta \mapsto \mathrm{E}_\theta[F]$ along a binary exponential family

$$p_\theta = \mathrm{E}\left(\sum_{j=1}^{d} \theta_j T_j - K(\theta)\right) \cdot p, \quad T_j^2 = 1.$$

- $\partial_j \mathrm{E}_\theta[F] = \mathrm{Cov}_\theta(F, T_j) = \mathrm{E}_\theta[F T_j] - \mathrm{E}_\theta[F]\,\mathrm{E}_\theta[T_j]$
- θ is is a critical point if $\mathrm{Cov}_\theta(F, T_j) = 0$, $j = 1, \dots, d$. This is not possible if F is a linear combnation of the T_j's or if the remainder after projection on the tangent space is small enough.
- At the critical point the Hessian matrix is

  $$\partial_i \partial_j = \mathrm{Cov}_\theta(F, T_i T_j)$$

  which is not zero if F is a linear combination of the T_j's and the interactions $T_i T_j$'s.
- The diagonal elements of the Hessian matrix are $\partial_i^2 = \mathrm{Cov}_\theta\left(F, T_i^2\right) = \mathrm{Cov}_\theta(F, 1) = 0$. The Hessian matrix is not sign-defined at the critical point.

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.

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

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

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

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

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 families

- 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 $V = \mathrm{Span}\,(u_i, \ldots, u_n)$, 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 $\mathrm{E}_r\left[u\right] = 0$ 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

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

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

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 $q \in \mathcal{M}_>(\mathbb{R}) \cap C^1(\mathbb{R})$
$$F : q \mapsto \frac{q'}{q}$$
3. For all strictly positive $q \in \mathcal{M}_>(\mathbb{R}) \cap C^2(\mathbb{R})$
$$F : q \mapsto q''/q$$

Evolution equation

Definition

A one-parameter statistical model in $\mathcal{M}_>$, $p(\theta)$, $\theta \in I$, solves the evolution 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

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

Heat equation

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 evolution 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. 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 evolution equation follows.
as

$$\mathrm{E}_{p_\theta}\left[\frac{\dot{p}_\theta}{p_\theta}v\right] + \mathrm{E}_{p_\theta}\left[F_0(p_\theta)v\right] = 0 \quad v \in D$$

where F_0 is the vector field associated to the translation model.

A-model: a definition?

- Let be given a (nonnegative) integer matrix $A \in \mathbb{Z}_{\geq}^{m+1,\mathcal{X}}$. The elements are denoted by $A_i(x)$, $i = 0 \dots m, x \in \mathcal{X}$. We assume the row A_0 to be the constant 1. Each row of A is the logarithm of a monomial term denoted $t^{A(x)} = t_0 t_1^{A_1(x)} \cdots t_m^{A_m(x)}$.
- We consider unnormalized probability densities

$$q(x;t) = t^{A(x)}, \quad x \in \mathcal{X}, t \in \mathbb{R}_{\geq}^{m+1}.$$

For each reference measure μ on $\mathcal{X}$ we define the probability density

$$p(x;t) = \frac{t^{A(x)}}{\sum_{x \in \mathcal{X}} t^{A(x)} \ \mu(x)}, \quad x \in \mathcal{X},$$

for all $t \in \mathbb{R}_{\geq}^{m+1}$ such that q_t is not identically zero.

- The parameter t_0 cancels out, i.e the density is parameterized by $t_1 \dots t_m$ only. The unnormalized density is a projective object.

C-constrained A-model; identification

- In some applications the statistical model is further constrained by a matrix $C \in \mathbb{Z}^{k,n}$.

$$\begin{cases} q(x;t) &= t^{A(x)}, \\ \sum_{x \in \mathcal{X}} C_i(x) q(x;t) &= 0, \end{cases}$$

  for $\quad x \in \mathcal{X}, t \in \mathbb{R}_{\geq}^{m+1}, i = 1 \ldots k.$

- Assume $s, t \in \mathbb{R}_{>}^{m}$ and $p_s = p_t$. Denote by Z the normalizing constant. Then $p_t = p_s$ if, and only if,

$$Z(s) t^{A(x)} = Z(t) s^{A(x)}, \quad x \in \mathcal{X}$$

  hence

$$\sum_{i=0}^{m} (\log t_i - \log s_i) A_i(x) = \log Z(t) - \log Z(s), \quad x \in \mathcal{X}.$$

  The confounding condition is

$$\delta^T A = 1, \quad \delta_i = (\log t_i - \log s_i)/(\log Z(t) - \log Z(s)),$$

  so that $\delta \in e_0 + \ker A^T$.

