The 2^{nd} International Symposium on Information Geometry and its Applications

Advances in the geometry of non-parametric exponential models

Giovanni Pistone

DIMAT Politecnico di Torino

giovanni.pistone@polito.it http://staff.polito.it/giovanni.pistone/igaia2-talk.pdf

Abstract 1

The theory of **exponential manifolds modeled on Orlicz spaces**, based on my joint work with C. Sempi, M.-P. Rogantin, P. Gibilisco (1995-1999), has been improved in the basic construction in the PhD thesis of A. Cena (2002). He also made some advancement in the study of the related Amari connections.

- G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, 23(5):1543–1561, October 1995;
- G. Pistone and M. P. Rogantin. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5(4):721–760, August 1999;
- P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.*, 1(2):325–347, 1998;
- A. Cena. *Geometric structures on the non-parametric statistical manifold*. PhD thesis, Dottorato in Matematica, Università di Milano, 2002.

The first part of the talk will review this basic improved construction and review the particular case of Wiener spaces and related joint work with P. Gibilisco and D. Imparato.

Abstract 2

The approximation of non parametric models with parametric ones recently raised some interest, especially with reference to applications to Finance Mathematics.

However, some of the proposed approximation methods do not converge in the sense on the manifold topology, which is a very strong topology. The approximation with **finite state space** appears to be more promising in this sense.

Abstract 3

Information geometry of finite state space models has special algebraic features that where first discussed by Pistone, Riccomagno and Wynn (2001) . The key-word is the notion of **toric ideal** firstly used in Statistics by P. Diaconis and B. Sturmfels (1998). An important feature of this theory is associated with the use of Symbolic Computational Software such as Singular or CoCoA. On the other side, the use of algebraic geometry ideas in continuous state space is just at the beginning, as in Pistone & Wynn "Finitely generated cumulants" (1999). This last and main part of the talk will discuss the connections between Information Geometry and Algebraic Geometry.

- P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 1998 (preprint 1993!)
- G. Pistone, E. Riccomagno, and H. P. Wynn. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman&Hall, 2001.
- L. Patcher and B. Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.
- G. Pistone and H. P. Wynn. Generalised confounding with Gröbner bases. *Biometrika*, 83(3):653–666, Mar. 1996.

Outline

Exponential statistical model : review and update of the basic construction. The main results relates with the functional analytic properties of the main objects of the theory.

Approximation : suggestion to approximate non parametric problem with finite state space models.

Finite state space : finite state space models and special models such as Gaussian models have special algebraic features. This features can be dealt with the use of modern constructive commutative algebra.

- D. A. Cox, J. B. Little, and D. O'Shea. *Ideal, Varieties, and Algorithms*. Springer-Verlag, New York, 2nd edition, 1997. 1st ed. 1992;
- M. Kreuzer and L. Robbiano. *Computational Commutative Algebra 1*. Springer, Berlin-Heidelberg, 2000;
- CoCoATeam. CoCoA: a system for doing Computations in Commutative Algebra. Available at `http://cocoa.dima.unige.it`, no date.

IG as a Banach manifold

Our construction of the exponential statistical manifold wants to be a functional framework for the development of IG in the sense of professor Amari seminal work, e.g.

- S. Amari. *Differential-geometrical methods in statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1985
- S. Amari and H. Nagaoka. *Methods of information geometry*. American Mathematical Society, Providence, RI, 2000. Translated from the 1993 Japanese original by Daishi Harada

Given a probability space $(X, \mathcal{X}, \mu)$, we will denote by $\mathcal{M}$ the set of all densities which are positive μ-a.s. $\mathcal{M}$ is thought to be the maximal regular statistical model. We want to give to this maximal model a manifold structure in such a way that each specific statistical model could be considered as a **submanifold** of $\mathcal{M}$.

The model space for the manifold are locally at each $p \in \mathcal{M}$ the Orlicz space of centered random variable, see e.g. M. M. Rao and Z. D. Ren. *Applications of Orlicz spaces*, volume 250 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., New York, 2002.

