Causality

Jean-Marc Azaïs, Université de Toulouse

April 12, 2012

According to Platon "everything that is born is born necessarily a from a cause". This view of a single universal causality seems to us poorly adapted to the statistical approach. In most cases, a complex phenomenum appears only by the conjunction of several causes, some are controllable and some are not (e. g. environmental variables). The statistician will search partial cause only. In this search is **confounding** will be his worst enemy. Confounding happens when different causes may have the same consequences and it is impossible to separate their influence.

Example 1 *A famous exemple : Death penalty Verdict by defendant's race . We consider the verdicts for murder in Florida over the period 1973-1979 (Agresti, 2010, Table 3.1).*

Over this period 4764 murders have been judged that we have dispatched following the race of the defendant

	Sentence		
	death	*other*	*%*
black	*59*	*2448*	*2.4*
white	*72*	*2185*	*3.2*

Let us analyze this table: we see a greater severity (almost significant) of jurors against whites. This does not correspond to what one has heard of the southern states of the United States in the 70's. The anti-white racism was not very developed. In the vocabulary that we will develop below it is not **consistent** *with the context .*

Drawn from this data set, the conclusion that in Florida white be exposed to greater severity of juries would be a hasty conclusion.

One way to go further in analyzing the phenomenon is to look at the race of the victim, giving

		Sentence		
victim	*defebdant*	*death*	*other*	*%*
black	*black*	*11*	*2209*	*.5*
black	*white*	*0*	*111*	*0*
white	*black*	*48*	*239*	*16.7*
white	*white*	*72*	*2074*	*3.4*

Now the interpretation is clear: what is expensive it is to kill a white. This interpretation is consistent with the context and moreover changes are much more pronounced

than in the previous table. This suggests strongly that this is the right interpretation, see Section 2

The exemple above is an example of confounding : there is a (partial) confounding between " being white" and " killing a white"".

We see clearly that conclusion drawn without precautions on data with confounding can be totally wrong.

1 Some definitions

Causality First we have to define causality. The most convenient way is trough random assignment. The following is borrowed from Van der Vaart

"If individuals are randomly assigned to a treatment and a control group, and the groups differ significantly after treatment, Then the treatment causes the difference. "

This definition is clearly related to the classical "treatment to control randomized experiment" that is routinely used is medicine. In the vocabulary of design of experiments it is a completely randomized design in which $2n$ units (rats, persons, ETC...) are used, n of which at random receive the treatment at the n others the control. The judgment of significativity is conducted by a T test (comparison of two means).

Observational data and interventional data

We define observational data as data that are collected by pure observation without any influence. This is the case, for example, in many of the economics data. But the example above (treatment versus control comparison) involves an intervention since there is a random choice that infers on the experiment.

Conditional effect and causal effect

This clasical example is due to Pearl : suppose I have a lawn in my garden without watering device. I can see an important link between the two events $W=$ "the lawn is wet" and $R=$"it rains". So in the distribution of probability observed the conditional probability

$$\mathbb{P}(R|W) \text{ is high.}$$

How do I know that if I water my lawn, it will of course not cause rain?

A more realistic example : meta-genome

Human beings have a genome, i.e. the genetics composition of their cells. But they host, mainly in their alimentary canal, a lot of micro-organisms the genome of which is very stable during our life. The genome of these micro-organisms, that can be sequenced, is called the meta-genome. It is a hot topic in medicine. The meta-genome is suspected to play an important role in the development of diseases, in particular in the evolution to obesity and to Type II diabetes (fat one). From observational data, it is easy to predict the evolution to these diseases but it is much more difficult, because of all we have seen, to predict the evolution of the person the meta-genome of which has been artificially modified. The reason in that this intervention creates an individual that does not belongs more to the observed population.

This has a lot of application in economics. If a government starts a new policy, for example putting a cap on the exchange rate with Euro, it created in some sense a completely new economics situation and which is very difficult to predict from observational data.

2 Hill Causality

Austin Bradford Hill was the scientific who co-discovered (Doll and Hill ,1954) the effect of tobacco on cancer. Obviously it was from observational data an if was even more difficult because of the influence of tobacco trusts. The definition of Hill is made on common sense observations .Basically the causality will be proved by an accumulation of presumptions. The methodology developed by Hill as become a standard in the definition of causality in epidemiology.

Basically the different criterions are the following

- Strength : for example lung cancer is observed to be 9 times more frequent for smokers than for non-smokers.
- Consistency : is the phenomenon repeatable ?This does not imply that one is necessarily in a lab experiment repeatable to infinity. Repeatability in epidemiology is observed for example when we observe the same phenomenon in several social groups, countries, continents, etc. ...
- Specificity: the link is found between the two variables and there is no connection with other variables. This criterion is poorly adapted to the study infectious diseases that can have a multiplicity of symptoms: there is causality without specificity.
- Temporality: -which is the cart and which the horse?
- biological gradient, for exemple dose response curves
- plausibility, coherence and analogy. This is a analysis of coherence with the context. Of course this is not an absolute criterion unless no unexpected discovery is possible.