Toric ideals; closure of the A-model

- The ker of the ring homomorphism

$$k[q(x) \colon x \in \mathcal{X}] \ni q(x) \mapsto t^{A(x)} \in k[t_0, \dots, t_m]$$

is the toric ideal of A, $\mathrm{I}(A)$. It has a finite basis made of binomials of the form

$$\prod_{x \colon u(x) > 0} q(x)^{u^+(x)} - \prod_{x \colon u(x) < 0} q(x)^{u^-(x)}$$

with $u \in \mathbb{Z}^{\mathcal{X}}$, $Au = 0$.

- As $\sum_{x \in \mathcal{X}} u(x) = 0$, all the binomials are homogeneous polynomials so that all densities p_t in the A-model satisfy the same binomial equation.

Theorem

- *The nonnegative part of the A-variety is the (weak) closure of the positive part of the A-model.*
- *Let $\mathcal{H}$ be the Hilbert basis of* $\mathrm{Span}\,(A_0, A_1, \dots) \cap Z_{\geq}^{\mathcal{X}}$. *Let H be the matrix whose rows are the elements of $\mathcal{H}$ of minimal support. The H-model is equal to the nonnegative part of the A-variety.*

Markov Chains MCs

- In a Markov chain with state space V, initial probability π_0 and stationary transitions $P_{u\to v}$, $u, v \in V$, the joint distribution up to time T on the sample space Ω_T is

$$P(\omega) = \prod_{v\in V} \pi_0(v)^{(X_0(\omega)=v)} \prod_{a\in\mathcal{A}} P_a^{N_a(\omega)}, \qquad \text{(M)}$$

  where $(V, \mathcal{A})$ is the directed graph defined by $u \to v \in \mathcal{A}$ if, and only if, $P_{u\to v} > 0$.

- A MC is an instance of the A model with $m = \#V + \#\mathcal{A}$, $n = \#\Omega_T$ and rows

$$A_0(\omega) = 1, A_v(\omega) = (X_0(\omega) = 1), A_a(\omega) = N_a(\omega)$$

  i.e the unnormalized density is

$$q(\omega; t) = t_0 \prod_{v\in V} t_v^{(X_0(\omega)=v)} \prod_{a\in\mathcal{A}} t_a^{N_a(\omega)} \qquad \text{(A)}$$

- The (MC) model is derived from the (A) model by adding the constrains

$$\sum_{v\in V} t_v = \sum_{v:\ u\to v\in\mathcal{A}} q_{u\to v}, \quad u \in V.$$

A-model of a MC

- The unconstrained A-model of the MC is a Markov proces with non-stationary transition probabilities.
- The unconstrained model is described probabilistically as follows. Define $a(v) = \sum_{v \to w \in \mathcal{A}} t_{v \to w}$; hence $P_{u \to v} = t_{v \to w}/a(v)$ is a transition probability. Also $\nu(v) = a(v)/\sum_v a(v)$ is a probability. Consider the change of parameters

$$b\pi(v) = t_v, a\nu(v)P_{v \to w} = t_{v \to w},$$

to get

$$\begin{aligned} q(\omega;) &= t_0 \prod_{v \in V} (b\pi(v))^{(X_0(\omega)=v)} \prod_{v \to w \in \mathcal{A}} (aa(v)P_{v \to w})^{N_{v \to w}(\omega)} \\ &= t_0 ba^N \prod_{v \in V} \pi(v)^{(X_0(\omega)=v)} \prod_{v \in V} \nu(v)^{N_{v+}} \prod_{v \to w \in \mathcal{A}} P_{v \to w}^{N_{v \to w}(\omega)} \end{aligned}$$

- It is a change in reference measure.

