UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 08-17
Business Economics Series 02
April 2008, 25

Departamento de Economía de la Empresa
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249607

A VALID THEORY ON PROBABILISTIC CAUSATION*

Jose M. Vidal-Sanz[1]

Abstract

In this paper several definitions of probabilistic causation are considered, and their main drawbacks discussed. Current notions of probabilistic causality have symmetry limitations (e.g. correlation and statistical dependence are symmetric notions). To avoid the symmetry problem, non-reciprocal causality is often defined in terms of dynamic asymmetry. But these notions are likely to consider spurious regularities. In this paper we present a definition of causality that does not have symmetry inconsistencies. It is a natural extension of propositional causality in formal logics, and it can be easily analyzed with statistical inference. The modeling problems are also discussed using empirical processes.

***Keywords**:* Primary 62M30; secondary 62M15, 62G20.
***AMS 2000 subject classifications**:* Causality, Empirical Processes and Classification Theory.

* This research has been supported by the Spanish Ministry of Education and Science (research project SEJ2007-65897), and the Spanish local government of Comunidad de Madrid (research project S-0505/TIC-0230).

[1] E-mail: jvidal@emp.uc3m.es

INTRODUCTION

The aim of science is first to determine whether a set of axiomatic events or propositions can be accepted as true, and then to derive the validity of more complex facts establishing causality relationships. As Wold (1954) pointed out: "The concept of causality is indispensable and fundamental to all sciences". But these axiomatic events and causal implications can be deterministic (invariable) regularities or be defined in terms of indeterministic (probabilistic) regularities. Probabilistic causality is a difficult topic, involving the controversial issues of certainty (irrevocability) of a cause-effect relationship, and the connection of causality with theories of induction.

In "formal logic" (in the deterministic context), a proposition A causes B if when A is true then B is also true, and we denote this fact by $A \Rightarrow B$. Another expression for "A causes B" is that B is necessary for A. It is also said that both propositions are equivalent when $A \Rightarrow B$ and also $B \Rightarrow A$, and this is denoted by $A \Leftrightarrow B$. It means that, by definition, causality is an asymmetric concept, and symmetry results into equivalence. Earman (1986) provides an introduction to determinism in physics. A drawback of this causality definition is that it cannot be applied to indeterministic contexts. Bertrand Russel (1913, 1948) criticised the deterministic concept of causation. He argues that the world is complex, and even though causal laws might hold true, they often fail because of preventing circumstances, and the fact that it is impractical to bring in innumerable "unless" clauses. But, in spite of the high complexity in the world, there are also causal lines of quasi-permanence that warrant our inferences. For instance, the statement "smoke causes lung cancer" is false when the logical definition is used, but we can find empirical regularities suggesting this effect, see e.g. Suppes (1970). Causality relationships are usually due to a constellation of factors that are jointly sufficient for B, where A is a relevant causal factor but not sufficient to ensure B. The complexity of the surrounding factors can be addressed by probability theory. For a discussion of indeterminism and causation see Humpreys (1989).

According to the Stanford Encyclopedia of Philosophy, "Probabilistic Causation" designates a group of philosophical theories that aim to characterize the relationship between cause and effect using the tools of probability theory. The central idea behind these theories is that causes raise the probabilities of their effects, *ceteris paribus*. This idea has been formalized in a variety of definitions, but often they are affected by symmetry (e.g., a high correlation between two random variables is a symmetric notion). To reduce the symmetry problem, some of these causality notions have been modified introducing conditional probabilities to some specific event. But these refinements are

not entirely satisfactory. Mackie (1974), Lewis (1986), and Hausman (1998) discuss different issues about asymmetry of causation.

To avoid symmetry, some philosophers have defined causality from a dynamic perspective , arguing that causes are invariably followed by their effects and therefore causation should be considered in terms of stable patterns of succession (see e.g. David Hume, 1748, section VII.) Following Hume, some proponents of probabilistic theories of causation have identified causal direction with temporal direction. But, a definition of causality based on dynamic regularities is likely to consider spurious regularities. For example, we observe that lightning is often followed by thunder, but the first does not cause the second, both are simultaneously caused by the same electric phenomena. However, even using dynamic specifications we can find vicious circularities (symmetries) that have to be avoided by ruling out the possibility of backwards-in-time causation a priori. Even without backwards-in-time causation, spurious associations are relevant (e.g., the sustained increment in atmospheric carbon dioxide during the last century correlates with posterior earth average warming which is often interpreted as a causal effect and explained using a greenhouse analogy, whereas for other authors it is a spurious association and the global warming trend marks the arrival of a glaciation period). Besides, the logical notion $A \Rightarrow B$ does not require the use of time, and we should require the same generality for any valid definition of probabilistic causation.

Overall, the current definitions of probabilistic causation can be objected, because they lack a sound logical basis, and/or involve symmetry (e.g., statistical dependence), and/or time-delay requirements which limit the general applicability. Probabilistic causality is still an undefined concept. However, it has a crucial relevance for scientists. The falsificationist theory of Karl Popper considers that a physical theory can be falsified if it can be rejected based on contradiction with empirical observation. In a deterministic context a simple counterexample can be used to reject a causal theory, but not in the context of probabilistic causality. Without a valid theory of probabilistic causality, empirical induction cannot give scientists clear rules to reject causal relationships.

This paper proposes a probabilistic theory of causation which does not suffer from any of these drawbacks. The rest of the paper is organized as follows. In section 2 the previous developments on probabilistic causation are discussed. In section 3, I propose a new concept of probabilistic causation, based on logic that is not affected by symmetry or time delay requirements, and consider asymptotic forms of causation. Section 4 introduces the empirical analysis of probabilistic causation. In Section 5 I present central concepts relating modelling causal analysis and statistical inference. In the concluding remarks section, I discuss how causal relationships can be used for optimal decision

making.

LITERATURE REVIEW

The idea behind the probabilistic causation theories is that, *ceteris paribus* causes raise the probabilities of their effects. This basic idea is generally formalized using conditional probabilities. Let $(\Omega, \mathcal{F}, P)$ be a probability space. For any sets $A, B \in \mathcal{F}$ with $P(B) > 0$, the conditional probability is defined as $P(A|B) = P(A \cap B)/P(B)$. Here we consider the probability function as a mathematical object satisfying the Kolmororov axioms. It might be interpreted as a personal degree of belief (often based on previous empirical analysis), or as a propensity law of physical events (Popper, 1983) that can be estimated using frequency limits (as considered by von Mises 1939), but the specific interpretation has little relevance for our purposes (for a discussion see, e.g. Dawid, 2004).

The first definition of probabilistic causation, known as "Probability-Raising" (PR), considers that A causes B, where $0 < P(A) < 1$, if and only if

$$P(B|A) > P(B|A^c) \tag{1}$$

where A^c denotes the complement of the set A, see Suppes (1970). It is equivalent to $P(B|A) > P(B)$[1], and also to the property of positive statistical dependence $P(A \cap B) > P(A)P(B)$. If the last inequality is reversed the events show negative statistical dependence, and if an equality holds the events are statistically independent. Clearly, the PR notion of causality introduces some flexibility on the deterministic formulations, but it is perfectly symmetric (if A causes B in a PR sense, then also B causes A in a PR sense). In other words, this formulation is closer to equivalence rather than causation.

Furthermore, PR is sensitive to spurious causality. If A and B are both caused by some third factor $C \in \mathcal{F}$, then it may be that $P(B|A) > P(B|A^c)$ even though B does not cause A, since A and B are positive dependent and simultaneously caused by C. Hans Reichenbach (1956) suggested the idea of "screening off" to apply to a particular type of probabilistic relationship. Given three sets $A, B, C \in \mathcal{F}$ and $P(A \cap C) > 0$, then C is said to screen A off from B if

$$P(B|A \cap C) = P(B|C)$$

[1] $P(B|A) > P(B|A^c) \Leftrightarrow P(B|A)P(A^c) > P(B|A^c)P(A^c) \Leftrightarrow$
$P(B|A)(1 - P(A)) > P(B|A^c)P(A^c) \Leftrightarrow P(B|A) > P(B|A)P(A) + P(B|A^c)P(A^c) = P(B)$

which means that C renders B irrelevant to A in probabilistic terms. (In fact, this is equivalent to "conditional statistical independence" $P(A \cap B|C) = P(A|C)\,P(B|C)$ between A and B). Using this notion, we can define the 'no screening off' probabilistic causation, using the probability raise condition $P(B|A) > P(B|A^c)$ together with the condition that there is no C that screens A off from B.

But the 'no screening off' condition is not sufficient to solve the spurious causality problem, since it just eliminates events C that conditionally make A and B statistically independent. But there might be other events C such that $P(A \cap B|C) > P(A|C)\,P(B|C)$ (or the opposite) that cause both A and B. Due to the "Simpson's Paradox," we may have that $P(B|A) > P(B|A^c)$ with probabilistic inequalities reversals $P(B|A \cap C) < P(B|A^c \cap C)$ and $P(B|A \cap C^c) < P(B|A^c \cap C^c)$ for some C.

Alternatively, other authors have considered causality conditional to some specific situation, to express that A raises the probability of B under some specific event $C \in \mathcal{F}$, called "test situation", with $P(A \cap C) > 0$ so that $P(B|A \cap C) > P(B|A^c \cap C)$. Often, this condition is required for all the event C in a class of test situations. For a review of this literature, see Cartwright (1979), Skyrms (1980), Eells (1991, chapters 2, 3, and 4) and Hitchcock (1993). But the definition of test situations introduces substantial complexity, and it does not solve the symmetry problem.

All the considered refinements of PR causation are based on (1), and therefore are affected by symmetry. More complex formulations of causality have been proposed to introduce asymmetries with little success, see e.g. Reichenbach (1956), Price (1991), Arntzenius (1993), Papineau (1993), and Hausman (1998). Other authors suggest that the necessary asymmetry is provided by our perspective as agents, see e.g. Price (1991). For instance, simple regression analysis is often interpreted in terms of causality by some researchers, even though correlation is a bi-directional relationship.

A useful argument in logic to establish causal relationships is the *"reductio absurdum"* equivalence that states: $A \Rightarrow B$ if and only if $no - B \Rightarrow no - A$. Some authors consider the "counterfactual causality", where causality is defined using the idea that A causes B if the probability that A does not occur is higher with B than it would be if B had not happened, i.e.

$$P(A^c|B) > P(A^c|B^c)$$

or equivalently $P(A|B^c) < P(A|B)$, for a review see Lewis (1986), Noordhof (1999), and Kvart (1997). But probabilistic counterfactual causality is actually equivalent to positive dependence[2], and therefore to PR causality, so that it is still affected by symmetry.

[2] $P(A|B^c) < P(A|B) \Leftrightarrow P(A|B^c)\,P(B^c) < P(A|B)\,P(B^c) \Leftrightarrow P(A|B^c)\,P(B^c) < P(A|B)\,(1 - P(B)) \Leftrightarrow$
$P(A|B^c)\,P(B^c) + P(A|B)\,P(B) < P(A|B) \Leftrightarrow P(A) < P(A|B) \Leftrightarrow P(A \cap B) > P(A)\,P(B)$

Alternative notions of probabilistic causality have been considered in the statistical literature, usually introducing random variables. In this setting, a measurable event A with $0 < P(A) < 1$ causes a random vector Y if the conditional probabilities satisfy that

$$P(Y \in B|A) \neq P(Y \in B|A^c) \tag{2}$$

for some measurable set B. Testing non-causality means using data to decide if the null hypothesis $H_0 : \sup_y |F(y|A) - F(y|A^c)| = 0$ is true, where $F(y)$ is the cumulative probability distribution of Y. In practice, a weaker property is usually studied, such as the homogeneity of conditional means $H_0 : E(Y|A) = E(Y|A^c)$, i.e.

$$H_0 : E(Y|d = 1) - E(Y|d = 0) = 0, \tag{3}$$

where $d = I(A)$ and $I(A)$ is the indicator function of the set A. In empirical applications, the null hypothesis can be tested using ANOVA and MANOVA methods. If additional regressors Z are included, the null hypothesis $H_0 : P(E(Y|d = 1, Z) = E(Y|d = 0, Z)) = 1$ can be tested with ANACOVA methods under linearity. As a consequence, (static) non-causality is often discussed using heterogeneity of linear regression models. In experimental settings, the regressors are dummy variables associated to some treatment, and the regression coefficients are interpreted as potential effects on the dependent variable. The significant heterogeneity of coefficients in the regression model is interpreted as a proof of causation, moving interpretations from association (correlation or statistical dependence) to causation. For an introduction to experimental design see Cochran and Cox (1957). The experimental analysis is the cornerstone of manipulability theories of causation considered by Gasking (1955), Collingwood (1940), von Wright (1971), and Menzies and Price (1993), and Sobel (1998). Pearl (1995) considers graphic structures to depict these causal relationships. More in general, a linear regression relationship between two random vectors X, Y is a model for linear association $Y = \Pi X + v$. Although the variables Y are often called endogenous, they are chosen by the researcher who could similarly regress X with respect to Y. Correlation, means dependence, and statistical dependence are essentially concepts of symmetric association and do not constitute a valid setting for causal analysis.