Orlicz spaces

- The Jung function $\Phi(x) = \cosh x - 1$ is used instead of the quivalent and more commonly used $\mathrm{e}^{|x|} - |x| - 1$.
- Ψ denotes its conjugate Jung function or the equivalent $(1+y)\log(1+y) - y$.
- A random variable u belongs to the vector space $L^{\Phi}(p)$ if for some $\alpha > 0$ $\mathrm{E}_p\left(\Phi(\alpha u)\right) < +\infty$.
- The closed unit ball of $L^{\Phi}(p)$ consists of all u's such that $\mathrm{E}_p\left(\Phi(u)\right) \leq 1$.
- The open unit ball $B(0,1)$ consists of those u's such that αu is in the closed unit ball for some $\alpha > 1$.
- The Banach space $L^{\Phi}(p)$ is not separable, like L^{∞}. In this sense it is an un-natural choice.
- However, $L^{\Phi}(p)$ is natural for statistics because for each $u \in L^{\Phi}(p)$ the Laplace transform of u is well defined at 0 and the one-dimensional exponential model $p(\theta) \propto \mathrm{e}^{\theta u}$ is well defined.
- The space $L^{\Psi}(p)$ is separable and it is the pre-dual of $L^{\Phi}(p)$, with pairing $\mathrm{E}_p\left(uv\right)$. For $1 < a < +\infty$, $L^{\Phi}(p) \subset L^{a}(p) \subset L^{\Psi}(p)$.
- For a given $p \in \mathcal{M}$, we define the **moment functional** to be $M_p(u) = \mathrm{E}_p\left(\mathrm{e}^{u}\right)$.

Moment functional

- $M_p(0) = 1$; otherwise, for each $u \neq 0$, $M_p(u) > 1$.
- M_p is convex and lower semi-continuous, and its proper domain

$$\operatorname{dom}(M_p) = \left\{ u \in L^\Phi(p \cdot \mu) : M_p(u) < \infty \right\}$$

  is a convex set which contains the open unit ball $B(0,1) \subset L^\Phi(p \cdot \mu)$.
- M_p is infinitely Gâteaux-differentiable in the interior of its proper domain, the nth-derivative at $u \in \overset{\circ}{\operatorname{dom}}(M_p)$ in the direction $v \in L^\Phi(p)$ being

$$\left. \frac{d^n}{dt^n} M_p(u + tv) \right|_{t=0} = \mathrm{E}_p\left(v^n \mathrm{e}^u\right);$$

- M_p is bounded, **infinitely Fréchet-differentiable and analytic on the open unit ball** of $L^\Phi(p)$, the nth-derivative at $u \in B(0,1)$ evaluated in $(v_1, \dots, v_n) \in L^\Phi(p) \times \cdots \times L^\Phi(p)$ is

$$D^n M_p(u)(v_1, \dots, v_n) = \mathrm{E}_p\left(v_1 \cdots v_n \mathrm{e}^u\right).$$

  In particular, $DM_p(0) = \mathrm{E}_p(\cdot)$.

Cumulant functional

- For a given $p \in \mathcal{M}$ and all p-centered u's in the interior of the proper domain of M_p, define the **cumulant functional** as $K_p(u) = \log M_p(u)$.
- K_p is infinitely Gâteaux-differentiable.
- K_p is bounded, **infinitely Fréchet-differentiable and analytic on the open unit ball** of $L_0^\Phi(p)$.
- If $\mathcal{V}_p$ is the open unit ball in $L_0^\Phi(p)$,

$$e_p : \begin{cases} \mathcal{V}_p \to \mathcal{M} \\ u \mapsto \mathrm{e}^{u - K_p(u)} p \end{cases}$$

  is a **local parameterization** of $\mathcal{M}$.
- If $e_p(\mathcal{V}_p) = \mathcal{U}_p$, the corresponding **chart** is

$$s_p : \begin{cases} \mathcal{U}_p \to \mathcal{V}_p \\ q \mapsto \log\left(\frac{q}{p}\right) - E_p\left[\log\left(\frac{q}{p}\right)\right] \end{cases}$$

- If $q = e_p(u)$, then $DK_p(u)v = \mathrm{E}_q(v)$ and $D^2K_p(u)vw = \mathrm{E}_q(vw)$

Connected component

- In the theory of statistical models, we associate to each density p a space of p-centered random variables: scores, estimating functions .... It is technically crucial to discuss how the relevant spaces depend on the variation of the density p.
- Given $p, q \in \mathcal{M}$, the exponential model $p(\theta) \propto p^{1-\theta} q^{\theta}$, $0 \leq \theta \leq 1$ connects the two given densities as end points of a curve, sometimes called Hellinger arc. This curve need not to be continuous for the topology we are going to put on $\mathcal{M}$.

D We say that $p, q \in \mathcal{M}$ are **connected by an open exponential arc** if there exist $r \in \mathcal{M}$, $u \in L_0^{\Phi}(r)$ and an open interval I that contains 0, and such that $p(t) \propto \mathrm{e}^{tu} \cdot r$, $t \in I$, is an exponential model containing both p and q

Th Let p and q be densities connected by an open exponential arc. **Then the Banach spaces $L^{\Phi}(p)$ and $L^{\Phi}(q)$ are equal as vector spaces and their norms are equivalent**.