Detailed balance

- Consider a simple graph $(V, \mathcal{A})$.
- A transition matrix $P_{v \to w}$, $v, w \in V$, satisfies the detailed balance conditions if $\kappa(v) > 0$, $v \in V$, and

$$\kappa(v) P_{v \to w} = \kappa(w) P_{w \to v}, \quad v \to w \in \mathcal{A}.$$

- It follows that $\pi(v) \propto \kappa(v)$ is an invariant probability and the Markov chain X_n, $n = 0, 1, \ldots,$ has reversible two-step joint distribution

$$\mathrm{P}\left(X_n = v, X_{n+1} = w\right) = \mathrm{P}\left(X_n = w, X_{n+1} = v\right), \quad v, w \in V, n \geq 0.$$

Reversibility on trajectories

Let $\omega = v_0 \cdots v_n$ be a trajectory (path) in the connected graph $\mathcal{G} = (V, \mathcal{E})$ and let $r\omega = v_n \cdots v_0$ be the reversed trajectory.

Proposition

If the detailed balance holds, the the reversibility condition

$$\mathrm{P}(\omega) = \mathrm{P}(r\omega)$$

holds for each trajectory ω.

Proof.

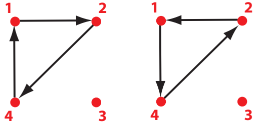
Write the detailed balance along the trajectory,

$$\begin{aligned}
\pi(v_0)P_{v_0 \to v_1} &= \pi(v_1)P_{v_1 \to v_0}, \\
\pi(v_1)P_{v_1 \to v_2} &= \pi(v_2)P_{v_2 \to v_1}, \\
&\vdots \\
\pi(v_{n-1})P_{v_{n-1} \to v_n} &= \pi(v_n)p_{v_n \to v_{n-1}},
\end{aligned}$$

and clear $\pi(v_1) \cdots \pi(v_{n-1})$ in both sides of the product. $\square$

Kolmogorov's condition

We denote by ω a closed trajectory, that is a trajectory on the graph such that the last state coincides with the first one, $\omega = v_0 v_1 \dots v_n v_0$, and by $r\omega$ the reversed trajectory $r\omega = v_0 v_n \dots v_1 v_0$



Theorem (Kolmogorov)

Let the Markov chain $(X_n)_{n \in \mathbb{N}}$ have a transition supported by the connected graph $\mathcal{G}$.

- *If the process is reversible, for all closed trajectory*

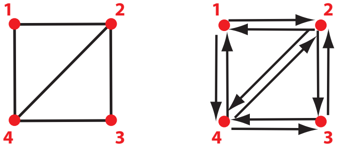
$$\boxed{P_{v_0 \to v_1} \cdots P_{v_n \to v_0} = P_{v_0 \to v_n} \cdots P_{v_1 \to v_0}}$$

- *If the equality is true for all closed trajectory, then the process is reversible.*

- The Kolmogorov's condition does not involve the π.
- Detailed balance, reversibility, Kolmogorov's condition are algebraic in nature and define binomial ideals.

Transition graph

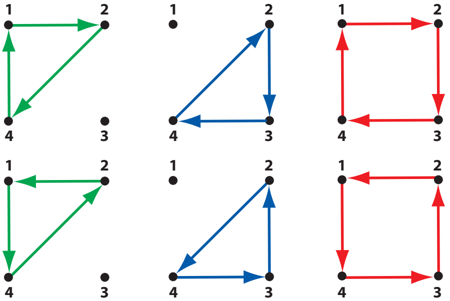
- From $\mathcal{G} = (V, \mathcal{E})$ an (undirected simple) graph, split each edge into two opposite arcs to get a connected directed graph (without loops) $\mathcal{O} = (V, \mathcal{A})$. The arc going from vertex v to vertex w is $(v \to w)$. The reversed arc is $r(v \to w) = (w \to v)$.