Some authors use even more dubious notions of probabilistic causality based on Simultaneous Equation Models (SEM), also known as structural models. These models were developed by econometricians in the 1940s (driven by the "Cowless Commission for research in Economics" and building on Haavelmo's work), to estimate the parameters in economic models whose equilibrium is determined by a system of equations conditional on exogenous elements. In essence, these models specify

a covariance structure. Structural models and the related path analysis, are nowadays commonly used to justify causality claims in social sciences, see e.g. Sobel (1995). SEM postulated relationships between variables in a vector $Y \in \mathbb{R}^G$ conditionally on certain environmental variables $X \in \mathbb{R}^K$, assuming a parametric relationship $g(Y, X, \theta_0) = \varepsilon$ where some assumptions are considered about the conditional distribution of $\varepsilon|X$, typically $E[\varepsilon|X] = 0$ and $Var[\varepsilon|X] = \Sigma$. For example, the linear SEM is given by the linear system

$$BY + CX = \varepsilon.$$

A structural model is well defined only if there is a locally unique relationship that can be regarded as an inverse of the model $Y = f(X, \varepsilon, \beta_0)$ known as the reduced form, and θ_0 can be obtained from β_0. In the linear SEM, for example, if $rank(B) = G$ ($\Leftrightarrow \det(B) \neq 0$) we can write the model as $Y = \Pi X + v$, where $\Pi = -B^{-1}C$ and $v = B^{-1}\varepsilon$ so that $\Omega = E[v_t v_t'] = B^{-1}\Sigma B^{-1\prime}$. This model is known as the reduced form associated with the structural form.

Typically the square matrix B is normalized to have ones in the main diagonal, and the other coefficients in $BY+CX = \varepsilon$ are often (wrongly) interpreted in terms of static causality. For example, if B is a triangular matrix some authors interpret the relationship as a chain of successive effects. A path analysis figure is often drawn depicting the linear SEM structure. This figure displays the variables Y, X and a flow diagram of causal effects in the form of arrows connecting variables related by non-null coefficients in B, C. I denote $B_{i,j}$ the coefficient (i, j) in the matrix B and similarly for $C_{i,k}$ in C, then if $B_{i,j} \neq 0$ and $B_{j,i} \neq 0$ a double headed arrow connect Y_i and Y_j, replaced by a single headed arrow if one of these coefficients is zero, and if $C_{i,k} \neq 0$ a single headed arrow follows from X_k to Y_i. The use of graphical models for drawing causal relationships, based on ideas of Sewal Wright back in the 1920s, is extensive in social sciences after the work of Dudley Duncan (1966), see Whittaker (1990).

Structural models are not identified. Given any multiequation regression model, $Y = \Pi X + v$, with $E[Xv] = 0$ and $E[vv'] = \Omega$, if we multiply the model by an arbitrary non-singular matrix B we obtain $BY + CX = \varepsilon$ with $\Sigma = B\Omega B'$ and $\Pi = -B^{-1}C$. Since there are infinitely many regular matrices B, we can obtain infinitely many structural forms associated to a multiequation regression model. From another perspective, if we take an arbitrary factorization $\Omega = \Psi\Sigma\Psi'$ where Ψ has complete rank and Σ is symmetric and positive definite (e.g. we can use a Choleski decomposition, obtaining a lower triangular structural form – a procedure popularized by Chirstopher A. Sims, and/or combine it with a permutation matrix to reorder the path arrows), then multiplying the

reduced form by $B = \Psi^{-1}$ we can define a structural form whose perturbations have covariance matrix Σ . Therefore, the concept of structural form has little meaning from a statistical point of view.

To solve the ambiguity of structural models, it is necessary to assume some identification constraints $W(B, C, \Sigma) = 0$, which are an arbitrary choice of the researcher based on his theoretical dogma. The SEM is identified if the system $W(B, C, \Sigma) = 0$, $B\Pi + C = 0$, $\Omega = B^{-1}\Sigma B^{-1\prime}$ has a unique solution $\{B, C, \Sigma\}$. The output is somewhat arbitrary. For instance, we can identify the structural form $\{B, C, \Sigma\}$ by requiring that $B_{i,j} = 0$, or a reciprocal effect $B_{j,i} = 0$, or even $B_{i,j} = -B_{j,i}$ meaning that both variables are affected by identical opposite structural relationships. Any choice has nothing to do with the causal laws in nature. Several authors (e.g., Cliff 1983, Holland 1988) have criticized the use of structural models to infer causation. Simultaneous equation models are artificial structures identified from the reduced form regression. Yet, we observe too often that research articles using regression models are rejected in social sciences because endogeneity was not taken into account. Structural forms can be a convenient method for adjusting theoretical models (with static-interactions) to empirical data, provided that the model is rich enough to ensure the identification of the parameters (setting additional constraints). But we cannot test if the model "is correct", nor can we interpret the coefficients in B, C (nor the coefficient signs) in terms of "causality". The ambiguities are even bigger with nonlinear structural models, where identification is often a local concept.

As if the subjectivity of SEMs was not important enough, some researchers (considering high dimensional vectors X,Y) combine ideas from structural models and factor analysis, see e.g. Jöreskog (1973, 1978) and Bollen (1989). In particular, factor structural models, consider that $B\eta + C\xi = \varepsilon$, where η and ξ are unobserved latent variables called "constructs", which are given by the factor model $Y = \Lambda_Y \eta + \varepsilon$, $X = \Lambda_X \xi + \delta$. Factor structural models are extremely ambiguous, usually identified by setting some rotation for the latent factors along with the structural identification assumptions. The resulting output is so ambiguous for the practitioner, that any result can be obtained, provided that we devote enough time to estimate under different identification schemes. Yet, causal relationships conclusions drawn from these models pervade some social sciences, enhanced by computer-friendly software such as LISREL, EQS or AMOS.

Some authors have proposed probabilistic theories of causation based on the idea that causes usually precede their effects in time, and the notion of dynamic causality can be applied only to cause-effect relationships that take place along some time horizon. But this idea in unrelated with

the causality concept in formal logic. Suppes (1970, chapter 2) and Eells (1991, chapter 5) define causal asymmetry in terms of temporal asymmetry, using time as a kind of test situation. Dynamic causality has been also approached by models based on automated procedures, see Spirtes, Glymour and Scheinnes (2000), Scheines (1997), Hausmand and Woodward (1999), and also Pearl (2000).

In the statistical literature, the simplest notion of dynamic causality can be considered in time series models where past is considered cause for the present. Assume that the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$ has a spectral density satisfying $\int_{\Pi^d} \log f(\lambda)\, d\lambda > -\infty$, $\int_{\Pi^d} f(\lambda)^{-1}\, d\lambda < \infty$. There are infinitely many factorizations of the spectral density,

$$f(\lambda) = (2\pi)^{-1} \sigma^2 |A(\lambda)|^{-1}$$

where $A(\lambda)$ is normalized with $a_0 = \int_{\Pi^d} A(\lambda)\, d\lambda = 1$. We usually identify a particular factor by setting some coefficients to zero. In particular, $A(\lambda)$ is known as causal representation when it has Fourier expansion with real coefficients $a_k = \int_{\Pi^d} A(\lambda)\, e^{ik'\lambda} d\lambda = 0$ for $k > 0$, and anticausal if $a_k = 0$ for $k < 0$. Each of these factorizations leads to linear autoregressive models expressed with respect to the past or the future of $\{Y_t\}$ respectively, as suggested by Wold (1954) and Wiener (1956). This notion of past autoregressive modelling must be interpreted in terms of predictability rather than causality. It is just an arbitrary interpretation of symmetric autocorrelations from one side, as a tool to identify time series models. Actually, it is not clear that past is a cause for future, particularly when we consider certain abstract phenomena in astrophysics and Quantum physics, where we can find paradoxes easily. For example it is commonly accepted that the Big-Bang is the cause for the universe dynamic expansion, but according to the current physical theories, time did not exist when the Big-Bang occurred (the time-dimension was caused by it), so we cannot consider the Big-Bang a "dynamic cause" of the universe expansion because it is a vicious circular definition.

The econometric literature has considered more complex notions of dynamic causality. Using ideas of stochastic processes, assume that a sequence of measurements $\{X_t\}$ of the "causal" phenomenon are regularly taken along the time horizon, and measures $\{Y_t\}$ of the "effect" phenomenon are similarly taken. Then, causality can be interpreted using recursive arguments: we say that there is dynamic causality if for all time periods, (1) the measurements of the "effect" Y_t are statistically dependent on lagged measures of the causal variable X_{t-1}, X_{t-2} conditionally on its own lags, and (2) there is no reverse symmetric relationship (i.e., measurements of the "cause" X_t are statistically independent from lags of the effect variable $Y_{t-1}, Y_{t-2}, ..$ conditionally on its own lags). This idea was put across by Granger (1969), who built on earlier work in statistics literature (Wold, 1954

and Wiener 1956), and it is nowadays known as "Granger's causality", for a review see Engle et al. (1983) and Geweke (1984). Granger's causality is essentially a notion of crossed predictability between two time series, and it is also close to vicious circularity: in order to assess whether X_t causes Y_t, we would already need to know whether Y_t causes X_t. Granger's causality has considerable operational significance in empirical analysis. Unfortunately, a definition of causality based on dynamic regularities is likely to consider spurious regularities when both $\{X_t, Y_t\}$ are caused by Z_t, but X_t shows the influence of Z_t before Y_t.

The problem of providing a valid probabilistic causation theory is still open, and there is no general definition avoiding symmetries and time conditional requirements yet. Many authors are uncomfortable when thinking about the notion of causality in an uncertainty context. In part, as Geweke (1984) points out, because "the idea is notoriously difficult to formalize, as casual reading in the philosophy of science will attest". Statisticians have essentially abandoned the quest for a valid concept of probabilistic causation. In the next section, a probabilistic causality notion is presented that avoids all the symmetries and conflicts inherent in previous definitions.

VALID PROBABILISTIC CAUSATION

Let Ω be the universal set, and $\mathcal{F}$ a σ-algebra of events that can be asserted as true or false with some probability. Using set theory, for any sets $A, B \subset \Omega$ we can also express the fact $A \Rightarrow B$ through the expression $A \subset B$, meaning that if any $\omega \in A$ occurs then also $\omega \in B$ is satisfied. Since we can also express the fact $A \Rightarrow B$ through the expression $A \subset B$, then it is clear that $P(A) \leq P(B)$, but the reciprocal is not true and that is the reason for failure in PR definitions of probabilistic causality. By contrast, I will introduce a concept $A \Rightarrow B$ almost sure, using the idea that $A \subset B$ almost sure with respect to a probability function P, i.e. the probability of $\omega \in A$ which are not in B have zero probability. Clearly, $A \subset B$ if and only if $A \cap B = A$, and if and only if $A \cap B^c = \varnothing$ where $B^c = \{\omega \in \Omega : \omega \notin B\}$ denotes the complement of set B. Therefore a valid definition of almost sure causation is that:

Definition almost sure causality. Given a probability space $(\Omega, \mathcal{F}, P)$, for any sets $A, B \in \mathcal{F}$ we say that A causes B almost surely if $A \cap B^c$ has null probability, i.e.

$$P(A \cap B^c) = 0.$$

I denote almost sure causality by $A \overset{a.s.[P]}{\Rightarrow} B$. If $A, B \subset \Omega$ are not measurables, the definition can be extended using $P^*(A \cap B^c) = 0$, where P^* denotes the outer probability.

I am not requiring A to be a minimal cause for B, and it could have some irrelevant components. I say that A is minimal cause if $A \cap B^c = \varnothing$, meaning that there is not any non empty subset in A that does not belong to B (note that I simply require this set to have a zero probability measure). This concept is compatible with the intuition underlying propositional implications.

The proposed definition is an asymmetric definition. If $P(A \cap B^c) = 0$ and $P(A^c \cap B) = 0$, then $A \cup B = A \cap B$ except for a set of probability null, meaning that both concepts A and B are essentially the same $P(\{A \cup B\} \setminus \{A \cap B\}) = 0$ (i.e. both are equal to the intersection $A \cap B$, and all other components are negligible in probability terms). It means that both events are almost surely equivalent, $P(A = B) = 1$.

