-Optimum design and related statistical issues

May 26–28 2005. Juan-les-Pins

Causality and toric models

Giovanni Pistone
Politecnico of Turin, Italy giovanni.pistone@polito.it

with Eva Riccomagno and Henry P. Wynn

Introduction

Consider the following simple model on binary variables:

$$\begin{array}{ccc} \text{Carcinogenic genotype } U & \longrightarrow & Y \text{ Lung cancer} \\ \downarrow & & \uparrow \\ \text{Smoking } X & \longrightarrow & Z \text{ tar deposit on lungs} \end{array}$$

U is unobservable, X, Y, Z are observable.

The question is: Is cigarettes' smoking a **cause** of lung cancer?

The checking of

$$\mathsf{P}\{Y = 1|X = 1\} > \mathsf{P}\{Y = 1|X = 0\}$$

does not really answer the question. The present talk is devoted to a preliminary discussion of the **geometry** of statistical models based on this DAG. Most of the material I am discussing is taken from various authors, but no proper references are given now. Particular thanks are due to Eva Riccomagno, Jim Smith and Henry Wynn for explaining to me some of the key concepts.

Philosophy

The discussion of causes of events, manipulability of reality, and of responsibility post-hoc, is a major (unsolved) philosophical question for mankind. A small sample of quotations:

- LORD God said: *"Have you eaten from the tree of which I commanded you not to eat?" The man said, "The woman whom you gave to be with me, she gave me fruit from the tree, and I ate."* Genesis 2:11b-12. Adam is suggesting the model

$$\text{you} \longrightarrow \text{she} \longrightarrow I$$

- *As to the frequent use of the words, Force, Power, Energy, &c.*, *which every where occur in common conversation, as well as in philosophy; that is no proof . . . with the connecting principle between cause and effect, or can account ultimately for the production of one thing to another. These words, as commonly used, have very loose meanings annexed to them; and their ideas are very uncertain and confused*. D. Hume 1748 EHU. David Hume suggests to avoid any mechanical explanation and to consider correlations only. This is the modern way.

- Post-modern thinkers try to associate the idea of cause with the idea of manipulability.

DAG's and $\perp\!\!\!\perp$

The previous DAG, together with the ordering

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

encodes the factorization

$$p(u,x,z,y) = p(u)p(x|u)p(z|x)p(y|u,z)$$

which, in turn, is equivalent to the following two statements of conditional independence

$$U \perp\!\!\!\perp Z | X \quad X \perp\!\!\!\perp Y | U, Z$$

Proof 1. Assume the conditional independence. Then

$$p(u,x,z,y) = p(u)p(x|u)p(z|u,x)p(y|u,x,z) = p(u)p(x|u)p(z|x)p(y|u,z)$$

2. Assume the factorization. Then

$$p(z|u,x) = \frac{p(u,x,z,+)}{p(u,x,+,+)} = p(z|x)$$

$$p(y|u,x,z) = \frac{p(u,x,z,y)}{p(u,x,z,+)} = p(y|u,x)$$

Intervention or experiment

There are at least two possible approaches to the observation of the world in the perspective of drawing conclusions about cause-effect.

Passive Assume we sample from the population described by our model for the variables U, X, Z, Y by stratifying along the values of the observable variable X. In the stratum $\{X = 0\}$, i.e. the smokers, the model will be the conditional probability

$$p(u, z, y | X = 0) = \begin{cases} \dfrac{p(u, 1, z, y)}{p(u, +, z, y)} & \text{on } \{X = 0\} \\ 0 & \text{on } \{X = 1\} \end{cases}$$

Active Assume we sample from the whole population, but now we force the sampled individuals to stop smoking. According some authors, the intervention hides the influence of U on Z, and the model becomes

$$p(u, z, y \| X = 0) = \begin{cases} p(u)p(z|0)p(y|u, z) & \text{on } \{X = 0\} \\ 0 & \text{on } \{X = 1\} \end{cases}$$

which correspond, on the restricted set $\{X = 0\}$, to the new DAG

$$U \longrightarrow Y \longleftarrow Z$$

Discussion of the factorization

- $p(u, x, z) = p(u, x, z, +) = p(u)p(x|u)p(z|x)$ is a DAG model, actually a Markov Chain.

- $p(x, z, y) = p(+, x, z, y) = \sum_u p(u)p(x|u)p(z|x)p(y|u, z)$ appears not to be ad DAG model.