- A path or trajectory is a sequence of vertices $\omega = v_0 v_1 \cdots v_n$ with $(v_{k-1} \to v_k) \in \mathcal{A}$, $k = 1, \ldots, n$. The reversed path is $r\omega = v_n v_{n-1} \cdots v_0$. Equivalently, a path is a sequence of inter-connected arcs $\omega = a_1 \ldots a_n$, $a_k = (v_{k-1} \to v_k)$, and $r\omega = r(a_n) \ldots r(a_1)$.

Circuits, cycles

- A closed path $\omega = v_0 v_1 \cdots v_{n-1} v_0$ is any path going from an initial v_0 back to v_0; $r\omega = v_0 v_{n-1} \cdots v_1 v_0$ is the reversed closed path. If we do not distinguish any initial vertex, the equivalence class of closed paths is called a circuit.
- A closed path is elementary if it has no proper closed sub-path, i.e. if does not meet twice the same vertex except the initial one v_0. The circuit of an elementary closed path is a cycle.



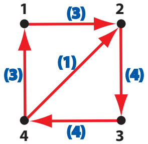
Kolmogorov's ideal

- With indeterminates $P = [P_{v \to w}]$, $(v \to w) \in \mathcal{A}$, form the ring $k[P_{v \to w} : (v \to w) \in \mathcal{A}]$. For a trajectory ω, define the monomial term

$$\omega = a_1 \cdots a_n \mapsto P^{\omega} = \prod_{k=1}^{n} P_{a_k} = \prod_{a \in \mathcal{A}} P_a^{N_a(\omega)},$$

with $N_a(\omega)$ the number of traversals of the arc a by the trajectory.

2→3 3→4 4→1 1→2
2→3 3→4 4→2
2→3 3→4 4→1 1→2
2→3 3→4 4→1 1→2



$P_{1\to 2}^{3} P_{2\to 3}^{4} P_{3\to 4}^{4} P_{4\to 1}^{3} P_{4\to 2}$

Definition (K-ideal)

The Kolmogorov's ideal or K-ideal of the graph $\mathcal{G}$ is the ideal generated by the binomials $P^{\omega} - P^{r\omega}$, where ω is any circuit.

Bases of the K-ideal

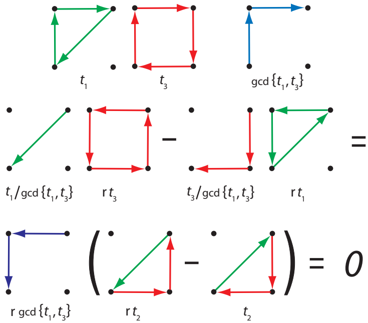
Finite basis of the K-ideal

The K-ideal is generated by the set of binomials $P^{\omega} - P^{r\omega}$, where ω is cycle.

Universal G-basis

The binomials $P^{\omega} - P^{r\omega}$, where ω is any cycle, form a reduced universal Gröbner basis of the K-ideal.

Six cycles: $\omega_1 = 1 \to 2\ 2 \to 4\ 4 \to 1$ (green), $\omega_2 = 2 \to 3\ 3 \to 4\ 4 \to 2$, $\omega_3 = 1 \to 2\ 2 \to 3\ 3 \to 2\ 4 \to 1$ (red), $\omega_4 = r\omega_1$, $\omega_5 = r\omega_2$,$\omega_6 = r\omega_3$.



Cycle space of $\mathcal{O}$

- For each cycle ω define cycle vector

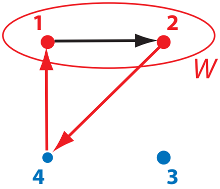
$$z_a(\omega) = \begin{cases} +1 & \text{if } a \text{ is an arc of } \omega, \\ -1 & \text{if } r(a) \text{ is an arc of } \omega, \\ 0 & \text{otherwise.} \end{cases} \qquad a \in \mathcal{A}.$$

- The binomial $P^{\omega} - P^{r\omega}$ is written as $P^{z^+(\omega)} - P^{z^-(\omega)}$.
- The definition of z can be is extended to any circuit $\bar{\omega}$ by $z_a(\bar{\omega}) = N_a(\omega) - N_a(r\omega)$.
- There exists a sequence of cycles such that $z(\bar{\omega}) = z(\omega_1) + \cdots + z(\omega_l)$.
- We can find nonnegative integers $\lambda(\omega)$ such that $z(\bar{\omega}) = \sum_{\omega \in \mathcal{C}} \lambda(\omega) z(\omega)$, i.e. it belongs to the integer lattice generated by the cycle vectors.
- $Z(\mathcal{O})$ is the cycle space, i.e. the vector space generated in $k^{\mathcal{A}}$ by the cycle vectors.