Clearly, for any measurable set A, it is satisfied that $A \overset{a.s.[P]}{\Rightarrow} \Omega$. We can consider the *"reductio absurdum"* equivalence ($A \Rightarrow B$ if and only if $B^c \Rightarrow A^c$). The same property is satisfied in almost sure causality, where $A \overset{a.s.[P]}{\Rightarrow} B$ if and only if $B^c \overset{a.s.[P]}{\Rightarrow} A^c$ since $P(B^c \cap (A^c)^c) = 0$. This is a form of "conterfactual analysis".

Note that $A = \varnothing$ implies any set $B \in \mathcal{F}$, since $P(A \cap B^c) = P(\varnothing) = 0$; although in general we will consider non empty sets A with $P(A) > 0$. If $B = \Omega$, then for any set $A \in \mathcal{F}$, since $P(A \cap B^c) = P(\varnothing) = 0$, but in general we will consider sets $B \subsetneq \Omega$ and with $P(B) < 1$. If A and B^c are statistically independent, then $P(A \cap B^c) = P(A) P(B^c)$ and causality means that A has zero probability or B has probability one, so that the relationship is meaningless.

The next table shows all the causal events that can be established between A and B,

	A	A^c
B	$A \cap B$	$A^c \cap B$
B^c	$A \cap B^c$	$A^c \cap B^c$

Then, we have the following possible causal probabilistic relationships:

Strict causality

$A \overset{a.s.[P]}{\Rightarrow} B,\ B \overset{a.s.[P]}{\not\Rightarrow} A$

P	A	A^c
B	$\pi \in [0,1)$	$1-\pi$
B^c	0	1

equivalence

$A \overset{a.s.[P]}{\Rightarrow} B,\ B \overset{a.s.[P]}{\Rightarrow} A$

P	A	A^c
B	1	0
B^c	0	1

Let $I(A) = I(A)(\omega)$ denote the indicator function for the set $A \in \mathcal{F}$ (i.e., $I(A) = 1$ if $\omega \in A$ and $I(A) = 0$ otherwise). Then, we can express $P(A \cap B^c) = E[I(A \cap B^c)]$, where any of the following

expressions can be considered:

$$\begin{aligned} I(A \cap B^c) &= I(A) \cdot I(B^c) = I(A) \cdot (1 - I(B)) = I(A) - I(A \cap B), \\ I(A \cap B^c) &= \min\{I(A), I(B^c)\} = \min\{I(A), 1 - I(B)\}, \\ I(A \cap B^c) &= \max\{0, (I(A) - I(B))\} := (I(A) - I(B))^+, \end{aligned}$$

all of which can be proved considering combinations of indicator values giving a value of 1. To assess almost sure causality in empirical context, we can replace the expectations $E[I(A \cap B^c)]$ by averages of observed events ω.

A smaller σ-algebra $\mathcal{A} \subset \mathcal{F}$ can cause almost surely the event $B \in \mathcal{F}$, if $E[Z \cdot I(B^c)] = 0$ for any bounded $\mathcal{A}$-measurable random variable Z (actually, it is enough to check it for all Z indicator functions of events in $\mathcal{A}$). Then we say $\mathcal{A} \overset{a.s.[P]}{\Rightarrow} B$.

Often, causal relationships are based on random variables. The introduction of random variables is useful because then we can relate probabilistic causality and empirical inferences. Most of the Physical laws, can be expressed in terms of systems of equations $B = \{f(X) = \varepsilon\}$, or inequality systems such as $B = \{f(X) \leq 0\}$.

Equations and Causality: Let us consider a measure space $(\Xi, \mathcal{B})$, and X a measurable measurable application $X : \Omega \to \Xi$, then $(\Omega, \mathcal{F}, P)$ induces a probability law $P_X = P \circ X^{-1}$ on $(\Xi, \mathcal{B})$. In particular, if we consider the Borel euclidean space $\left(\mathbb{R}^d, \mathbb{B}^d\right)$ then a measurable application $X : \Omega \to \mathbb{R}^d$ is a random vector. Given two events $\alpha, \beta \in \mathcal{B}$ and setting

$$A = \{\omega \in \Omega : X(\omega) \in \alpha\}, \qquad B = \{\omega \in \Omega : X(\omega) \in \beta\}$$

then $A \overset{a.s.[P]}{\Rightarrow} B$ (i.e. $E[I(X \in \alpha) \cdot (1 - I(X \in \beta))] = 0$) is trivially equivalent to $\alpha \overset{a.s.[P_X]}{\Rightarrow} \beta$. This approach is particularly relevant for empirical purposes. In particular, considering the Borel euclidean space, we can consider theories expressed by systems of equations such as $A = \{g(X) = 0\}$ and $B = \{g(X) = 0\}$; or alternatively by inequalities $A = \{g(X) \leq 0\}$, and $B = \{f(X) \leq 0\}$.

A variety of cases can be studied here, for example the experimental design framework. Assume that the event A denotes a specific treatment that can be set by the researcher. Then, we can consider if B is satisfied when the treatment A is applied, and study the causal hypothesis $E[I(A \cap B^c)] = 0$. In particular, we can consider an event defined by a conditional expectation such as $B = \{E[Y|Z] = 0\}$, and the null causal hypothesis is

$$E[I(A) - I(A \cap \{E[Y|Z] = 0\})] = 0.$$

Note that the classical experimental design framework considers if $E[Y|Z,A] = E[Y|Z,A^c]$ a.s., which is a relevant but different problem.

In most empirical applications, we assume that the probability function of X is unknown but we observe a sample of data identically distributed as the random vector X. Alternatively, some random variables could be observed but others could be regarded as latent variables or random shocks (about which probability distribution we have some information). For instance, consider events given by

$$A = \{\omega \in \Omega : g(X) \geq \varepsilon\}, \qquad B = \{\omega \in \Omega : f(X) \geq \epsilon\}, \tag{4}$$

where (ε, ϵ) are unobserved random variables, and define $H(\varepsilon, \epsilon|x)$ as the conditional cumulative probability distribution of $(\varepsilon, \epsilon)|X = x$. Applying the law of iterated expectations, the causality statement $A \overset{a.s.[P]}{\Rightarrow} B$ can be expressed by the identity,

$$\begin{aligned} E[E[I(\varepsilon \leq g(X))|X] - E[I(\varepsilon \leq g(X)) I(\epsilon \leq f(X))|X]] &= 0 \Leftrightarrow \\ E[H(g(X)|X) - H(g(X), f(X)|X)] &= 0. \end{aligned}$$

were $H(\varepsilon|x) = \lim_{\epsilon \to \infty} H(\varepsilon, \epsilon|x)$ is the marginal conditional distribution of $\varepsilon|X = x$. The distribution H can be specified by a parametric model or be unknown (e.g., in a semiparametric setup).

The notion of probabilistic causality can be extended to the dynamic framework. Let us consider Ξ a topological space defined by the Cartesian product of a non empty family of complete separable metric spaces $\{\Xi_t\}_{t \in \mathbb{T}}$ each one with a Borel σ-algebra $\mathcal{B}_t$. Let $\mathcal{B}$ denote the cylindrical σ-algebra generated by the projections (containing the Cartesian product of all the Borel σ-algebras $\mathcal{B}_t$), which is equal to the Borel σ-algebra for the product topology due to the separability assumption. Let X be measurable applications from $(\Omega, \mathcal{F})$ into $(\Xi, \mathcal{B})$, i.e. a stochastic processes. Then we can consider

$$A = \{\omega \in \Omega : X(\omega) \in \alpha\}, \qquad B = \{\omega \in \Omega : X(\omega) \in \beta\}$$

for α, $\beta \in \mathcal{B}$. In particular, we can consider events α and β about the ocurrence of some projections $A = \{X_t = a_0, X_{t-1} = a_1, ..., X_{t-L} = a_L\}$ and $B = \{X_{t+1} = b_1\}$, or more complex events such as $A = \{g(X_t, X_{t-1}, ..., X_{t-L}) = 0\}$ and $B = \{f(X_{t+1}, X_{t+2}, ..., X_{t+K}) = 0\}$. A variety of dynamic probabilistic causal relationships can be considered, including random shocks $\{\varepsilon_t\}$ in the expressions g and f.

Finally, notice that causality can be considered as a limit case in the probability space $(\Omega, \mathcal{F}, P)$. This is useful in contexts where causality is difficult to assess but can be studied by a series of related situations.

Definition Asymptotic probabilistic causality. Given a probability space $(\Omega, \mathcal{F}, P)$, consider a sequence of events $\{A_n\} \subset \mathcal{F}$, and $\{B_n\} \in \mathcal{F}$. I say that $\{A_n\}$ cause asymptotically $\{B_n\}$ in probability if $\lim_{n\to\infty} P(A_n \cap B_n^c) = 0$, and $\{A_n\}$ cause asymptotically $\{B_n\}$ almost surely if $P(\limsup_{n\to\infty} \{A_n \cap B_n^c\}) = 0$.

The use of limit ideas allows us to consider asymptotic forms of dynamic causality, such as $A = \{g(X_t, X_{t-1}, X_{t-2}....) = 0\}$ and $B = \{f(X_{t+1}, X_{t+2}, ...) = 0\}$, involving an infinite number of lags.

Although conceptually correct, the proposed notion of almost sure causality is too strong. Usually, scientists can relay on the strength of a causal relationship that only fails in quite rare events. In this sense, for the notion of causality it would be sufficient that the event $A \cap B^c$ has a small probability, even if it is not equal to zero.

Definition $\epsilon-$probabilistic causality. Given a probability space $(\Omega, \mathcal{F}, P)$, for any sets $A, B \in \mathcal{F}$ and any $\epsilon \in (0,1)$, I say that A causes B in terms of $\epsilon-$probabilistic causality if $\Pr(A \cap B^c) \leq \epsilon$. Asymptotic causality can be also defined on these terms requiring that $\lim_{n\to\infty} P(A_n \cap B_n^c) \leq \epsilon$ (in probability) or $P(\limsup_{n\to\infty} \{A_n \cap B_n^c\}) \leq \epsilon$ (almost surely), respectively.

Since we can consider several values $\epsilon \in [0,1]$, I define the causality ε-tolerance as

$$\varepsilon = \inf \{\epsilon \in [0,1] : P(A \cap B^c) \leq \epsilon\} = P(A \cap B^c).$$

If it is equal to zero, there is almost sure causality.

Physical laws can be described at different coarser levels of detail, and there might be causal relationships that are valid relatively to a specific level, but not in general. In Appendix A, I discuss the concept of conditional probabilistic causation.

Sequential causality

The probabilistic causality notion has been defined as a stable fact: if we think that A implies B almost surely, this notion is not subject to any change. However, what is believed true or false can change with time (e.g., due to Bayesian learning, or other exogenous changes). We introduce sequential causation to study causality for cases where the basic beliefs evolve, often due to information arrival.

In this section I will consider a form of sequential causality. Thus, I consider a sequence of probability spaces $\{(\Omega, \mathcal{F}_n, P_n)\}$, where the $\{\mathcal{F}_n\}$ is a filtration (i.e., a non decreasing sequence of

σ-algebras $\mathcal{F}_n \subset \mathcal{F}_m$ if $n < m$), and define $\mathcal{F}$ as the smallest σ-algebra containing the union of all the $\mathcal{F}_n$. A particular case is the non-learning situation, where $\mathcal{F}_n = \mathcal{F}$ for all n and all the P_n are defined on the same $\mathcal{F}$. We can define a type of probabilistic causation in the following sense.

Definition Sequential Causality in Probability. Given a sequence of probability spaces $(\Omega, \mathcal{F}_n, P_n)$.

If $A, B \subset \mathcal{F}$ I say A causes B sequentially if $\lim_{n\to\infty} P_n(A \cap B^c) = 0$. Furthermore, for any sequence $\{A_n\}$ with $A_n \in \mathcal{F}_n$ for all n, and any $B \in \mathcal{F}$, I say that that $\{A_n\}$ causes B sequentially with respect to $\{P_n\}$ if,

$$\lim_{n\to\infty} P_n(A_n \cap B^c) = 0.$$

Similarly, if $A_n, B_n \in \mathcal{F}_n$ for each n, I can say that that $\{A_n\}$ causes $\{B_n\}$ sequentially with respect to $\{P_n\}$ if, $\lim_{n\to\infty} P_n(A_n \cap B_n^c) = 0$.