Th For all q that are connected to p by an open exponential arc, the Orlicz space of centered random variables at q is $L_0^{\Phi}(q) \sim \left\{u \in L^{\Phi}(p) \,|\, \mathrm{E}_p\left(\frac{q}{p} u\right) = 0\right\}$, then it is equivalent to the **orthogonal space of** $\left(\frac{q}{p} - 1\right)$.

Exponential manifold

D For every $p \in \mathcal{M}$, the **maximal exponential model at** p is defined to be the family of densities

$$\mathcal{E}(p) := \left\{ \mathrm{e}^{u - K_p(u)} p : u \in \overset{\circ}{\operatorname{dom}} K_p \right\} \subseteq \mathcal{M}_\mu.$$

Th The maximal exponential model at p, $\mathcal{E}(p)$ is equal to the set of all densities $q \in \mathcal{M}$ connected to p by an open exponential arc.

Th The collection of charts $\{(\mathcal{U}_p, s_p) : p \in \mathcal{M}\}$ is an affine $\mathcal{C}^\infty$ atlas on $\mathcal{M}_\mu$.

- The transition maps are

$$s_{p_2} \circ e_{p_1} : \begin{cases} s_{p_1}(\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}) & \to s_{p_2}(\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}) \\ u & \mapsto u + \log\left(\frac{p_1}{p_2}\right) - \mathrm{E}_{p_2}\left[u + \log\left(\frac{p_1}{p_2}\right)\right] \end{cases}$$

- The derivative of the transition map $s_{p_2} \circ e_{p_1}$ is

$$L_0^\phi(p_1) \ni u \mapsto u - \mathrm{E}_{p_2}(u) \in L_0^\phi(p_2)$$

which is a top-linear isomorphism, because p_1 and p_2 are connected by an open exponential arc.

Divergence

- As $\overset{\circ}{\operatorname{dom}} K_p \ni u \leftrightarrow e^{u-K_p(u)} \cdot p \in \mathcal{E}(p)$, the manifold is actually defined by **global charts**.

Th For each $p \in \mathcal{M}$, the KL divergence

$$D(\cdot \| \cdot) : \mathcal{E}(p) \times \mathcal{E}(p) \to \mathbb{R}$$

is of class $\mathcal{C}^\infty$.

- Proof:
  1. Given the charts $(\mathcal{U}_{p_1}, s_{p_1})$ and $(\mathcal{U}_{p_2}, s_{p_2})$ of $\mathcal{E}(p)$, we consider the local representative $D_{p_1,p_2} = D \circ (e_{p_1}, e_{p_2}) : \mathcal{V}_{p_1} \times \mathcal{V}_{p_2} \to \mathbb{R}$.
  2. $$D(q_1 \| q_2) = \mathbf{D}K_{p_1}(u_1) \cdot (u_1 - s_{p_1} \circ e_{p_2}(u_2)) - K_{p_1}(u_1) + K_{p_2}(u_2) - \mathrm{E}_{p_1}\left[u_2 + \log \frac{p_2}{p_1}\right]$$
  3. $$\mathbf{D}_1 D_{p_1,p_2}(u_1, u_2) \cdot w_1 = \mathbf{D}^2 K_{p_1}(u_1) \cdot (u_1 - s_{p_1} \circ e_{p_2}(u_2), w_1)$$
     $$\mathbf{D}_2 D_{p_1,p_2}(u_1, u_2) \cdot w_2 = -\mathbf{D}K_{p_1}(u_1) \cdot (w_2 - \mathrm{E}_{p_1}[w_2]) + \mathbf{D}K_{p_2}(u_2) \cdot w_2 - \mathrm{E}_{p_1}[w_2].$$

The exponential geometry

The previous theory is intended to capture the essence and to generalize the idea of **curved exponential model** as defined by

- B. Efron. Defining the curvature of a statistical problem (with applications to second–order efficiency). *The Annals of Statististics*, 3:1189–1242, 1975. (with discussion)
- B. Efron. The geometry of exponential families. *Ann. Statist.*, 6(2):362–376, 1978;
- A. P. Dawid. Discussion of a paper by Bradley Efron. *The Annals of Statistics*, 3:1231–1234, 1975
- A. P. Dawid. Further comments on a paper by Bradley Efron. *The Annals of Statistics*, 5:1249, 1977

From the work of Professor Amari, we know that there is a second geometry on probabilities, whose geodesics are mixtures. This structure is a *connection* on a special vector bundle of the $\mathcal{M}$-manifold. A related manifold on the set $\mathcal{P}$ of *normalized* random variables is defined by the charts

$$\boxed{q \mapsto \frac{q}{p} - 1}$$

Locally, the q's have finite divergence $D(q|p)$.