Cocycle space of $\mathcal{O}$

- For each subset W of V, define cocycle vector

$$u_a(W) = \begin{cases} +1 & \text{if } a \text{ exits from } W, \\ -1 & \text{if } a \text{ enters into } W, \\ 0 & \text{otherwise.} \end{cases} \qquad a \in \mathcal{A}.$$



- The generated subspace of $k^{\mathcal{A}}$ is the cocycle space $U(\mathcal{O})$
- The cycle space and the cocycle space orthogonally split the vector space $\{y \in k^{\mathcal{A}}\colon y_a = -y_{r(a)}, a \in \mathcal{A}\}$.
- Note that for each cycle vector $z(\omega)$, cocycle vector $u(W)$, $z_a(\omega)u_a(W) = z_{r(a)}(\omega)u_{r(a)}(W)$, $a \in \mathcal{A}$, hence

$$z(\omega)\cdot u(W) = 2\sum_{a\in\omega} u_a(W) = 2\left[\sum_{a\in\omega, u_a(W)=+1} 1 - \sum_{a\in\omega, u_a(W)=-1} 1\right] = 0.$$

Toric ideals

- Let U be the matrix whose rows are the cocycle vectors $u(W)$, $W \subset V$. We call the matrix $U = [u_a(W)]_{W \subset V, a \in \mathcal{A}}$ the cocycle matrix.
- Consider the ring $k[P_a \colon a \in \mathcal{A}]$ and the Laurent ring $k(t_W \colon W \subset V)$, together with their homomorphism h defined by

$$h \colon P_a \longmapsto \prod_{W \subset V} t_W^{u_a(W)} = t^{u_a}.$$

- The kernel $I(U)$ of h is the toric ideal of U. It is a prime ideal and the binomials $P^{z^+} - P^{z^-}$, $z \in \mathbb{Z}^{\mathcal{A}}$, $Uz = 0$ are a generating set of $I(U)$ as a k-vector space.
- As for each cycle ω we have $Uz(\omega) = 0$, the cycle vector $z(\omega)$ belongs to $\ker_{\mathbb{Z}} U = \left\{z \in \mathbb{Z}^{\mathcal{A}} \colon Uz = 0\right\}$. Moreover, $P^{z^+(\omega)} = P^{\omega}$, $P^{z^-(\omega)} = P^{r\omega}$, therefore the K-ideal is contained in the toric ideal $I(U)$.

The K-ideal is toric

Theorem

The K-ideal is the toric ideal of the cocycle matrix.

Definition (Graver basis)

$z(\omega_1)$ is conformal to $z(\omega_2)$, $z(\omega_1) \sqsubseteq z(\omega_2)$, if the component-wise product is non-negative and $|z(\omega_1)| \leq |z(\omega_2)|$ component-wise, i.e. $z_a(\omega_1)z_a(\omega_2) \geq 0$ and $|z_a(\omega_1)| \leq |z_a(\omega_2)|$ for all $a \in \mathcal{A}$. A Graver basis of $Z(\mathcal{O})$ is the set of the minimal elements with respect to the conformity partial order $\sqsubseteq$.