Sequential causality can be related with convergence of probability measures. Clearly, if $P_n \to P$ in the sense that $\lim_{n\to\infty} P_n(A) = P(A)$ for all set $A \in \mathcal{F}$, then A causes B sequentially respect to $\{P_n\}$ where $A, B \in \mathcal{F}$, implies that $A \overset{a.s.[P]}{\Rightarrow} B$. If Ω is a topological space, and $\{\mathcal{F}_n\}$ is included in the Borel σ-algebra, the idea can be extended to weak convergence. Assume that $P_n \to_{w.} P$ (i.e. $\lim_{n\to\infty} P_n(A) = P(A)$ for all A in the Borel σ-algebra with $P(\partial A) = 0$ where ∂A is the frontier of A), if A causes B sequentially with respect to $\{P_n\}$ and $P(\partial(A \cap B^c)) = 0$, then $A \overset{a.s.[P]}{\Rightarrow} B$. In some cases we can also consider sequences of causal events: if P_n converges to P in the variational norm (i.e. $\sup_{A\in\mathcal{F}} |P_n(A) - P(A)| = 0$) then $\lim_{n\to\infty} P_n(A_n \cap B_n^c) = 0$ implies that $\{A_n\}$ cause asymptotically $\{B_n\}$ in probability P (i.e. $\lim_{n\to\infty} P(A_n \cap B_n^c) = 0$). In Appendix B, I discuss the robustness of probabilistic and sequential causation.

The events A_n, B_n can be defined in terms of random variables. We define a sequence of random variables $\{X_n\}$, i.e. a sequence of measurable applications from $(\Omega, \mathcal{F}_n, P_n)$ on the Borel measurable space $\left(\mathbb{R}^d, \mathbb{B}^d\right)$, and a measurable application X from $(\Omega, \mathcal{F}, P)$ on the $\left(\mathbb{R}^d, \mathbb{B}^d\right)$. For example, let us consider the events $A_n = \{X_n \in \alpha\}$ and $B = \{X \in \beta\}$, where $\alpha, \beta \in \mathbb{B}^d$, then we can consider the asymptotic causality by the requirement $\lim_{n\to\infty} E_{P_n}[I(X_n \in \alpha)(1 - I(X \in \beta))] = 0$. These expectations can be estimated empirically, and related to empirical stochastic processes for triangular arrays. The rest of the paper is focused on the empirical study of probabilistic causation.

EMPIRICAL ANALYSIS AND CAUSALITY

Karl Popper (1959) classifies theories as either metaphysical or physical, where the physical ones make predictions about the real world and can be empirically tested, in contrast to the metaphysical theories. For example, this paper discusses metaphysical concepts about causality. But not every physical theory is scientific. Popper used the notion of falsifiability to "demarcate" what science is: a physical theory is falsifiable (refutable or testable) if it can be shown false by real experience (from direct observation or a controlled physical experiment). For example, theological dogmas are unfalsifiable. According to Popper, if the physical theory fails it must be discarded or reshaped, but passing empirical tests does not ensure that the falsifiable theory is true. In this section I formalize these concepts, and study probabilistic causality from an empirical point of view.

Here I formalize some of these notions. Therefore, I will consider a universe of contingent elementary events Ω, in this set I define a σ-algebra $\mathcal{F}$ of subsets or events that we are interested in studying. We say that the universe Ω is physical if there exists a set of $\mathcal{X}$ applications $X : \Omega \to \mathbb{R}$ of (observable) empirical signals about the universe, and name $(\Omega, \mathcal{X})$ a physical space. Then, I say that $\mathcal{F}$ is a testable σ-algebra (or falsifiable) if $\mathcal{F} \subset \sigma(\mathcal{X})$, where $\sigma(\mathcal{X})$ is the smallest σ-algebra such that the applications in $\mathcal{X}$ are measurable (i.e. $\sigma(\mathcal{X}) = \left\{X^{-1}(U) : B \in \mathbb{B}, X \in \mathcal{X}\right\}$ with $\mathbb{B}$ the Borel real σ-algebra). For us, scientific knowledge is demarcated by physical and testable measure spaces $(\Omega, \mathcal{F})$. Then, we can consider (probabilistic) scientific knowledge from two sources: deduction and induction.

1. **Probabilistic Deduction:** Given a measurable space $(\Omega, \mathcal{F})$ (physical and testable in sciences), let us consider some measurable assumptions $\{B_j\}_{j \in J}$ that we call "premises" with known probabilities $\{P(B_j)\}$. Deduction means the computation of probabilities for other events that can be expressed in terms of the premises by using countable intersections, unions or conjugations and the probability axioms.

For a rich enough class of premises, $P(A)$ can be deduced for any $A \in \mathcal{F}$, whilst if not, we can only deduce probabilities for events in $\sigma\left(\{B_j\}_{j \in J}\right)$, and the events excluded from this σ-algebra will be called "conjectures". In deduction we take the probabilities $\{P(B_j)\}$ as known, but in the scientific method this knowledge comes from induction.

2. **Probabilistic Induction:** Given a physical and testable measurable space $(\Omega, \mathcal{F})$, we perform statistical inference to assign probabilities $P(\bullet)$ to some measurable events $A \in \mathcal{F}$ from

empirical data. In some cases, the inductive inference estimates $P(\bullet)$ for a few simple events from which the probability of A can be **deduced**, but often we can directly estimate $P(A)$. More often, we try to estimate the whole distribution P.

Deduction and induction can be combined. Given a testable physical space $(\Omega, \mathcal{X})$, the process of selecting a set of premises $\{B_j\}_{j\in J} \subset \Omega$ generating a σ-algebra $\mathcal{F} = \sigma\left(\{B_j\}_{j\in J}\right)$ such that $\mathcal{F} \subset \sigma(\mathcal{X})$ is called "abduction." Then, induction can be used to allocate probabilities to the premises, and then to deduce the probability of more complex events in $\mathcal{F}$. This is central to the scientific method. Note also that the physical space $(\Omega, \mathcal{X})$ can change with new information. We can consider a monotonously increasing sequence $\{\mathcal{X}_n\}$ of sets of signals about Ω, and define a filtration $\{\mathcal{F}_n\}$ with $\mathcal{F}_n \subset \sigma(\mathcal{X}_n)$ of refined testable theories. We can explain the progress of scientific knowledge as the permanent process of getting information $\mathcal{X}_n$, and updating the probability inferences of contingent assertions by conditioning on the σ-algebra $\mathcal{F}_n$ refined by the new information $\mathcal{X}_n$. Popper's discussion about falsified/unfalsified theories can be reconsidered under the light of the presented framework.

Regarding the estimation of P in the induction process, classical inference methods usually quantify the probability measure $P(\bullet)$, considering a parametric model or family $\{P_\theta : \theta \in \Theta\}$, and setting a single value $\theta_0 \in \Theta$ in a separable metric space[3] that minimizes some adjusting function or specific distance $D(P, P_\theta)$ between the model and the true probability. Then, the induction process estimates θ_0 by minimizing the adjustment $D(\mathbb{P}, P_\theta)$ of P_θ to the empirical distribution $\mathbb{P}$ (or smoothed version of it) of the collected data. These procedures are compatible with the frequentist view about probability laws, providing sound ground for scientific analysis.

Members of the "Bayesian statistics" school, however, think that the probability or degree of belief $P(\bullet)$ cannot be "quantified" precisely. Instead, they postulate a set of possible probability laws usually parametrized by some model $\{P_\theta : \theta \in \Theta\}$, quantifying the probability of each law with an arbitrary prior belief distribution $\pi(\theta)$, and updating their prior assumption with new data, computing $\pi(\theta|data)$ with the Bayes theorem[4]. Therefore, Bayesian statistics computes a "diffuse" or imprecise quantification P_θ of the degree of belief for falsifiable events. The posterior distribution is dependent upon the initial prior π assumption, although its influence is reduced when the dataset

[3] In classical inference Θ is included in an euclidean space. But it could alternatively be included in an infinito-dimensional space (as in the nonparametric and semiparametric literature, for example)

[4] It is not clear why Bayesians think that it is possible to quantify the prior. One could similarly parametrize a family of "priors" $\{\pi_\gamma(\theta) : \gamma \in \Gamma\}$ and postulate a higher order prior probability $\mu(\gamma)$ for these parameters, and so on, developing a hierarchically complex Bayesian structure.

increases. The approach is somewhat useless and therefore, paradoxically, Bayesians are forced to use P_{θ^e} as a quantification of the probability $P(\bullet)$ where $\theta^e = E[\theta|data]$. Then, classical and Bayesian methods show striking analogies in a few models. The Bayesian approach is not accepted by many scientists because it introduces more subjectivity in the induction process than the classical approach, and will be avoided in this paper.

The analysis of probabilistic causal relationships can be tackled from basic premises using deduction, or it can be studied directly from empirical data using induction. This is the aim of the second part of this paper. If we use empirical data, the analysis can lead to the rejection/acceptance of a probabilistic causal relationship through hypothesis testing. The empirical tests are performed with a significance level (quantifying the probability of rejecting a valid theory) and power (quantifying the probability of rejecting a false theory). From the empirical point of view, the study of reciprocal causality is a similar problem, and I will focus on one-directional analysis.

To study an almost sure causal relationship $A \overset{a.s.[P]}{\Rightarrow} B$, I consider the Bernoulli random variable $\delta = I(\omega \in A \cap B^c)$ with $\pi = P(\delta = 1) = P(A \cap B^c)$. The simplest empirical tests of causation must study the relative frequency from a sample of independent observations $\{\delta_1, ..., \delta_n\}$. If I define $D_n = \sum_{i=1}^{n} \delta_i$ then $\widehat{\pi}_n = D_n/n$ is the relative frequency. A simple decision test is the classical technique of counter-example: if $\widehat{\pi}_n > 0$ (i.e. $D_n > 0$) then we reject the a.s. causality. The significance level of this test is 100% since $\Pr(D_n > 0) = 0$ under the null $\pi = 0$, and the test is consistent as the power is given by

$$\Pr(D_n > 0) = 1 - \Pr(D_n = 0) = 1 - (1 - \pi)^n \to 1,$$

for any alternative $\pi \in (0, 1]$. Therefore, a causal relationship is rejected if it has at least one counterexample, and the power of this reasoning increases with n.

In empirical analysis, even if the causal relationship $A \overset{a.s.[P]}{\Rightarrow} B$ is true, experiments are often affected by noisy elements corrupting the probability π, so that for a large enough sample we can easily observe a relative frequency $\widehat{\pi}_n$ small but positive. In other words, if we apply the counter-example rule, in many cases we would end up rejecting causality. The counter-example rule is too strict for studying causality in real world situations where the value $P(A \cap B^c)$ can be small but not too clear based on empirical data. Two strategies can be considered to introduce more flexibility:

The first approach is based on sequential causality concepts. Let us consider a sequence $\{(\Omega, \mathcal{F}_n, P_n)\}_{n=1}^{\infty}$, and $\{A_n\}$ causes B sequentially for $\{P_n\}$, (in particular we can consider $A_n = A$ for all n). We

define $\pi_n = P_n(A_n \cap B^c) \in (0,1)$, and consider that asymptotically

$$\pi_n = c + \lambda/n + o\left(n^{-1}\right)$$

for a small value $\lambda > 0$ and some $c \geq 0$. The null assumption of sequential causality can be stated by $H_0 : c = 0$ so that $n\pi_n \to \lambda$, whilst the alternative assumption is $H_1 : c > 0$ and $n\pi_n \to \infty$. Define a Bernoulli variable $\delta_n = I(\omega \in A_n \cap B^c)$ with $P_n(\delta_n = 1) = \pi_n$, and consider a triangular array of i.i.d observations $\{\delta_{n1}, ..., \delta_{nn}\}$ distributed as δ_n for each n. Then I define $D_n = \sum_{i=1}^{n} \delta_{ni}$. By the De Moivre-Laplace theorem, under the null $D_n \longrightarrow_{w.[P_n]} Poiss(\lambda)$ where $Poiss(\lambda)$ denotes a Poisson random variable with parameter λ. Therefore, if we reject the null for $D_n > 0$, the probability of rejection under the null is approximately

$$\Pr(D_n > 0) = 1 - \Pr(D_n = 0) \to 1 - e^{-\lambda}.$$

close to zero if λ is very small ($e^{-\lambda} \to 1$ when $\lambda \downarrow 0$). Actually, we can control the significance level in the testing process, rejecting the null if $D_n > z_\alpha$, where z_α is chosen so that $\Pr(Poiss(\lambda) > z_\alpha) = \alpha$ under the null assumption. The value $\lambda > 0$ determines our robustness requirements. Under the alternative $\pi_n \to c > 0$ and $D_n \xrightarrow{p} \infty$ so that $\Pr(D_n > z_\alpha) \to 1$, i.e. the test is consistent.