Mixture manifold 1

- We enlarge $\mathcal{M}$ considering the sets

$$\mathcal{P}_{\geq} = \left\{ p \in L^1(\mu) : p \geq 0, \int p d\mu = 1 \right\}$$

$$\mathcal{P} = \left\{ p \in L^1(\mu) : \int p d\mu = 1 \right\}.$$

- For each $p \in \mathcal{P}_{\geq}$, we define the sets $L_0^{\Psi}(p)$ and

$$^{*}\mathcal{U}_p = \left\{ q \in \mathcal{P} : \frac{q}{p} \in L^{\Psi}(p) \right\}$$

and the map

$$\eta_p : \begin{cases} ^{*}\mathcal{U}_p \to L_0^{\Psi}(p) \\ q \mapsto \frac{q}{p} - 1 \end{cases}$$

with inverse:

$$L_0^{\Psi}(p) \ni u \mapsto (u+1)\, p \in {}^{*}\mathcal{U}_p.$$

- The collection of sets $\{^{*}\mathcal{U}_p\}_{p \in \mathcal{P}}$ is a covering of $\mathcal{P}$:

Mixture manifold 2

- If $p \in \mathcal{M}$, then $\mathcal{U}_p \subset {}^*\mathcal{U}_p$.
- If $p_1, p_2 \in \mathcal{E}(p)$, then ${}^*\mathcal{U}_{p_1} = {}^*\mathcal{U}_{p_2}$.

D For each $p \in \mathcal{M}$, we define ${}^*\mathcal{E}\left(p\right) \subset \mathcal{P}$ as

$$ {}^*\mathcal{E}\left(p\right) = \left\{ q \in \mathcal{P} : \frac{q}{p} \in L^{\Psi}\left(p\right) \right\} $$

Th The set of charts

$$ \left\{ \left({}^*\mathcal{U}_q, \eta_q \right) : q \in \mathcal{E}\left(p\right) \right\} $$

is an affine $\mathcal{C}^\infty$-atlas on ${}^*\mathcal{E}\left(p\right)$, so it has the structure of a manifold modeled on the Banach space $L_0^{\Psi}(p)$.

- For each pair $p_1, p_2 \in \mathcal{E}\left(p\right)$ the transition map is

$$ \eta_{p_2} \circ \eta_{p_1}^{-1} : \begin{cases} L_0^{\psi} p_1 \to L_0^{\psi} p_2 \\ \qquad u \mapsto u \dfrac{p_1}{p_2} + \dfrac{p_1}{p_2} - 1 \end{cases} $$

Mixture manifold 3

Th Let $p \in \mathcal{M}_\mu$ be given. For each $q \in \mathcal{P}$, the divergence $D(\tilde{q}\|p)$ of the probability density $\tilde{q}$ with respect to p is definite if and only if $\tilde{q} \in {}^*\mathcal{U}_p$:

$$D(\tilde{q}\|p) = \mathrm{E}_p\left[\frac{\tilde{q}}{p}\log\left(\frac{\tilde{q}}{p}\right)\right] < \infty \quad \Leftrightarrow \quad q \in {}^*\mathcal{U}_p$$

where $\tilde{q} := |q| / \int |q|\, d\mu$.

Th For each density $p \in \mathcal{M}_\mu$, the inclusion $j : \mathcal{E}(p) \hookrightarrow {}^*\mathcal{E}(p)$ is of class $\mathcal{C}^\infty$.

- More work is needed to fully clarify the basic structure of the mixture manifold.
- For a different construction, including all α-connections, cfr. Gibilisco & Pistone (1998).

Exponential models

- Let V be a closed subspace of L_0^ϕ. We call exponential model based on V

$$\mathcal{E}_V(p) = \left\{ \mathrm{e}^{u - K_p(u)} \,|\, u \in V \right\}$$

- Let

$$V^\perp = \left\{ v \in L_0^\Psi(p) \,|\, \mathrm{E}_p(vu) = 0, v \in V \right\}$$

be the orthogonal space of V. Then

$$q \in \mathcal{E}_V(p) \quad \text{iff} \quad \mathrm{E}_p\left(v \log \frac{q}{p} \right) = 0, v \in V^\perp$$

- Note that $V + V^\perp$ in not a splitting of the model space $L_0^\phi(p)$. In fact, the proper notion of statistical model appears to be different from what is technically termed a sub-manifold, because there is no orthogonal splitting of subspaces in $L^\Phi(p)$. The proper spitting consists of the orthogonal space of the tangent space of the model in the pre-dual space $L_0^\Psi(p)$.