Theorem

1. *For each cycle vector* $z \in Z(\mathcal{O})$, $z = \sum_{\omega \in \mathcal{C}} \lambda(\omega) z(\omega)$, *there exist cycles* $\omega_1, \ldots, \omega_n \in \mathcal{C}$ *and positive integers* $\alpha(\omega_1), \ldots, \alpha(\omega_n)$, *such that* $z^+ \geq z^+(\omega_i)$, $z^- \geq z^-(\omega_i)$, $i = 1, \ldots, n$ *and* $z = \sum_{i=1}^{n} \alpha(\omega_i) z(\omega_i)$.
2. *The set* $\{z(\omega) \colon \omega \in \mathcal{C}\}$ *is a Graver basis of* $\mathcal{Z}(\mathcal{O})$. *The binomials of the cycles form a Graver basis of the K-ideal.*

Positive K-ideal

- The strictly positive reversible transition probabilities on $\mathcal{O}$ are given by:

$$\begin{aligned} P_{v \to w} &= s(v,w) \prod_S t_S^{u_{v \to w}(S)} \\ &= s(v,w) \prod_{S:\ v \in S, w \notin S} t_S \prod_{S:\ w \in S, v \notin S} t_S^{-1}, \end{aligned}$$

  where $s(v,w) = s(w,v) > 0$, $t_S > 0$.
- The first set of parameters, $s(v,w)$, is a function of the edge.
- The second set of parameters, t_S, represent the deviation from symmetry. The second set of parameters is not identifiable because the rows of the U matrix are not linearly independent.
- The parametrization can be used to derive an explicit form of the invariant probability.

Parametric detailed balance

Theorem

Consider the strictly non-zero points on the K-variety.

1. *The symmetric parameters* $s(e)$, $e \in \mathcal{E}$, *are uniquely determined. The parameters* t_S, $S \subset V$ *are confounded by* $\ker U = \{U^t t = 0\}$.
2. *An identifiable parametrization is obtained by taking a subset of parameters corresponding to linearly independent rows, denoted by* t_S, $S \subset \mathcal{S}$*:*

$$P_{v \to w} = s(v, w) \prod_{S \subset \mathcal{S} :\ v \in S, w \notin S} t_S \prod_{S \subset \mathcal{S} :\ w \in S, v \notin S} t_S^{-1}$$

3. *The detailed balance equations,* $\kappa(v) P_{v \to w} = \kappa(w) P_{w \to v}$, *are verified if, and only if,*

$$\kappa(v) \propto \prod_{S :\ v \in S} t_S^{-2}$$

Detailed balance ideal

Definition

The detailed balance ideal is the ideal

$$\text{Ideal}\left(\prod_{v\in V}\kappa(v)-1, \kappa(v)P_{v\to w}-\kappa(w)P_{v\to w},\ (v\to w)\in\mathcal{A}\right).$$

in $k[\kappa(v) : v\in V, P_{v\to w}, (v\to w)\in\mathcal{A}]$

1. The matrix $[P_{v\to w}]_{v\to w\in\mathcal{A}}$ is a point of the variety of the K-ideal if and only if there exists $\kappa=(\kappa(v)\colon v\in V)$ such that (κ, P) belongs to the variety of the detailed balance ideal.
2. The detailed balance ideal is a toric ideal.
3. The K-ideal is the κ-elimination ideal of the detailed balance ideal.

Parameterization of reversible transitions

- There exist a (non algebraic) parametrization of the non-zero K-variety of the form

$$P_{v \to w} = s(v,w)\kappa(w)^{1/2}\kappa(v)^{-1/2}$$

- Such a P is a reversible transition probability strictly positive on the graph $\mathcal{G}$ with invariant probability proportional to κ if, and only if,

$$\kappa(v)^{1/2} \geq \sum_{w \neq v} s(u,w)\kappa(w)^{-1/2}.$$

- In the Hastings-Metropolis algorithm, we are given an unnormalized positive probability κ and a transition $Q_{v \to w} > 0$ if $(v \to w) \in \mathcal{A}$. We are required to produce a new transition $P_{v \to w} = Q_{v \to w}\alpha(v,w)$ such that P is reversible with invariant probability κ and $0 < \alpha(v,w) \leq 1$. We have

$$Q_{v \to w}\alpha(v,w) = s(v,w)\kappa(w)^{1/2}\kappa(v)^{-1/2}$$

and moreover we want

$$\alpha(v,w) = \frac{s(v,w)\kappa(w)^{1/2}}{Q_{v \to w}\kappa(v)^{1/2}} \leq 1.$$