We can consider a second procedure for situations where the causal relationship is only approximately satisfied, using the notion of $\epsilon-$probabilistic causality. For instance, if $\delta = I(\omega \in A \cap B^c)$ with $\pi = P(\delta = 1) = P(A \cap B^c) > 0$, using an i.i.d. sample $\{\delta_1, ..., \delta_n\}$ and $\widehat{\pi}_n = n^{-1} \sum_{i=1}^{n} \delta_i$ then we can test if $\pi < \epsilon$ for some specific $\epsilon \in (0,1)$, e.g. using a normal asymptotic distribution if n is large. Also, we can consider $\widehat{\pi}_n = n^{-1} \sum_{i=1}^{n} \delta_i$, as an estimate of the causality-tolerance ε. This second approach will be considered in the causal modelling section.

MODELLING AND CAUSALITY

So far I have discussed the empirical analysis of causality between fully specified theories A and B. But usually, scientific theories are not totally specific. They are defined by a general model, including parameters or flexible components to play with in order to empirically strengthen the postulated causal relationships. Researchers usually choose the parameters that are more likely to have a causal relationship from a pre-specified class, relating causality to modelling problems. Following this idea we can consider, for example, the events $A_\theta = \{g_\theta(X) \geq 0\}$ and $B_\theta = \{f_\theta(X) \geq 0\}$ (equalities could be alternatively considered) with $\theta \in \Theta \subset \mathbb{R}^K$ and modelers can seek the parameter $\theta_0 \in \Theta$ for which a causal relationship $A_{\theta_0} \Rightarrow B_{\theta_0}$ is most likely; i.e. we minimize the ε-tolerance in $\theta \in \Theta$.

Therefore, we are faced with the problem of solving

$$\min_{\theta\in\Theta} P\left(A_\theta \cap B_\theta^c\right) = \min_{\theta\in\Theta} \int l_\theta\left(X\right) dP.$$

where $l_\theta\left(X\right) = I\left(X \in A_\theta\right)\left(1 - I\left(X \in B_\theta\right)\right)$. These problems are central to a variety of physical sciences. Given a sample $\{X_1, ..., X_n\}$ from P, and appropriate identification assumptions, θ_0 can be estimated by $\widehat{\theta}_n$ minimizing the empirical analogous $\min_{\theta\in\Theta} \mathbb{P}_n\left(A_\theta \cap B_\theta^c\right)$, where $\mathbb{P}_n\left(A\right) = n^{-1}\sum_{i=1}^n I\left(X_i \in A\right)$ denotes the empirical distribution function, i.e.

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \int l_\theta\left(X\right) d\mathbb{P}_n.$$

Notice that causal modelling can be considered in terms of minimization of $\int l_\theta dP$ for a class of functions l_θ in a variety of cases.

1. First, in some contexts, it might be convenient to replace $I\left(X \in A_\theta \cap B_\theta^c\right)$ by a smooth approximation to the indicator function. For example, if A_θ, B_θ are closed sets in a separable metric space (Ξ, d) such as the euclidean space, then an indicator function $I\left(A\right)\left(x\right)$ can be approximated by the uniformly continuous function,

$$f_h\left(x, A\right) = \left(1 - d\left(x, A\right)/h\right)^+ \tag{5}$$

where $h > 0$, and it is satisfied that $I\left(A\right)\left(x\right) \leq f_h\left(x\right) \leq I\left(A^h\right)\left(x\right)$, where $A^h = \left\{x : d\left(x, A\right) < h\right\}$. Therefore, we can consider the class of smooth functions

$$\mathcal{L} = \left\{f_h\left(x, A_\theta \cap B_\theta^c\right) : \theta \in \Theta\right\}$$

or alternatively consider an approximation to each set A_θ and B_θ,

$$\mathcal{L} = \left\{l_\theta\left(x\right) = f_h\left(x, A_\theta\right)\left(1 - f_h\left(x, B_\theta\right)\right) : \theta \in \Theta\right\}.$$

2. We can also consider a convolution smoothing, considering the sets $A_\theta \cap B_\theta^c$ as elements of an L_1 space and approximate them by means of convolution, e.g. considering the class of functions

$$\mathcal{L} = \left\{l_{\theta,h}\left(x\right) = \int I\left(z \in A_\theta \cap B_\theta^c\right)\phi_h\left(z - x\right) dz : \theta \in \Theta\right\} \tag{6}$$

where the kernel ϕ_h denotes the $N\left(0, h^2 I\right)$ density (other kernels can be considered, alternatively), so that

$$\int l_{\theta,h}\, d\mathbb{P}_n = n^{-1}\sum_{i=1}^n \int_{A_\theta\cap B_\theta^c} \phi_h\left(z - X_i\right) dz = \int_{A_\theta\cap B_\theta^c} \left(n^{-1}\sum_{i=1}^n \phi_h\left(z - X_i\right)\right) dz,$$

meaning that we integrate $A_\theta \cap B_\theta^c$ with respect to a smooth density estimator. To improve the approximation behaviour we can alternatively consider a class of functions

$$\mathcal{L} = \left\{ l_{\theta,h_1,h_2}(x) = f_{h_1} * \phi_{h_2} : \theta \in \Theta \right\}$$

where l_{θ,h_1,h_2} is defined as the convolution of $f_{h_1}(x, A_\theta \cap B_\theta^c)$ with a Gaussian kernel ϕ_{h_2}.

3. For example, if we consider $A_\theta = \{g_\theta(X) \geq \varepsilon\}$ and $B_\theta = \{f_\theta(X) \geq \epsilon\}$ where (ε, ϵ) are unobserved random variables independent from X with distribution function $H(\varepsilon, \epsilon)$, then the optimal θ solves

$$\min_{\theta \in \Theta} E\left[l_\theta(X)\right]$$

where $l_\theta(X) = H(g_\theta(X_i), \infty) - H(g_\theta(X_i), f_\theta(X_i))$, and can be estimated using the estimator $\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \int l_\theta \, d\mathbb{P}_n$.

All the previous examples show that causal modelling consists of minimizing the expectation $E[l]$ for a class of parametrized functions $\mathcal{L}$. Let $(\Omega, \mathcal{F})$ be a testable physical measurable space, X an observable random vector with unknown probability P, and $\{X_1, ..., X_n\}$ a random sample from P. Assume that the class $\{(A_\theta \cap B_\theta^c)\}$ can be identified with a class $\mathcal{L}_\Theta = \{l_\theta : \theta \in \Theta\}$ in such way that $\min_{\theta\in\Theta} P(A_\theta \cap B_\theta^c) = \min_{\theta\in\Theta} \int l_\theta \, dP$. To simplify the notation, I will use the notation $P\, l = \int l \, dP$, and $\mathbb{P}_n \, l = \int l \, d\mathbb{P}_n$. We assume that for a class of measurable functions $\mathcal{L}$ the causal modelling problem is solved by some $l^* \in \mathcal{L}$,

$$l^* \in \arg\min_{l\in\mathcal{L}} \; P \, l$$

and l^* is estimated with the empirical analogous

$$\widehat{l}_n \in \arg\min_{l\in\mathcal{L}} \; \mathbb{P}_n \, l$$

We define the "sub-causality" associated to $l \in \mathcal{L}$ as

$$C_P(l) = P \, l - \min_{l\in\mathcal{L}} P \, l = P \, l - P \, l^*.$$

Then I say that $\widehat{l}_n$ is a consistent estimator of l^* in terms of probabilistic causality if $C_P\left(\widehat{l}_n\right)$ tends to zero almost surely when the sample size $n \to \infty$.

We focus on the $\epsilon-$probabilistic causality, which is the relevant case for empirical applications. Assume henceforth that $Pl^* > 0$. Notice that if the model allows to establish almost-sure causality $Pl^* = 0$, then $\Pr\{\mathbb{P}l^* = 0\} = 1$, so that $\mathbb{P}\widehat{l}_n \leq \mathbb{P}l^* = 0$ with probability one and $\Pr\left\{C_P\left(\widehat{l}_n\right) = 0\right\} = 1$ trivially. In the general case, though, $C_P\left(\widehat{l}_n\right) > 0$ for finite samples.

Consistency

Here I study the almost sure convergence of $C_P\left(\widehat{l}_n\right)$ to zero when $n \to \infty$, meaning that $P\,\widehat{l}_n$ is a good estimate for Pl^*. Central to the analysis is the study of the supremum norm of the empirical process $\{(\mathbb{P}_n - P)\ l\}_{l\in\mathcal{L}}$.

Using that $\mathbb{P}_n\,\widehat{l}_n \leq \mathbb{P}_n\,l^*$, it can be deduced that

$$\begin{aligned} C_P\left(\widehat{l}_n\right) &= P\,\widehat{l}_n - \mathbb{P}_n\,\widehat{l}_n + \mathbb{P}_n\widehat{l}_n - Pl^* \leq P\,\widehat{l}_n - \mathbb{P}_n\,\widehat{l}_n + (\mathbb{P}_n - P)\,l^* \\ &= (\mathbb{P}_n - P)\left(l^* - \widehat{l}_n\right). \end{aligned}$$

Using that

$$C_P\left(\widehat{l}_n\right) \leq \sup_{l,l'\in\mathcal{L}} |(\mathbb{P}_n - P)\ (l - l')| \leq 2\sup_{l\in\mathcal{L}} |(\mathbb{P}_n - P)\ l|\,,$$

an upper bound for $C_P\left(\widehat{l}_n\right)$ is obtained, which is determined by the supremum norm of the empirical process $\{(\mathbb{P}_n - P)\ l\}_{l\in\mathcal{L}}$. Furthermore, since $\mathbb{P}_n\,\widehat{l}_n - P\,\widehat{l}_n \leq \sup_{l\in\mathcal{L}} |(\mathbb{P}_n - P)\ l|$, bounding the supremum of the empirical process we get an estimate of the error made when $\mathbb{P}_n\,\widehat{l}_n$ is used to estimate $P\,\widehat{l}_n$. If I define the "empirical sub-causality",

$$C_{\mathbb{P}_n}\ (l) = \mathbb{P}_n\,l - \min_{l\in\mathcal{L}} \mathbb{P}_n\,l = \mathbb{P}_n\ \left(l - \widehat{l}_n\right),$$

then using that $Pl^* \leq P\widehat{l}_n$, I conclude that

$$\begin{aligned} C_{\mathbb{P}_n}\ (l) - C_P\,(l) &= \mathbb{P}_n\ \left(l - \widehat{l}_n\right) - P\ (l - l^*) = (\mathbb{P}_n - P)\ l + \left(Pl^* - \mathbb{P}_n\widehat{l}_n\right) \\ &\leq (\mathbb{P}_n - P)\ l - (\mathbb{P}_n - P)\,\widehat{l}_n = (\mathbb{P}_n - P)\left(l - \widehat{l}_n\right), \end{aligned}$$

and therefore $|C_{\mathbb{P}_n}\ (l) - C_P\,(l)|$ is also uniformly bounded by $2\sup_{l\in\mathcal{L}} |(\mathbb{P}_n - P)\ l|$.

The rest of the section is devoted the analysis of conditions ensuring that the supremum norm of the empirical process $\{(\mathbb{P}_n - P)\ l\}_{l\in\mathcal{L}}$ tends to zero, and its relation with the causal analysis. Let $(\mathcal{L}, \rho)$ be a semimetric space of functions $l : \mathbb{R}^d \to \mathbb{R}$, and $N(\varepsilon, \mathcal{L}, \rho)$ the covering number or minimum number of balls or radius ε needed to cover $\mathcal{L}$, and $\log N(\varepsilon, \mathcal{L}, \rho)$ is known as the metric entropy. An envelope for the class $\mathcal{L}$ is any function $L(x) \geq 0$ such that $|l(x)| \leq L(x)$ for all $x \in \mathbb{R}^d$ and all $l \in \mathcal{L}$. Since $\mathbb{P}_n$ puts all its mass on the set of observations $\{X_1, ..., X_n\}$, it is sufficient to our purposes to measure distances at this set. In this paper I consider the semi-metric space $(\mathcal{L}, \rho_n)$, for the stochastic distance

$$\rho_n(l, l') = \|l - l'\|_{L_1(\mathbb{P}_n)} = n^{-1}\sum_{i=1}^{n} |l(X_i) - l'(X_i)|\,.$$

We could consider the $L_p(\mathbb{P}_n)$ semi-noms. The covering numbers are useful tools for the analysis of empirical processes. Pollard (1984) proved that if $\mathcal{L}$ has a measurable envelope $L < M$, then

$$P\left\{\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l| > \delta\right\} \leq 8E\left[N\left(\delta/8, \mathcal{L}, \rho_n\right)\right] e^{-n\delta^2/128M^2},$$

which generalizes a popular bound by Vapnik and Chervonenkis (1971) and implies the ULLN when $E\left[\log N\left(\delta, \mathcal{L}, \rho_n\right)\right] = o(n)$ for all $\varepsilon > 0$ (actually, the condition is also necessary for the ULLN). In the case of non measurable events, we should replace $E[\cdot]$ and $P\{\cdot\}$ by outer expectations and probabilities, respectively. We will discuss later how the metric entropy condition can be checked by combinatorial arguments.