- The discussion of the correlation between X and Y is based on $p(x, y) = p(+, x, +, y)$ and the algebra is not nice. Non-correlation and independence is the same in our example, because we deal with binary variables.

- It has been rediscovered recently that two discrete random variables X and Y are independent if and only if the 2-way table of probabilities has all its 2×2-minors equal to zero:

$$p_{xy}p_{x'y'} - p_{xy'}p_{x'y} = 0 \quad x \neq x', y \neq y'$$

- The intervention is a mapping from a factorization to a factorization. The new density is dominated by the original one and the operation on the densities is multiplication by a factor.

DAG's and quadrics

Assume for a moment that U, X, Z, Y are the canonical random variables on the sample space $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3 \times \Omega_4$, with $\#\Omega = N = N_1 N_2 N_3 N_4$. Then the conditional independence statements are equivalent to the following system of quadratic equations

$$U \perp\!\!\!\perp Z | X \begin{cases} p(u, x, z, +)p(u', x, z', +) - p(u, x, z', +)p(u', x, z, +) = 0 \\ u, u' \in \Omega_1, x \in \Omega_2, z, z' \in \Omega_3 \quad u \neq u' \text{ and } z \neq z' \end{cases}$$

$$X \perp\!\!\!\perp Y | U, Z \begin{cases} p(u, x, z, y)p(u, x', z, y') - p(u, x, z, y')p(u, x', z, y') = 0 \\ u \in \Omega_1, x, x' \in \Omega_2, z \in \Omega_3, y, y' \in \Omega_4 \quad x \neq x' \text{ and } y \neq y' \end{cases}$$

The total number of such equations is

$$\left(\binom{N_1}{2} N_2 \binom{N_3}{2}\right) + \left(N_1 \binom{N_2}{2} N_3 \binom{N_4}{2}\right)$$

Considering algebraic dependencies, the number of sufficient equations can be smaller when the number of states is greater than 2. Note that the equations are homogeneous of degree 2. The equations of the second group are binomial i.e. consists of two monomial.

DAG's and quadrics: binary case

If $\Omega_i = \{0, 1\}$, then the equations are

$$U \perp\!\!\!\perp Z | X \begin{cases} p(0,0,0,+)p(1,0,1,+) - p(0,0,1,+)p(1,0,0,+) = 0 \\ p(0,1,0,+)p(1,1,1,+) - p(0,1,1,+)p(1,1,0,+) = 0 \end{cases}$$

$$X \perp\!\!\!\perp Y | U, Z \begin{cases} p(0,0,0,0)p(0,1,0,1) - p(0,0,0,1)p(0,1,0,0) = 0 \\ p(0,0,1,0)p(0,1,1,1) - p(0,0,1,1)p(0,1,1,0) = 0 \\ p(1,0,0,0)p(1,1,0,1) - p(1,0,0,1)p(1,1,0,0) = 0 \\ p(1,0,1,0)p(1,1,1,1) - p(1,0,1,1)p(1,1,1,0) = 0 \end{cases}$$

The total number of such equations is

$$\left(\binom{2}{2} 2 \binom{2}{2}\right) + \left(2 \binom{2}{2} 2 \binom{2}{2}\right) = 6$$

and the number of degrees of freedom is $\mathrm{df} = 15 - 6 = 9$.

Intervention and quadrics

The conditional independence model after the intervention is

$$U \perp\!\!\!\perp Z | Y \begin{cases} p(0,0,0,0)p(1,0,1,0) - p(0,0,1,0)p(1,0,0,0) = 0 \\ p(0,0,0,1)p(1,0,1,1) - p(0,0,1,1)p(1,0,0,1) = 0 \end{cases}$$

$$\{X = 1\} \begin{cases} p(u,1,z,y) = 0 \\ \text{for } u,z,y = 0,1 \end{cases}$$

I cannot see any relation between this system and the previous one.

DAG's and algebraic varieties

In the ring of polynomials with rational coefficients and indeterminates $p(u, x, z, y)$, $u, x, z, y \in \{0, 1\}$ the previous equations define a zero-set (algebraic variety). If we add the conditions

$$\begin{cases} p(+,+,+,+) - 1 = 0 \\ \qquad p(u, x, z, y) \geq 0 \quad u, x, z, y \in \{0, 1\} \end{cases} \quad \underline{\geq}!$$

we get a description of the model as a manifold with borders.