Metropolis–Hastings algorithm

Proposition

Let Q be a probability on $V \times V$, strictly positive on $\mathcal{E}$, and let $\pi(x) = \sum_y Q(x,y)$. If $f :]0,1[\times]0,1[\to]0,1[$ is a symmetric function such that $f(u,v) \leq u \wedge v$ then

$$P(x,y) = \begin{cases} f(Q(x,y), Q(y,x)) & \{x,y\} \in \mathcal{E} \\ \pi(x) - \sum_{y \colon y \neq x} P(x,y) & x = y \\ 0 & \text{otherwise}, \end{cases}$$

is a 2-reversible probability on $\mathcal{E}$ such that $\pi(x) = \sum_y P(x,y)$, positive if Q is positive.

The proposition applies to

- $f(u,v) = u \wedge v$. This is the Hastings case: $u \wedge v = u(1 \wedge (v/u))$
- $f(u,v) = uv/(u+v)$. This is the Barker case: $uv/(u+v) = u(1 + u/v)^{-1}$
- $f(u,v) = uv$. This is one of the Hastings general form.

Gröbner basis: recap I

- The K-ideal is generated by a finite set of binomials. A Gröbner basis is a special class of generating set of an ideal. We refer to and for the relevant necessary and sufficient conditions.
- The theory is based on the existence of a monomial order, i.e. a total order on monomial term which is compatible with the product. Given such an order, the leading term $\mathrm{LT}(f)$ of the polynomial f is defined. A generating set is a Gröbner basis if the set of leading terms of the ideal is generated by the leading terms of monomials in the generating set. A Gröbner basis is reduced if the coefficient of the leading term of each element of the basis is 1 and no monomial in any element of the basis is in the ideal generated by the leading terms of the other element of the basis. The Gröbner basis property depend on the monomial order. However, a generating set is a universal Gröbner basis if it is a Gröbner basis for all monomial orders.

Gröbner basis: recap II

- The finite algorithm for computing a Gröbner basis depends on the definition of sygyzy. Given two polynomial f and g in the polynomial ring K, their sygyzy is the polynomial

$$S(f,g) = \frac{\mathrm{LT}(g)}{\gcd(\mathrm{LT}(f),\mathrm{LT}(g))} f - \frac{\mathrm{LT}(f)}{\gcd(\mathrm{LT}(f),\mathrm{LT}(g))} g.$$

  A generating set of an ideal is a Gröbner basis if, and only if, it contains the sygyzy $S(f,g)$ whenever it contains f and g, see Chapter 6 in or Theorem 2.4.1 p. 111 of .

Bibliography I

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.

Alberto Cena and Giovanni Pistone. Exponential statistical manifold. Ann. Inst. Statist. Math., 59(1):27–56, 2007. ISSN 0020-3157.

David Cox, John Little, and Donal O'Shea. Ideals, varieties, and algorithms: An introduction to computational algebraic geometry and commutative algebra. Undergraduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1997. ISBN 0-387-94680-2.

Paolo Gibilisco and Giovanni Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. IDAQP, 1(2):325–347, 1998. ISSN 0219-0257.

Martin Kreuzer and Lorenzo Robbiano. Computational commutative algebra. 1. Springer-Verlag, Berlin, 2000. ISBN 3-540-67733-X.

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`.

Giovanni Pistone. κ-exponential models from the geometrical viewpoint. The European Phisical Journal B Condensed Matter Physics, 71(1):29–37, July I 2009. ISSN 1434-6028. doi10.1140/epjb/e2009-00154-y. URL `http://dx.medra.org/10.1140/epjb/e2009-00154-y`.

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.

Giovanni Pistone and Maria Piera Rogantin. The algebra of reversible markov chains. arXiv:1007.4282v2 [math.ST], 2011.

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.

P. Suomela. Invariant measures of time-reversible Markov chains. J. Appl. Probab., 16(1):226–229, 1979. ISSN 0021-9002.