In some cases we consider a class of probability functions $\mathcal{P}$, for example we can consider a neighborhood of the true probability (an approach introduced by Le Cam). The ULLN can be proved also uniformly in the class $\mathcal{P}$. If,

$$\begin{aligned} \lim_{M\to\infty}\sup_{P\in\mathcal{P}} P\left(F\cdot I\left(F > M\right)\right) &= 0, \\ \sup_{Q\in Q_n}\log N\left(\varepsilon\left\|L\right\|_{L_1(Q)}, \mathcal{L}, L_1\left(Q\right)\right) &= o_p\left(n\right), \end{aligned} \tag{7}$$

for all $\varepsilon > 0$, then the ULLN is satisfied uniform in $P \in \mathcal{P}$, see van der Vaart and Wellner (1996, Th 2.8.1). This result is particularly useful in the context of sequential causality. For an appropriate sequence $\mathcal{P} = \{P_n\}$ of probability functions, $C_{P_n}\left(\widehat{l}_n\right) \to_{a.s.} 0$, as a consequence of the inequality $\sup_{P\in\mathcal{P}} C_P\left(\widehat{l}_n\right) \leq 2\sup_{P\in\mathcal{P}}\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l|$.

Error bounds and convergence rates

We can obtain more insightful bounds for $\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l|$, using that this variable is concentrated around its mean, since for all $\delta > 0$,

$$\begin{aligned} \Pr\left\{\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l| - E\left[\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l|\right] \geq \delta\right\} &\leq 2e^{-2n\delta^2} \Leftrightarrow \\ \Pr\left\{\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l| \leq E\left[\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l|\right] + \sqrt{\frac{2\ln(1/\delta)}{n}}\right\} &\geq 1-\delta, \end{aligned}$$

by McDiarmid's (1989) bounded difference inequality. The next result, based on Pollard's work, provides a bound $E\left[\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l|\right]$. Under appropriate conditions, this bound is $O\left(n^{-1/2}\right)$, which can be combined with McDiarmid's (1989) bounded difference inequality to obtain convergence rates for $C_P\left(\widehat{l}_n\right)$.

Theorem: Assume that $\mathcal{L}$ has a measurable envelope L such that $PL < \infty$. Then,

$$E\left[\sup_{l\in\mathcal{L}}\left|\left(\mathbb{P}_n - P\right)\ l\right|\right] \leq \frac{24}{\sqrt{n}} \sup_{Q\in Q_n} \int_0^1 \sqrt{\log 2N\left(\varepsilon \left\|L\right\|_{L_1(Q)}, \mathcal{L}, L_1\left(Q\right)\right)} d\varepsilon$$

where Q_n is the class of distribution functions with finite mass at n arbitrary points $\{x_1, ..., x_n\}$.

If for a class of probability functions $\mathcal{P}$, the measurable envelope satisfies (7), then

$$\sup_{P\in\mathcal{P}} E_P\left[\sup_{l\in\mathcal{L}}\left|\left(\mathbb{P}_n - P\right)\ l\right|\right] \leq \frac{24}{\sqrt{n}} \sup_{Q\in Q_n} \int_0^1 \sqrt{\log 2N\left(\varepsilon \left\|L\right\|_{L_1(Q)}, \mathcal{L}, L_1\left(Q\right)\right)} d\varepsilon.$$

NOTE: Usually $\sup_{x_1...,x_n} N\left(\varepsilon, \mathcal{L}, \rho_n\right) \to \infty$ when $\varepsilon \downarrow 0$, and since $\int_0^1 \varepsilon^{-r} d\varepsilon < \infty$ for all $r < 1$, for the convergence of the integral $\int_0^1 \sqrt{\log 2N\left(\varepsilon, \mathcal{L}, \rho_n\right)} d\varepsilon$ it suffices that $\sup_{Q_n} N\left(\varepsilon, \mathcal{L}, \rho_n\right) = o\left(\varepsilon^{-2}\right)$.

PROOF

Without loss of generality assume that the covering number and the integral are finite. Applying a standard symmetrization argument,

$$\begin{aligned} E\left[\sup_{l\in\mathcal{L}}\left|\left(\mathbb{P}_n - P\right)\ l\right|\right] &\leq 2E_X E_\varepsilon\left[\sup_{l\in\mathcal{L}}\left|n^{-1}\sum_{i=1}^n \varepsilon_i l\left(X_i\right)\right|\right] \\ &\leq 2E_X E_\varepsilon\left[\sup_{l\in\mathcal{L}_M}\left|n^{-1}\sum_{i=1}^n \varepsilon_i l\left(X_i\right)\right|\right] + 2P\left(F \cdot I\left(F > M\right)\right) \end{aligned}$$

where $\mathcal{L}_M = \left\{l \cdot I\left(F \leq M\right) : l \in \mathcal{L}\right\}$, for all $M > 0$ by the triangle inequality, see e.g. van der Vaart and Wellner (1996). We will consider the Rademacher process $E_\varepsilon\left[\sup_b\left|n^{-1}\sum_{i=1}^n \varepsilon_i b_i\right|\right]$ for an appropriate class of n-dimensional vectors b. This expectation can be bounded using different techniques of covering numbers (for example, uniform covering numbers, random covering numbers, bracketing numbers, etc.), and applying the associated Dudley's entropy integrals.

We define

$$\mathcal{L}_M\left(\mathbb{P}_n\right) = \left\{b \in \mathbb{R}^n : \exists l \in \mathcal{L}_M, b_i = l\left(X_i\right), i = 1, ..., n\right\}$$

and consider $d\left(b, b'\right) = n^{-1}\sum_{i=1}^n \left|b_i - b_i'\right|$, so that the isometry between $\left(\mathcal{L}_M\left(\mathbb{P}_n\right), d\right)$ and $\left(\mathcal{L}_M, \rho_n\right)$ is maintained. Then, we consider

$$E\left[\sup_{l\in\mathcal{L}_M}\left|n^{-1}\sum_{i=1}^n \varepsilon_i l\left(X_i\right)\right|\right] = E\left[\max_{b\in\mathcal{L}_M\left(\mathbb{P}_n\right)}\left|n^{-1}\sum_{i=1}^n \varepsilon_i b_i\right|\right] = E\left[\left|n^{-1}\sum_{i=1}^n \varepsilon_i b_i^*\right|\right],$$

and I study the expression on the right using the Kolmogorov chaining trick.

Let us define a sequence $\left\{\mathcal{L}_k\right\}$ such that $\mathcal{L}_k$ is a minimal cover of $\mathcal{L}_M\left(\mathbb{P}_n\right)$ of radius 2^{-k}, increasing to the value $k = M$ such that $\mathcal{L}_M$ is a minimal covering of radius 1 for the set $\mathcal{L}_M\left(\mathbb{P}_n\right)$. For each

$1 \leq k \leq M$ I define b^k as the nearest neighbor of b^* in the $k-th$ cover, so that $d\left(b^k, b^*\right) = \min\left\{d\left(b^k, b^*\right) : b \in \mathcal{L}_k\right\}$, and clearly $d\left(b^k, b^*\right) \leq 2^{-k}$, which means that

$$d\left(b^k, b^{k-1}\right) \leq d\left(b^k, b^*\right) + d\left(b^*, b^{k-1}\right) \leq 3 \cdot 2^{-k}.$$

Thus, setting $b^0 = 0$ I can consider a telescopic sum $\sum_{i=1}^n \varepsilon_i b_i^* = \sum_{k=1}^M \sum_{i=1}^n \varepsilon_i \left(b_i^k - b_i^{k-1}\right)$, and

$$\begin{aligned} E\left[\left|\sum_{i=1}^n \varepsilon_i b_i^*\right|\right] &\leq \sum_{k=1}^M E\left[\left|\sum_{i=1}^n \varepsilon_i \left(b_i^k - b_i^{k-1}\right)\right|\right] \\ &\leq \sum_{k=1}^M E\left[\max_{b \in \mathcal{L}_k, b' \in \mathcal{L}_{k-1}, d(b,b') \leq 3 \cdot 2^{-k}} \left|\sum_{i=1}^n \varepsilon_i \left(b_i - b_i'\right)\right|\right]. \end{aligned}$$

Applying the Hoeffding inequality, for each pair $b \in \mathcal{L}_k$ and $b' \in \mathcal{L}_{k-1}$ such that $d\left(b, b'\right) \leq 3 \cdot 2^{-k}$,

$$E\left[\exp\left\{x \sum_{i=1}^n \varepsilon_i \left(b_i - b_i'\right)\right\}\right] \leq 2e^{\frac{1}{2}x^2 n\left(3 \cdot 2^{-k}\right)^2}$$

and the number of such pairs is bounded by $\left|\mathcal{L}_k\right| \left|\mathcal{L}_{k-1}\right| = N\left(2^{-k}, \mathcal{L}_M\left(\mathbb{P}_n\right), d\right)^2$ (see e.g. Lugosi, 2002). Then, the Jensen inequality implies that for $1 \leq k \leq M$

$$\begin{aligned} & E\left[\max_{b \in \mathcal{L}_k, b' \in \mathcal{L}_{k-1}, d(b,b') \leq 3 \cdot 2^{-k}} \left|\sum_{i=1}^n \varepsilon_i \left(b_i - b_i'\right)\right|\right] \\ \leq\ & \ln E\left[\exp\left\{\max_{b \in \mathcal{L}_k, b' \in \mathcal{L}_{k-1}, d(b,b') \leq 3 \cdot 2^{-k}} \left|\sum_{i=1}^n \varepsilon_i \left(b_i - b_i'\right)\right|\right\}\right] \\ \leq\ & \ln\left(\sum_{b \in \mathcal{L}_k, b' \in \mathcal{L}_{k-1} d(b,b') \leq 3 \cdot 2^{-k}} E\left[\exp\left|\sum_{i=1}^n \varepsilon_i \left(b_i - b_i'\right)\right|\right]\right) \\ \leq\ & \ln\left\{N\left(2^{-k}, \mathcal{L}_M\left(\mathbb{P}_n\right), d\right)^2 \cdot 2\exp\left\{\sqrt{n} 3 \cdot 2^{-k}\right\}\right\} \end{aligned}$$

leading to

$$E\left[\left|\sum_{i=1}^n \varepsilon_i b_i^*\right|\right] \leq \sum_{k=1}^M 3\sqrt{n} 2^{-k} \sqrt{\log 2N\left(2^{-k}, \mathcal{L}_M\left(\mathbb{P}_n\right), d\right)^2}$$

which implies that

$$\begin{aligned} E\left[\sup_{l \in \mathcal{L}_M} \left|n^{-1} \sum_{i=1}^n \varepsilon_i l\left(X_i\right)\right|\right] &\leq \frac{3}{\sqrt{n}} \sum_{k=1}^M 2^{-k} \sqrt{\log 2N\left(2^{-k}, \mathcal{L}_M\left(\mathbb{P}_n\right), d\right)^2} \\ &\leq \frac{12}{\sqrt{n}} \sum_{k=1}^M 2^{-k} \sqrt{\log 2N\left(2^{-k}, \mathcal{L}_M\left(\mathbb{P}_n\right), d\right)} \\ &\leq \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\log 2N\left(\varepsilon, \mathcal{L}_M\left(\mathbb{P}_n\right), d\right)}, \end{aligned}$$

using that $N(\varepsilon, \mathcal{L}_M(\mathbb{P}_n), d)$ is a monotonously decreasing function of ε. Finally, since $PL < \infty$, then $\mathbb{P}_n L = O_P(1)$ and I can replace $N(\varepsilon, \mathcal{L}_M(\mathbb{P}_n), d)$ by $N\left(\varepsilon \|L\|_{L_1(\mathbb{P}_n)}, \mathcal{L}(\mathbb{P}_n), d\right)$ where

$$\mathcal{L}(\mathbb{P}_n) = \{b \in \mathbb{R}^n : \exists l \in \mathcal{L}, b_i = l(X_i), i = 1, ..., n\}.$$

By the isometry, I can consider equivalently $N\left(\varepsilon \|L\|_{L_1(\mathbb{P}_n)}, \mathcal{L}, \rho_n\right)$. The expectation of $\sup_{l\in\mathcal{L}} |(\mathbb{P}_n - P)\ l|$ is bounded by two times the upper bound for the Rademacher process, so that

$$E\left[\sup_{l\in\mathcal{L}} |(\mathbb{P}_n - P)\ l|\right] \leq \frac{24}{\sqrt{n}} \sup_{x_1...,x_n} \int_0^1 \sqrt{\log 2N\left(\varepsilon \|L\|_{L_1(\mathbb{P}_n)}, \mathcal{L}, \rho_n\right)} d\varepsilon.$$

This proves the first part of the theorem.