Girsanov

- The Information Geometry of a sample space $(X, \mathcal{X}, \mu)$ has no relation with any structure of the sample space itself. This is a limitation of the theory, because it makes no use of the specific sample space, e.g. Gaussian, Stationary, Bernoulli, Wiener . . .
- For each $q \in \mathcal{E}(p)$ let us call **Girsanov transformation** a measurable mapping $T_q : X \to X$ such that $T_q^*(q \cdot \mu) = p \cdot \mu$. Such transformations, when available, give a **state space representation** of statistical models and could inherit the geometry of the statistical models.
- An example are the affine function transforming a general multivariate normal to the multivariate standard distribution.
- In a Wiener space, the Girsanov density

$$\exp\left(\int_0^T b(x(t))dx(t) - \frac{1}{2}\int_0^T b^2(x(t))dt\right)$$

is an exponential model and is associated with the transformation

$$x(\cdot) \mapsto x(\cdot) - \int_0^{\cdot} b(x(t))dt$$

cfr. the poster by D. Imparato in this Symposium.

Approximation and algebra

- The idea to approximate complex models in the numerical sense, not the asymptotic sense, has no unique solution, because it depends upon the divergence or topology we use to compute the approximation error.
- In applications of Girsanov transformations to Finance Mathematics, it has been suggested to use Wiener-Hermite expansions of the exponent. This does not apply to our framework, because there is no general convergence of such expansions in $L^{\phi}(p)$.
- Let us assure that there exist an increasing sequence finite partitions Δ_n, $n \in \mathbb{N}$ that generate the σ-algebra $\mathcal{X}$. For each n, let us consider a vector basis $T_{n,j}$ of the ring of Δ_n-measurable functions. We assume that all the elements of such a basis are integer valued.
- Consider the following approximation scheme:
  1. For $u \in L^{\Phi}(p)$, $u_n = \mathrm{E}_p\left(u|\Delta_n\right) \to u$ as $n \to \infty$.
  2. Approximate $u \in L_0^{\Psi}(q)$ with $u_n - \mathrm{E}_q\left(u_n\right)$.
- Then, **there exist a way to approximate a general $\mathcal{M}$ with sub-models which are essentially defined on a finite state space**

Manifold vs Variety 1

A factorial finite sample space (a *design*) is a set-product of finite spaces. Using integer coding of *levels*, we consider finite sets of the form

$$F \subseteq D = \times_{i=1}^{d} \{1, \ldots, n_i\}, \quad \text{for example} \quad F = \begin{array}{cccc} \bullet & \bullet & \bullet & \bullet \\ \bullet & + & + & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{array} \subseteq D = \begin{array}{cccc} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{array}$$

If we perform repeated sampling N times, the sample space will be $F_N = F^N \ni \omega = (\omega_1, \ldots, \omega_N)$. If the sampling of $a \in F$ is performed with probability p_a, then

$$p(\omega) = \prod_{i=1}^{N} p_{\omega_i} = \prod_{a \in D} p_a^{N_a(\omega)}$$

where $[N_a(\omega)]_{a \in F}$ is the table of counts observed in the sample. In the following we assume $N = 1$ and $F = \Omega$.

Manifold vs Variety 2

If $p(a) > 0$, $a \in F$, $\#\Omega = n$, and the reference probability is uniform, then the chart formula gives

$$u(\omega) = \log\left(\frac{p(\omega)}{1}\right) - E_0\left(\log\left(\frac{p(\cdot)}{1}\right)\right) = \sum_{a \in F} \log p(a) \left(N_a(\omega) - \frac{1}{n}\right)$$

Let B_0 be the space of centered random variables for the uniform probability (contrasts). Let V_0 be a linear subspace of B_0 with a basis $T_1, \dots, T_k$, and let $V_1 = V_0^\perp$ be the orthogonal space. Then, an exponential model is specified either by

$$u(\omega; \theta_j : j = 1 \dots k) = \sum_{a \in F} \log p(a; \theta_j : j = 1 \dots k) \left(N_a(\omega) - \frac{1}{n}\right) = \sum_{j=1}^{k} \theta_j T_j(\omega)$$

or by

$$\sum_{\omega \in \Omega} v(\omega) \sum_{a \in F} \log p(a; \theta) \left(N_a(\omega) - \frac{1}{n}\right) = \sum_{a \in F} \log p(a; \theta) \sum_{\omega \in \Omega} v(\omega) \left(N_a(\omega) - \frac{1}{n}\right)$$