The null hypothesis $X \perp\!\!\!\perp Y$ is obtained adding the equation

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

Apart from the inequalities, all the equations can be studied with the help a Computer Algebra software, such as Singular, CoCoA .... Substitutions and eliminations can be done on the machine to ease the algebraic computations.

Information geometry

The study of the geometry of statistical models that started in the 40's with Rao's paper on the interpretation of Fisher Information as a metric tensor. An important further step was done by Efron in the 70's by introducing the geometry of (curved)-exponential models. Amari in the 80' introduced a bigger picture including all previous results in a unified theory he called IG.

Let $(\Omega, \mathcal{F}, \mu)$ (no assumptions) a probability space. We denote by $\mathcal{P}_+$ the set of all μ-a.s. positive probability densities w.r.t μ. It is possible to show that there are two atlas of charts, each chart associated to a $p \in \mathcal{P}_+$, both defining on $\mathcal{P}_+$ a manifold structure together with its connections. The charts are

1. $q \mapsto u = \ln \frac{q}{p} - \mathbb{E}_p\left(\ln \frac{q}{p}\right)$
2. $q \mapsto \frac{q}{p} - 1$

The first atlas defines the exponential geometry, the second atlas defines the mixture geometry.

Exponential models

An exponential model is a family of densities in $\mathcal{P}_+$ of the form

$$q \propto \exp(u) \quad u \in V$$

where V is a linear subspace of μ-centered and exponentially integrable random variables. Introducing the normalization constant, we write

$$q = \exp(u - K(u)) \quad u \in V$$

where K is the cumulant functional. In the non parametric case, i.e. V if infinite dimensional, the cumulant functional has all the usual properties, e.g.

- K is convex and analytic.
- $\operatorname{grad} K(u) \cdot v = \mathbb{E}_q(v)$.
- $\operatorname{hess} K(u) \cdot v \circ w = \operatorname{Cov}_q(v, w)$.
- The Jensen conjugate of K is the entropy of q.

Toric models

Let Ω be a finite sample space, $n = \#\Omega$. If we consider an exponential model were the (finite dimensional) space V is spanned by integer valued random variables $T_j : \Omega \to \mathbb{Z}$, $j = 1, \ldots, l$, we have

$$p(\omega) \propto \exp\left(\sum_{j=0}^{l} \psi_j T_j(\omega)\right) \tag{1}$$

As in this case the sufficient statistics T_j are integer valued, the model can be given, by the parameters change

$$\mathrm{e}^{\psi_j} = \zeta_j, \tag{2}$$

the following form, that we call toric.

$$p(\omega) \propto \prod_{j=0}^{l} \zeta_j^{T_j(\omega)} \tag{3}$$

This name comes from Commutative Algebra. A statistical name could be Generalized Multinomial. The exponential model and the toric models are not exactly the same, because the toric model could have zero probabilities. Precisely, if $\zeta_j = 0$, the $T_j(\omega) \neq 0$ implies zero probability at ω.

Exponential vs toric vs algebraic

We can prove prove that

- *The exponential model is the strictly positive part of the toric model.*

- *Let* $T = [T_1 \cdots T_l]$*, let* $u_1 \cdots u_k$ *be a integer valued vectors forming a linear basis of* $\ker T^t$*, and let* $u = u^+ - u^-$*. Then the set of probabilities described by the quadrics*

$$\prod_\omega p(\omega)^{u^+(\omega)} - \prod_\omega p(\omega)^{u^-(\omega)} = 0$$

  is the closure of the exponential model. Let us call this model the algebraic model.

- *There exist a parameterization of the algebraic model as a toric model.*

- *Toric models are "union" of exponential models with variable support.*

The exponential model of the DAG

The DAG model

$$p(u, x, z, y) = p(u)p(x|u)p(z|x)p(y|u, z)$$

is associated with the additive decomposition (the log-linear model)