For the second part, consider that

$$\sup_{l\in\mathcal{L}} |(\mathbb{P}_n - P)\ l| \leq \sup_{l\in\mathcal{L}_M} |(\mathbb{P}_n - P)\ l| + (\mathbb{P}_n + P) F \cdot I(F > M)$$

and therefore,

$$\sup_{P\in\mathcal{P}} E\left[\sup_{l\in\mathcal{L}} |(\mathbb{P}_n - P)\ l|\right] \leq 2 \sup_{P\in\mathcal{P}} E_P E_\varepsilon \left[\sup_{l\in\mathcal{L}_M} \left| n^{-1} \sum_{i=1}^n \varepsilon_i l(X_i) \right|\right] + 2 \sup_{P\in\mathcal{P}} P(F \cdot I(F > M))$$

then apply the same type of argument.

END-PROOF

The only issue is to compute the covering number $N(\varepsilon, \mathcal{L}, \rho_n)$. We use the combinatorial method introduced by Vapnik and Chervonenkis (1974) and Vapnik (1998). For a given class $\mathcal{A}$ of subsets of $\mathbb{R}^d$, for an arbitrary set $x_1^n = \{x_1, ..., x_n\}$ of n points in $\mathbb{R}^d$, we define

$$\Delta^{\mathcal{A}}(x_1, ..., x_n) = card\{\{x_1, ..., x_n\} \cap A : A \in \mathcal{A}\}.$$

Then, the Vapnik-Chervonenkis (VC) n-shatter coefficient of $\mathcal{A}$ as the maximal number of different subsets of a set of n points $\{x_1, ..., x_n\}$ which can be obtained by intersecting it with the elements of $\mathcal{A}$,

$$\mathbb{S}_{\mathcal{A}}(n) = \sup\left\{\Delta^{\mathcal{A}}(x_1, ..., x_n) : x_1, ..., x_n \in \mathbb{R}^d\right\}.$$

The VC n-shatter coefficient can be equivalently defined as $\mathbb{S}_{\mathcal{A}}(n) = \sup_{x_1,...,x_n} |\mathcal{A}(x_1^n)|$, where $\mathcal{A}(x_1^n) = \{b \in \{0,1\}^n : \exists A \in \mathcal{A} : b_i = I(x_i \in A)\}$. The Vapnik-Chervonenkis (VC) dimension (or index) of the set $\mathcal{A}$ is defined as

$$V_{\mathcal{A}} = \inf\{n \geq 1 : \mathbb{S}_{\mathcal{A}}(n) < 2^n\},$$

and $V = \infty$ if $\mathbb{S}_{\mathcal{A}}(n) = 2^n$ for all n. Note that if $\mathbb{S}_{\mathcal{A}}(n) < 2^n$ then $\mathbb{S}_{\mathcal{A}}(m) < 2^m$ for all $m > n$ and the VC dimension is well defined. We say that $\mathcal{A}$ is a VC class if $V_{\mathcal{A}} < \infty$. The specific value $V_{\mathcal{A}}$ depends on the complexity of the considered model $\mathcal{A}$, and it has been computed for commonly used parametrized sets, see e.g. Devroye et al. (1996, Chapter 13). For example, if $\mathcal{A} = \{x : g(x) \geq 0, g \in \mathcal{G}\}$ and $\mathcal{G}$ is an $m-$dimensional vector space of real valued functions defined on $\mathbb{R}^d$, then $V_{\mathcal{A}} \leq m$. If $\mathcal{A}$ is the class of all linear half-spaces in $\mathbb{R}^d$, the VC dimension is $d+2$. If $\mathcal{A}$ is the class of all closed balls in $\mathbb{R}^d$, the VC dimension is $d+1$. If $\mathcal{A}$ is the class of rectangles in $\mathbb{R}^d$, the VC dimension is $2d$.

By definition, a VC class of sets picks out strictly less than 2^n subsets from any set of $n \geq V_{\mathcal{A}}$ elements, however in practice the value is much slower than the $2^n - 1$ possible ones. Sauer's lemma states that if $V_{\mathcal{A}} < \infty$, then $\mathbb{S}_{\mathcal{A}}(n) \leq \sum_{i=1}^{V_{\mathcal{A}}} \binom{n}{i}$ for all n, implying a that $\mathbb{S}_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}$ for all n, and also that $\mathbb{S}_{\mathcal{A}}(n) \leq (ne/V_{\mathcal{A}})^{V_{\mathcal{A}}}$ for $n \geq V_{\mathcal{A}}$. Therefore $\mathbb{S}_{\mathcal{A}}(n) = O\left(n^{V_{\mathcal{A}}-1}\right)$. Covering numbers and VC dimension are be related. Applying the Sauer lemma, Dudley (1978) proved that if $\mathcal{A}$ has VC dimension $V_{\mathcal{A}} < \infty$, then

$$N(\varepsilon, \mathcal{A}, \rho_n) \leq \left(4e/\varepsilon^2\right)^{V_{\mathcal{A}}/(1-1/e)}.$$

Haussler (1995) refined the bound to $N(\varepsilon, \mathcal{A}, \rho_n) \leq e(V_{\mathcal{A}}+1)\left(2e/\varepsilon^2\right)^{V_{\mathcal{A}}}$.

To apply these results in the context of the class of real valued functions $\mathcal{L}$, we define the subgraph of a function $l \in \mathcal{L}$ as the set $\{(t,x) : t < l(x)\}$. Then the VC dimension of the function class $\mathcal{L}$ is defined as the VC dimension of all the subgraphs of functions l in $\mathcal{L}$, denoted by $V_{\mathcal{L}}$. In particular, when $\mathcal{L} = \{I(A_\theta \cap B_\theta^c) : \theta \in \Theta\}$, i.e. it is defined by indicator functions, then the subgraph VC dimension of $\mathcal{L}$ is equal to the VC dimension of the class of sets family $\mathcal{A} = \{(A_\theta \cap B_\theta^c) : \theta \in \Theta\}$. When smooth parametric classes $\mathcal{L}$ are considered we should compute the VC dimension directly, but $V_{\mathcal{L}}$ has been already computed in the nonparametric literature for the most commonly used families of smooth functions. For for all $0 < \varepsilon < 1$, and $p \geq 1$, and a universal constant $K > 0$,

$$\sup_Q N\left(\varepsilon \|L\|_{L_p(Q)}, \mathcal{L}, \|\cdot\|_{L_r(Q)}\right) \leq KV_{\mathcal{L}}(16e)^{V_{\mathcal{L}}}\left(\frac{1}{\varepsilon^2}\right)^{p(V_{\mathcal{A}}-1)}.$$

the supremum over all probabilities Q such that $\|L\|_{L_p(Q)} > 0$ (the proof can be found in van der Vaart and Wellner, 1996, Th. 2.6.7). If $PL < \infty$, applying the Theorem 1 leads to

$$E\left[\sup_{l \in \mathcal{L}} |(\mathbb{P}_n - P)\ l|\right] \leq C\sqrt{\frac{V_{\mathcal{L}}}{n}}, \tag{8}$$

for a universal constant $C > 0$. Under this premise, with probability $1 - \delta$,

$$P\,\widehat{l}_n \leq \mathbb{P}_n\,\widehat{l}_n + C\sqrt{\frac{V_{\mathcal{L}}}{n}} + 2\sqrt{\frac{2\ln(1/\delta)}{n}}.$$

by McDiarmid's (1989) inequality. Therefore

$$C_P\left(\widehat{l}_n\right) \leq 2C\sqrt{\frac{V_{\mathcal{L}}}{n}} + 4\sqrt{\frac{2\ln(1/\delta)}{n}}$$

with probability $1 - \delta$.

Interestingly, we can use directly the combinatorial arguments on the rademacher process to obtain that

$$E\left[\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\,l|\right] \leq 2E_X E_\varepsilon\left[\sup_{l\in\mathcal{L}}\left|n^{-1}\sum_{i=1}^{n}\varepsilon_i l(X_i)\right|\right] \leq 2\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{L}}(n)}{n}},$$

see Devroye and Lugosi (2001, Th. 3.1.), and by McDiarmid's (1989) inequality,

$$P\,\widehat{l}_n \leq \mathbb{P}_n\,\widehat{l}_n + 4\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{L}}(n)}{n}} + 2\sqrt{\frac{2\ln(1/\delta)}{n}}$$

with probability $1 - \delta$, implying that $C_P\left(\widehat{l}_n\right) = O_P\left(\sqrt{V_{\mathcal{L}}\ln n/n}\right)$ for VC classes by the Sauer lemma. For a small size n, this bound could be better than the main one with rate $O_P\left(\sqrt{V_{\mathcal{L}}/n}\right)$.

We can derive a sharper upper bound on $C_P\left(\widehat{l}_n\right)$ using Talagrand's (1996 a, b) concentration inequality instead of McDiarmid's (1989) one. Talagrand's (1996 a, b) concentration inequality guaranties that, with probability $1 - \delta$,

$$\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l| \leq E\left[\sup_{l\in\mathcal{L}}|(\mathbb{P}_n - P)\ l|\right] + \sqrt{\frac{2\sup_{l\in\mathcal{L}}Var(l)\log\left(\frac{1}{\delta}\right)}{n}} + 4\frac{\log\left(\frac{1}{\delta}\right)}{3n}.$$

where $Var(l) = Pl^2 - (Pl)^2$. The advantage of Talagrand's inequality is that it can be used to control oscillations of the empirical process locally. Assume that $Var(l) \leq c(Pl)^\alpha$, for some $c, \alpha > 0$ and all $l \in \mathcal{L}$, so that

$$Var(l) \leq c\left(C_P(l) + Pl^*\right)^\alpha.$$

For example, $Var(l) = Pl = C_P(l) + Pl^*$ when l is defined by indicator function of Borel sets. Then, if we define the set $\mathcal{L}_r = \{l \in \mathcal{L} : C_P(l) \leq r\}$, and $\phi_n(\mathcal{L}, r) = E\left[\sup_{l\in\mathcal{L}_r}|(\mathbb{P}_n - P)\ l|\right]$, we have that with probability at least $1 - \delta$,

$$\sup_{l\in\mathcal{L}_r}|(\mathbb{P}_n - P)\ l| \leq \phi_n(\mathcal{L}, r) + \sqrt{\frac{2c\left(r + P\ l^*\right)^\alpha\log\left(\frac{1}{\delta}\right)}{n}} + 4\frac{\log\left(\frac{1}{\delta}\right)}{3n}.$$

We denote two times the right hand side of the inequality by $\varphi_n(r)$; it is an increasing nonnegative function. Therefore, for all $l \in \mathcal{L}_r$,

$$C_P(l) \leq 2 \sup_{l \in \mathcal{L}_r} |(\mathbb{P}_n - P)\ l| \leq \varphi_n(r)$$

Applying this argument to the subclass $\mathcal{L}_r$ containing all functions with ε-tolerances less than that of $\widehat{l}_n$, we obtain that with probability at least $1-\delta$

$$C_P\left(\widehat{l}_n\right) \leq \varphi_n(r).$$

Taking a smaller r we can improve the bound. The sharpest bound is obtained by considering a fixed point $r^* = \varphi_n(r^*)$, leading to $C_P\left(\widehat{l}_n\right) \leq r_n^*$ with probability at least $1-\delta$, which can be formally proved applying the arguments in Massart and Nédélec (2006) and Koltchinskii (2006).

Parametric and semiparametric estimation

As I have previously discussed, causal models can be defined by classes of measurable sets $\mathcal{A} = \{(A_\theta \cap B_\theta^c) : \theta \in \Theta\}$, and/or classes of functions $\mathcal{L} = \{l_\theta : \theta \in \Theta\}$ with $P(A_\theta \cap B_\theta^c) = E[l_\theta]$, (the simplest case is $l_\theta = I(A_\theta \cap B_\theta^c)$). Let us define the optimal parameter $\theta^* = \{\theta \in \Theta : l_\theta = l^*\}$, and the estimator $\widehat{\theta}_n = \left\{\theta \in \Theta : l_{\widehat{\theta}_n} = \widehat{l}_n\right\}$. Next I discuss the consistency of the parametric estimators and the convergence rate.