$$= 0 \qquad v \in V_1$$

Manifold vs Variety 3

If $V_1 = \operatorname{span}(v_1, \dots, v_h)$, $h + k = n - 1$, and

$$U = \left[\sum_{\omega \in F} v_j(\omega)\left(N_a(\omega) - \frac{1}{n}\right)\right]_{\substack{a \in F \\ j=1\dots h}} = \left[\sum_{\omega \in F} v_j(\omega) N_a(\omega)\right]_{\substack{a \in F \\ j=1\dots h}}$$

then

$$\sum_{a \in F} U_{aj} \log p(a) = \log\left(\prod_{a \in F} p(a)^{U_{aj}}\right) = 0 \qquad j = 1 \dots h$$

or

$$\prod_{a \in F} p(a)^{U_{aj}} = 1 \qquad j = 1 \dots h$$

or, taking the positive and negative part of $U = U^+ - U^-$,

$$\boxed{\prod_{a \in F} p(a)^{U^+_{aj}} - \prod_{a \in F} p(a)^{U^-_{aj}} = 0} \qquad j = 1 \dots h \tag{1}$$

If U takes values in the integers $\mathbb{Z}$, then (1 is a binomial in the ring $\mathbb{Q}[p(a) : a \in F]$.

Exponential vs Algebraic

The **explicit form**

$$u(\omega;\theta) = \sum_{a\in F} \log p(a) \left(N_a(\omega) - \frac{1}{n}\right) = \sum_{j=1}^k \theta_j T_j(\omega)$$

leads to the usual writing of the exponential model: here $N_a(\omega) = 1(a=\omega)$, then

$$\begin{aligned} \log p(\omega;\theta) &= \frac{1}{n}\sum_{a\in F} \log p(a;\theta) + \sum_{j=1}^k \theta_j T_j(\omega) \\ &= \sum_{j=1}^k \theta_j T_j(\omega) - \Psi(\theta_1,\dots,\theta_k) \end{aligned} \qquad (2)$$

The **implicit form**

$$\prod_{a\in F} p(a)^{U^+_{aj}} - \prod_{a\in F} p(a)^{U^-_{aj}} = 0 \qquad j = 1\dots h$$

leads to a model which is **not** restricted to positive probabilities. We call this model **the algebraic model of (2)**. **The positive part of the algebraic model is the exponential model**.

Extended Exponential

We call **extended exponential** the exponential model and its limit points.

The extended exponential model is contained in the algebraic model, possibly strictly.

Example

Consider $F = \{0,1\}^2$ and the log-linear model $\log p(x,y) = \theta_1 x + \theta_2 y - \psi(\theta_1,\theta_2)$. The matrix U has to be orthogonal to the column space of $Z = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$, then $U = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$ and the algebraic model is

$$p(0,0)p(1,1) - p(0,1)p(1,0) = 0$$

A limit probability can be zero only if either x or y is equal 1. The the case $p(0,+) = 0$ is in the algebraic model, but it is not a limit of the exponential model.

Commutative Algebra 1

The term **Algebra** in the title is ambiguous: here we mean **Computational Commutative Algebra** or **CCA** or CoCoA. The subject is relatively new, being about 20 years old.

What is CCA is explained in many good textbooks and software manuals:

- D. A. Cox, J. B. Little, and D. O'Shea. *Ideal, Varieties, and Algorithms*. Springer-Verlag, New York, 2nd edition, 1997. 1st ed. 1992;
- M. Kreuzer and L. Robbiano. *Computational Commutative Algebra 1*. Springer, Berlin-Heidelberg, 2000;
- CoCoATeam. CoCoA: a system for doing Computations in Commutative Algebra. Available at `http://cocoa.dima.unige.it`, no date.

The use of CCA as a tool in Statistics has been advocated for the first time in a pre-print dated 1993, published much later as P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 1998.

Commutative Algebra 2

The use of CCA as a tool in statistical design theory (DOE) has been suggested first in

- G. Pistone and H. P. Wynn. Generalised confounding with Gröbner bases. *Biometrika*, 83(3):653–666, Mar. 1996;
- E. M. Riccomagno. *Algebraic Geometry in Experimental Design and Related Fields*. Phd thesis, Department of Statistics, University of Warwick, August 1997.

It consists of the description of finite sets of points in an affine space k^d, where k is a computable number field, usually $k = \mathbb{Q}$, as the set of zeroes of a system of polynomial equations. Such a system is then processed through a CCA software to solve typical identifiability problems.

A presentation of this theory is contained in Chapters 1-3 of

- G. Pistone, E. Riccomagno, and H. P. Wynn. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman&Hall, 2001,

Reference to relevant literature in the last 5 years is contained in a recent talk by Maria-Piera Rogantin at the ICODOE 2005 Memphis Conference.