$$\ln p(u, x, z, y) = \ln p(u) + \ln p(x|u) + \ln p(z|x) + \ln p(y|u, z)$$

and the linear spaces of random variables involved are those generated by

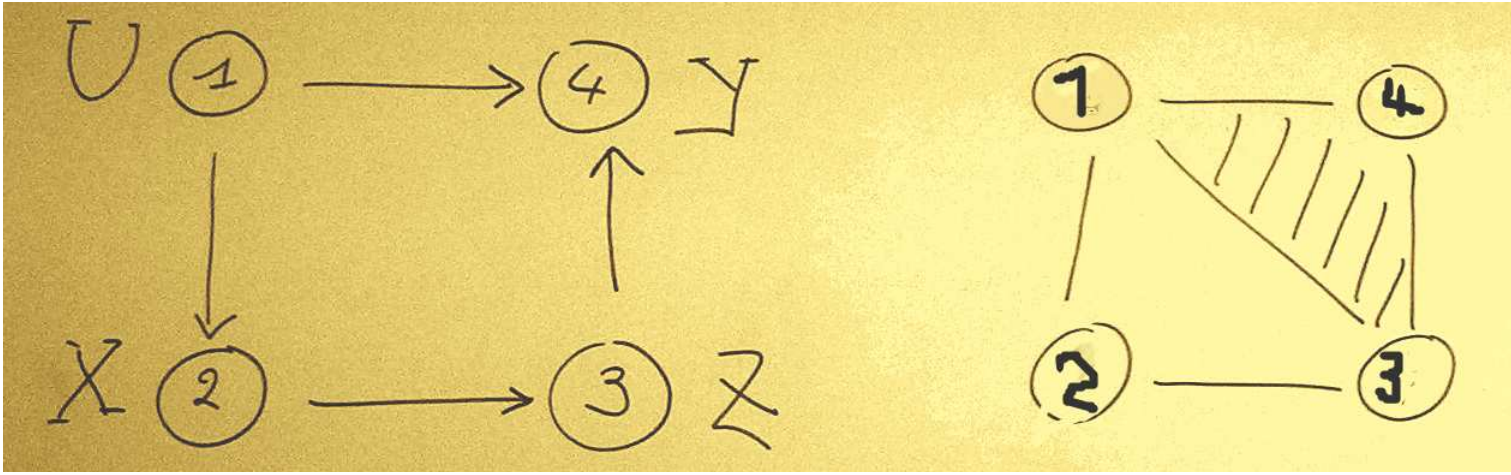
$$\{U\}, \{U, X\}, \{X, Z\}, \{U, Z, Y\}$$

On finite sample space each of these spaces of random variables has a finite basis as vector spaces, e.g. the indicator functions of points. The dimension of the space in the minimal exponential model containing our DAG model is

$$\mathrm{df} = (2 + 3 + 3 + 3) - 1 = 10$$

i.e. in the DAG model one degree of freedom is missing. The DAG model is a curved exponential model

Simplicial complexes



The best way to represent the fact that our DAG model is not an exponential model is the comparison between the two graphical representation. In the directed graph representation one arrow, i.e. one interaction is missing.

The use of simplicial complexes is even more useful in case we use monomial bases as bases of the vector space in the exponential model. The picture is a representation of the monomial basis

$$1, u, x, z, y, ux, xz, uy, zy, uzy$$

Probabilities in the mixture geometry

The mixture geometry is the geometry where the parameters are the raw probabilities, or any linear transformation od the probabilities. In the mixture geometry the geodesic between two probabilities is the mixture model. In the exponential geometry the geodesic is a 1-dimensional exponential model connecting two probabilities (the Hellinger arc).

In the mixture geometry it is highly convenient to change the coding $(0,1) \mapsto (-1.1)$. In such a coding, the generic probability has the form

$$\begin{aligned} p(u,x,z,y) = 1 + & \\ & \mu_1 u + \mu_2 x + \mu_3 z + \mu_4 y + \\ & \mu_{12} ux + \mu_{13} uz + \mu_{14} uy + \mu_{23} xz + \mu_{24} xy + \mu_{34} zy + \\ & \mu_{123} uxz + \mu_{124} uxy + \mu_{134} uzy + \mu_{234} xzy + \\ & \mu_{1234} uxzy \end{aligned}$$

where μ_α is the α-moment in the uniform probability. The generic conditional probability has the form

$$p(x|u) = 1 + (a_0 + a_1 u)x$$

so that, in principle, the DAG model can be written in polynomial form.

Discussion

- Statistical models have a very rich mathematical structure, and many alternative frameworks are available: algebraic geometry, differential geometry . . .

- The geometry of models used in modelling causal effect is of interest from the conceptual point of view.

- Computational Algebra may help in small-to-medium problems.

- The differential geometric methods should suggest efficient numerical methods.