Theses criterions are each a component of causality.

3 Granger Causality

This concept was developed in econometrics to study links between time series. For example: you will have an influence the outside temperature on the electrical consumption? Its strength and weakness lies in the need for many regularly spaced observations of our variables in a more or less stationary framework

The principle is to help the time to distinguish cause from consequence. As Granger himself said in the article of Scholarpedia: "The definition leans heavily on the idea that the cause occurs before the effect, which is the basis of most, but not all, causality definition".

Suppose we want to study the influence of the series X_t on the series Y_t. X_t can be for example the annual GNP of a country and Y_t the investment rate.

As a precaution we include in our model an additional series Z_t which is, in the general case multivariate, representing all variables other than X_t likely to have an influence on Y_t. It is in this research, of course impossible to satisfy exactly in practice that lies the main weakness of the method.

The simplest model proposed is the " simple causal model. "This is an Auto-regressive model which in practice must be not too long. More precisely if we assume for simplicity that Z_t is scalar, we set

$$\begin{aligned} X_t &= a_1(B)X_t + b_1(B)Y_t + c_1(B)Z_t + \epsilon_{1,t} \\ Y_t &= a_2(B)X_t + b_2(B)Y_t + c_2(B)Z_t + \epsilon_{2,t} \\ Z_t &= a_3(B)X_t + b_3(B)Y_t + c_3(B)Z_t + \epsilon_{3,t}, \end{aligned}$$

where a_i, b_i, c_i $i = 1, \ldots, 3$ are polynomials in the backward operator (Backward, B) which are of degree p and contain no term of degree zero. For example

$$a_1(B)X_t = a_1^1 X_{t-1} + \cdots + a_1^p X_{t-p}.$$

The strong assumption is the causal hypothesis of the AR model. This means that we assume that the vector ϵ_t is an innovation, and that the model above is correct:

- the dependence must be linear
- p the length of the auto-regression should be correct
- no relevant variables must be forgotten.

The Granger test is actually defined as the test of nullity of the polynomial $a_2(B)$ that links the present of Y_t to the past of X_t.

This model has a generalized classical model with " Instantaneous causality "in which, for example, Y_t may depend on the values of X_t or Z_t at the same time.

In this model we can arrive at a finding a " feedback "in which both variables inter act mutually without possibility of knowing in which way the causality is exercised.

In conclusion, the Granger causality is a standard reference to define causality. The longer the series the better the method. But it does not provide absolute protection against confouding . For example, one can construct counterexamples in which the influence of X on Y is confounded with the einfluence of some unobserved variable Z which in facts affects Y.

4 Intervention calculus and directed acyclic graphs

A way to represent a distribution and causal relation is to use a graphs for example the graph

$$R \to W$$

and

$$W \to R$$

can both represent the link between the variable R:"rain" and W " the lawn is wet "(supposed to be quantitative). Of course only the first is realistic Suppose that all our variables are jointly Gaussian. In that case all the relations between these variables are linear. In the first case, we write

$$R = \sigma_R Z_1 \quad W = \rho R + \sigma_W Z_2$$

where Z_1, Z_2 are two independent standard Gaussian variables.

The intervention calculus is baed on fixing some variables a computing the consequence on the variables that are placed after in the graph.

For example if by visiting a shaman we are able to impose rain : $R = 1$ then the distribution of the wetness of the lawn receives a shift of ρ. But in the contrary if we water the lawn $W = 1$ nothing happens.

This asymmetry of causality , in a case where the joint distribution can be symmetric (if we tune conveniently the parameters) clearly shows that the intervention calculus is distinct from the conditional calulus.

The Graph is directed in the sense that edges are oriented to indicate Causality. And to avoid pathology it must contain no cycle. In that sense it is acyclic.

The major problem is that in general, as in the example of the lawn, several Directed Acyclic Graphs DAG can be associated to the same probability distribution.

But under some hypothesis and using some clever algorithms it is possible to list all the possible DAG and to construct a envelope of the possible influence of some variable over an other.

This is done in a high dimension sparse model.

5 Bibliography

Lok ,Gill, van der Vaart, Robins. (2004) Estimating the causal effect of a time-varying treatment on time to event using structural nested failure time models.Statistica Neerlandica

Heckman. Econometric Causality (2008). Discussion paper 3525 IZA

C. W. J. Granger Investigating Causal Relations by Econometric Models and Cross-spectral Methods Econometrica, Vol. 37, No. 3. (Aug., 1969), pp. 424-438.

Maathuis, Kalisch & Bhlmann Estimating high-dimensional intervention effects from observational data. The Annals of Statistics. 2009, Vol. 37, No. 6A, 3133?3164 DOI: 10.1214/09-AOS685

Maathuis, Diego Colombo, Kalisch Bhlmann Predicting causal effects in large-scale systems from observational data. Nature Methods 7,4 ,247-248.