Computational Commutative Algebra

CoCoA

is a CCA software developed at the University of Genova by a team coordinated by Lorenzo Robbiano.

A number of other CCA systems is available:

Maple `http://www.maplesoft.com`,

Mathematica `http://www.wolfram.com`,

Singular `http://www.singular.uni-kl.de` ....

We use in particular CoCoA and Maple for general purpose algebraic computations. In special cases we use

R `http://www.r-project.org` for computations oriented to statistics;

4ti2 `http://www.4ti2.de` for special computations needed for combinatorial problems.

A CCA software makes exact computations on number fields e.g. $\mathbb{Q}$, $\mathbb{Z}_p$ ..., on rings of polynomials e.g. $k[x, y, z]$, on ideals e.g. $I = \langle g_j \rangle$, $I + J$, IJ, quotient rings e.g. $\frac{k[x,y,z]}{\langle x-y, y-z \rangle}$.

Computing the quotient ring

Given a ring $R = k[x_1 \dots x_d]$ and an ideal I, the **quotient ring** $\frac{R}{I}$ is, in particular, a k-vector space. We exploit frequently these facts:

- The quotient ring $\frac{R}{I}$ has (at least) one linear basis E of monomials $x_1^{\alpha_1} \cdots x_d^{\alpha_d} = x^\alpha$, $\alpha \in L$.
- This basis is hierarchical, i.e. if $x^\alpha \in E$ and x^β divides x^α, then $x^\beta \in E$. In turn, this is true if its set of *logarithms* $\alpha \in L$ is the complement of a lattice.
- A set of generators of such a lattice is computed by the CCA systems. The key word is **Gröbner basis**.
- If the zero-set of the ideal I is finite and has n distinct points, then $\#L = n$ and vice-versa. In this sense, a CCA system is able to compute the number of distinct solution.

Toric statistical models

An important class of models is both m-algebraic and e-algebraic.

Let be given k linearly independent and integer-valued random variables, e.g. monomial functions on a grid of integers.

If $B = [x^\alpha : \alpha \in L]$ is saturated monomial basis, consider a subset $M \subseteq L$, $1 \in M$, and the model exponential model

$$p(x,\psi) = \exp\left(\sum_{\beta \in M} \psi_\beta x^\beta\right) = \prod_{\beta \in M} \zeta_\beta^{x^\beta}$$

where $\zeta_\beta = \exp \psi_\beta$. By dropping ζ_0 we can write a **monomial parametric model for the unnormalized probabilities** q:

$$q(x,\psi) = \prod_{\beta \in M_0} \zeta_\beta^{x^\beta}$$

where $M_0 = M \setminus 1$, $\psi = (\psi_\beta : \beta \in M_0)$, $\zeta_\beta \geq 0$, $\beta \in M_0$.

Such a model is called (by algebrists) **Toric**.

Term ordering

- We call *monomial* a polynomial of of the ring $k[x]$, $x = x_1, \dots, x_s$, with one term. If ax^α is a monomial, we call

$$x^\alpha = x_1^{\alpha_1} \cdots x_s^{\alpha_s}$$

  its **term**, or power-product.
- Let T^s be the set of all terms, possibly identified with the set of **logarithms** $\alpha \in \mathbb{Z}^s_{\geq 0}$.
- Note that univariate polynomials are linear combinations of univariate terms x^n, which are ordered by their degree. All computations for one dimensional polynomials exploit this fact.
- In more than one dimension it is necessary to introduce the concept of a term-ordering to order terms.
- Note that the terms of are naturally pre-ordered according to simplification of terms. For example $x_1^2 x_3$ precedes $x_1^3 x_3^2$ as the "fraction" $\frac{x_1^3 x_3^2}{x_1^2 x_3} = x_1 x_3$ is still in T^s and $(2,1) \leq (3,2)$ component-wise.

Term ordering: definition

A **monomial** or **term-ordering** on the polynomial ring $k[x]$ is a total and **well ordering relation** $\succ_\tau$ (or τ or $\succ$) on T^s, that is on the terms of $k[x]$, such that

1. $x^\alpha \succ 1$ for all x^α with $\alpha \neq 0$ and
2. for all $\alpha, \beta, \gamma \in \mathbb{Z}_+^s$ such that $x^\alpha \succ x^\beta$, then $x^\alpha x^\gamma \succ x^\beta x^\gamma$.

Note that the restriction of a term-ordering to the terms of the type x_i gives an initial ordering of the indeterminates $x_1, \dots, x_s$.

- Any two terms are comparable, that is for any x^α, x^β either $x^\alpha \succ x^\beta$ or $x^\alpha = x^\beta$ or $x^\beta \succ x^\alpha$. This property characterizes total orderings.
- Given a ring of polynomials and a term ordering, $k[x]$ and τ, each polynomial is identified with an **ordered list** of elements of the number field (the coefficients) and of logarithms (the support).
- There is no infinite descending chain, that is any subset of terms contains a minimum element with respect to the ordering. This property is known as well-ordering.
- The ordering is compatible with the simplification of terms, that is for any pair of terms x^α and x^β, if x^α divides x^β then $x^\beta \succ x^\alpha$.

Term Orderings: Lexicographic

There are two basic term-orderings: lexicographic and degree-reverse-lexicographic. In CoCoA they are called `Lex` and `DegRevLex`. In Maple `lex` and `tdeg`.

The **lexicographic** term-ordering is defined as the order for which $x^\alpha \succ_{\texttt{lex}} x^\beta$ if, with reference to the $\log$ representation in the vector $\alpha - \beta$, **the left-most nonzero entry is positive**. That is

$$x^\alpha \succ x^\beta \quad \text{if and only if} \quad \begin{cases} \alpha_1 > \beta_1 \\ \text{or there exist } p \leq s \text{ such that} \\ \alpha_i = \beta_i \text{ for } i = 1, \dots, p-1 \text{ and } \alpha_p > \beta_p \end{cases}$$

For example, in $\mathbb{Q}[x_1, x_2, x_3, x_4]$ the initial ordering is $x_1 \succ x_2 \succ x_3 \succ x_4$.
The square-free terms in increasing lexicographic order are

$$1, x_4, x_3, x_3x_4, x_2, \\ x_2x_4, x_2x_3, x_2x_3x_4, x_1, x_1x_4, x_1x_3, x_1x_3x_4, x_1x_2, \\ x_1x_2x_4, x_1x_2x_3, x_1x_2x_3x_4$$

In a lexicographic ordering an indeterminate dominates over the others.

Gröbner basis

The **Hilbert basis theorem** states that any ideal is finitely generated, even if the generating set is not necessarily unique. Some bases are **special**.

Definition Let τ be a term-ordering on $k[x]$. A subset $G = \{g_1, \dots, g_t\}$ of an ideal I is a **Gröbner basis** of I with respect to τ if and only if

$$\langle \mathrm{LT}_\tau(g_1), \dots, \mathrm{LT}_\tau(g_t) \rangle = \langle \mathrm{LT}_\tau(I) \rangle$$

where $\mathrm{LT}_\tau(I) = \{\mathrm{LT}_\tau(f) : f \in I\}$.

In general the following inclusion holds

$$\langle \mathrm{LT}(g_1), \dots, \mathrm{LT}(g_t) \rangle \subseteq \langle \mathrm{LT}(I) \rangle$$

and unless $\{g_1, \dots, g_t\}$ is a Gröbner basis, the inclusion may be strict. Indeed in $\langle x_1^3 - 2x_1x_2, x_1^2x_2 - 2x_2^2 + x_1 \rangle \subset \mathbb{Q}[x_1, x_2]$ with the `tdeg`($x_1 \succ x_2$) ordering we have that $x_1^2 \in \langle \mathrm{LT}(I) \rangle$ but $x_1^2 \notin \langle \mathrm{LT}(x_1^3 - 2x_1x_2), \mathrm{LT}(x_1^2x_2 - 2x_2^2 + x_1) \rangle = \langle x_1^3, x_1^2x_2 \rangle$.

Toric vs Algebraic

A more statistical name than "Toric" could be "Generalized Multinomial" GMN.

The toric model is located between the exponential model and the algebraic model. Moreover, we can show it is a special algebraic model.

The elimination in the toric model of the ζ's indeterminates leads to a **toric variety**; with properly computed degrees, all polynomials in generator's sets are homogeneous binomial.

CoCoA has special functions `Elim` and `Toric` to perform this task.

`4ti2` has a special executables called `groebner` and `markov` to perform similar tasks.

Some times the algebraic and the toric coincides. We see that this is related with the existence of structural zeros compatible with the model. For a discussion with examples, see

F. Rapallo. Toric statistical models: Parametric and binomial representations. Technical report, Dipartimento di Matematica. Università di Genova, 2004. Submitted

From algebraic to maximal GMN

The procedure from toric to algebraic could be reversed, producing a **maximal** GMN such that

- The elimination or Toric procedures lead to the algebraic model.
- All compatible structural zeros are included in the model;
- All embedded exponential model are parameterized.

This is research work in progress to be published 2006. See related work in

- D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *The Annals of Statistics*, 29, 2001
- D. Geiger, C. Meek, B. Sturmfels *On the toric algebra of graphical models* to appear in *Annals of Statistics*