We will assume that the parameter set Θ belongs to a separable semi-metric space with a pseudo-distance d (which may be perfectly dependent on the unknown distribution P). In a variety of cases, the pseudo-distance and the causality tolerance are related in such way that $C_P(l_\theta) \geq c \cdot d(\theta, \theta^*)$ for all $\theta \in \Theta$. For example, if we consider $d^2\left(\theta, \theta'\right) = P\left[\left|l_\theta - l_{\theta'}\right|^2\right]$, and the measurable envelope of $\mathcal{L}$ is bounded by $k > 0$, then

$$d^2(\theta, \theta^*) = P\left[\left|l_\theta - l_{\theta^*}\right|^2\right] \leq 2k \cdot P\left[\left|l_\theta - l_{\theta^*}\right|\right] = 2k \cdot P\left(l_\theta - l_{\theta^*}\right) = 2k \cdot C_P(l),$$

and the inequality holds with $c = 1/2k$. If $d^2\left(\theta, \theta'\right) = Var\left[\left(l_\theta(X) - l_{\theta'}(X)\right)\right]$ also $d^2(\theta, \theta^*) \leq P\left[\left|l_\theta - l_{\theta^*}\right|^2\right] \leq 2k \cdot C_P(l)$.

The condition $C_P(l_\theta) \geq c \cdot d^2(\theta, \theta^*)$ implies that θ^* is locally identified as a minimum of $C_P(l_\theta)$ (i.e., $\forall \varepsilon > 0$, $\exists \eta > 0$ such that $\inf_{d(\theta,\theta^*)>\varepsilon} C_P(l_\theta) > \eta$, considering $\eta = c\varepsilon^2$). These bounds, combined with bounds for the expected value of the empirical process modulus of continuity, can be used to obtain consistency and convergence rates:

Theorem: Assume that for all $\theta \in \Theta$, there exists a constant $C > 0$,

$$C_P(l_\theta) \geq c \cdot d^2(\theta, \theta^*).$$

A) Assume that for every n and for all $\delta > 0$, the centered process $(\mathbb{P}_n - P)\ l$ satisfies that

$$E\left[\sup_{d(\theta,\theta^*)>\delta} |(\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta)|\right] \leq \frac{\zeta_n(\delta)}{\sqrt{n}}, \tag{9}$$

for functions $\zeta_n(\delta)$ such that $\delta \longmapsto \zeta_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ not depending on n. If $r_n^2 \cdot \zeta_n(1/r_n) = O(\sqrt{n})$ for all n, then $\widehat{\theta}_n \to^p \theta^*$ and $d\left(\widehat{\theta}_n, \theta^*\right) = O_P\left(r_n^{-1}\right)$.

B) Alternatively, assume that

$$E\left[\sup_{d(\theta,\theta^*)<\delta} |(\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta)|\right] \leq \frac{\gamma_n(\delta)}{\sqrt{n}}$$

for functions $\gamma_n(\delta)$ such that $\delta \longmapsto \gamma_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ not depending on n. If $r_n^2 \cdot \gamma_n(1/r_n) = O(\sqrt{n})$ for all n, then $\widehat{\theta}_n \to^p \theta^*$ and $d\left(\widehat{\theta}_n, \theta^*\right) = O_P\left(r_n^{-1}\right)$.

PROOF

A) Note first that for all $\varepsilon > 0$, if

$$\begin{aligned} d\left(\widehat{\theta}_n, \theta^*\right) &\geq \varepsilon/r_n \Longrightarrow \widehat{\theta}_n \notin B(\theta^*, \varepsilon/r_n) \Leftrightarrow \\ \inf_{d(\theta,\theta^*)\geq\varepsilon/r_n} \mathbb{P}_n\ l_\theta &\leq \inf_{\theta\in\Theta} \mathbb{P}_n\ l_\theta = \mathbb{P}_n\ l_{\widehat{\theta}_n} \leq \mathbb{P}_n\ l_{\theta^*}, \end{aligned}$$

and therefore,

$$\begin{aligned} P^*\left\{r_n \cdot d\left(\widehat{\theta}_n, \theta^*\right) \geq \varepsilon\right\} &\leq \Pr\left\{\inf_{d(\theta,\theta^*)>\varepsilon/r_n} \mathbb{P}_n\ (l_\theta - l_{\theta^*}) \leq 0\right\} \\ &= \Pr\left\{\inf_{d(\theta,\theta^*)>\varepsilon/r_n} \{C_P(l_\theta) - (\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta)\} \leq 0\right\} \\ &\leq \Pr\left\{\inf_{d(\theta,\theta^*)>\varepsilon/r_n} C_P(l_\theta) - \sup_{d(\theta,\theta^*)>\varepsilon/r_n} (\mathbb{P}_n - P)\ (l_\theta - l_{\theta^*}) \leq 0\right\} \end{aligned}$$

using that

$$\mathbb{P}_n\ (l_\theta - l_{\theta^*}) = P(l_\theta - l_{\theta^*}) + (\mathbb{P}_n - P)(l_\theta - l_{\theta^*}) = C_P(l_\theta) - (\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta)$$

Therefore, if

$$\begin{aligned} &P^*\left\{r_n \cdot d\left(\widehat{\theta}_n, \theta^*\right) \geq \varepsilon\right\} \leq \Pr\left\{\sup_{d(\theta,\theta^*)>\varepsilon/r_n} (\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta) \geq \inf_{d(\theta,\theta^*)>\varepsilon/r_n} C_P(l_\theta)\right\} \\ \leq\ &P^*\left\{\sup_{d(\theta,\theta^*)>\varepsilon/r_n} |(\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta)| > c(\varepsilon/r_n)^2\right\} \leq \frac{\zeta_n(\varepsilon/r_n)\, r_n^2}{c\varepsilon^2\sqrt{n}} = O\left(\varepsilon^{\alpha-2}\right), \end{aligned}$$

Since $\zeta_n(c\delta) \leq c^\alpha \zeta_n(\delta)$ for all $c > 1$ by the assumption on ζ_n, the result follows. For a consistency result, it suffices that $\sup_{\theta\in\Theta} |(\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta)| \to_p 0$ which holds if $\mathcal{L}$ is a Glivenko-Cantelly class (i.e., ULLN holds).

B) For the second part, consider that

$$\begin{aligned} & P^*\left\{r_n \cdot d\left(\widehat{\theta}_n, \theta^*\right) \geq \varepsilon\right\} \leq \Pr\left\{\inf_{r_n d(\theta,\theta^*)>\varepsilon} \mathbb{P}_n\ (l_\theta - l_{\theta^*}) \leq 0\right\} \\ \leq & \sum_{s=S}^{\infty} \Pr\left\{\inf_{2^{s-1}<r_n d(\theta,\theta^*)<2^s} \mathbb{P}_n\ (l_\theta - l_{\theta^*}) \leq 0\right\}, \end{aligned}$$

with $S = \min\{s \geq 1 : 2^s > \varepsilon\}$. The last expression is equal to

$$\begin{aligned} & \sum_{s=S}^{\infty} \Pr\left\{\inf_{2^{s-1}<r_n d(\theta,\theta^*)<2^s} (C_P(l_\theta) - (\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta)) \leq 0\right\} \\ = & \sum_{s=S}^{\infty} \Pr\left\{\inf_{2^{s-1}<r_n d(\theta,\theta^*)<2^s} (\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta) \geq \sup_{2^{s-1}<r_n d(\theta,\theta^*)<2^s} C_P(l_\theta)\right\} \\ \leq & \sum_{s=S}^{\infty} \Pr\left\{\inf_{\frac{2^{s-1}}{r_n}<d(\theta,\theta^*)<\frac{2^s}{r_n}} (\mathbb{P}_n - P)\ (l_{\theta^*} - l_\theta) > c\frac{2^{2s}}{r_n^2}\right\} \\ \leq & \sum_{s=S}^{\infty} \frac{\gamma_n(2^s/r_n)\, r_n^2}{\sqrt{n} c 2^{2s}} = O\left(\sum_{s=S}^{\infty} 2^{(\alpha-2)s}\right) \end{aligned}$$

The expression tends to zero for $\varepsilon \to \infty$ (as $S \to \infty$).

END-PROOF

Remark: By the fixed point refinement obtained from the Talagrand's inequality, teh assumption $d(\theta, \theta^*) \leq c \cdot C_P(l_\theta)$ for all $\theta \in \Theta$ implies that

$$d^2\left(\widehat{\theta}_n, \theta^*\right) \leq c \cdot \varphi_n(r_n^*) = c \cdot r_n^*$$

with probability $1 - \delta$, and $d\left(\widehat{\theta}_n, \theta^*\right) = O_P\left(\sqrt{r_n^*}\right)$.

It is usually easier to obtain faster convergence rates when $\mathcal{L}$ contains smooth functions with moderate complexity. For example, if we study the events (4) and consider the function $l_\theta(X) = H(g_\theta(X)|X) - H(g_\theta(X), f_\theta(X)|X)$, with $\theta \in \Theta \subset \mathbb{R}^K$ and a known smooth cumulative distribution $H(\varepsilon, \epsilon|X)$, the parameter solving $\min_{\theta\in\Theta} P\ l_\theta$ can be studied using the standard theory of M-estimators, and consistency at $\sqrt{n}$ rate can be easily derived. However, the specification of H is a risky assumption for an unobservable variable, by contrast to more flexible semiparametric approaches. When non-smooth functions $\mathcal{L}$ are considered, e.g. if $\mathcal{L} = \{I(A_\theta)(1 - I(B_\theta)) : \theta \in \Theta\}$,

the consistency of $\widehat{\theta}$ is harder to obtain, and even if we can prove consistency the convergence rate can be too slow. In this context, we can consider smooth functions to improve it, e.g. considering the approximation (5).

Unfortunately, models $\mathcal{L}$ with a finite VC dimension are often too small, and we can easily find classes with infinite VC dimension (e.g., when Θ belongs to an infinito-dimensional separable metric space). In these cases, we can consider an approximation method penalizing complexity. A simple alternative is the sieves method, where I consider an increasing sequence a sequence $\{\mathcal{L}_k\} \subset \mathcal{L}$ monotonously increasing to $\mathcal{L}$, and all of the $\mathcal{L}_k$ with finite VC dimension. For example, in the parametric model $\mathcal{L}_\Theta$ I can consider $\{\mathcal{L}_k\}$ as the constraint of $\mathcal{L}$ to Θ_k for a sequence of finito-dimensional subspaces Θ_k increasing to Θ. Thus, the sieves method solves

$$\mathbb{P}_n\, l \to \min, \qquad l \in \mathcal{L}_k.$$

The solution $\widehat{l}_{nk}$ is compared with that of $\{P\, l : l \in \mathcal{L}_k\}$. Let us define

$$C_k(l) = P\, l - \min_{l \in \mathcal{L}_k} P\, l,$$

for all $l \in \mathcal{L}_k$. Notice that the global causality default satisfies,

$$C\left(\widehat{l}_{nk}\right) = \left(\inf_{l \in \mathcal{L}_k} P\, l - P\, l^*\right) + C_k\left(\widehat{l}_{nk}\right),$$

both terms are playing the role of bias and variance, respectively. Note that $\inf_{l \in \mathcal{L}_k} (P\, l - P\, l^*) \to 0$ when $k \to \infty$, and if $\mathcal{L}_k$ satisfies that $E[\log N(\delta, \mathcal{L}_k, \rho_n)] = o(n)$ for all $\delta > 0$, then $C_k\left(\widehat{l}_{nk}\right) \to 0$ almost surely when $k \to \infty$. To ensure that both components tend to zero, typically we require that $k = k_n$ increases with the sample size at a particular rate, and k_n is known as the smoothing number. In particular if $\mathcal{L}_k$ is a VC class $C_k\left(\widehat{l}_{nk}\right) = O_P\left(\sqrt{V_{\mathcal{L}_k}/n}\right)$, so that setting k_n such that $\sqrt{V_{\mathcal{L}_{k_n}}/n} + k_n \to \infty$ when $n \to \infty$ I conclude that $C\left(\widehat{l}_{nk}\right) \to_{a.s.} 0$. For example, in the context of smoothed models like (5), we can consider that $\mathcal{L}$ is indexed by a smoothing parameter h, letting $h \downarrow 0$ with $n \to \infty$ in such way that $nh_n \to \infty$ we can usually obtain consistency, and we can also obtain fast convergence rates for the estimator $\widehat{\theta}_n$ in classes of events $\{(A_\theta \cap B_\theta^c) : \theta \in \Theta\}$ with moderate complexity. This is essentially a semiparametric procedure.

Alternatively, if we consider an upper bound $\pi_k(l) \geq \inf_{l \in \mathcal{L}_k} P\, l - P\, l^*$, then an alternative to the sieves approach is the regularization method, defining $\widehat{l}_{nk}$ as the minimizer of the penalized functional,

$$\mathbb{P}_n\, l + \pi_k(l) \to \min, \qquad l \in \mathcal{L}.$$

These ideas can be connected to the literature on classification theory, see e.g. Boucheron, Bousquet and Lugosi (2005), Devroye et al. (1996, Chapter 18).

CONCLUDING REMARKS

The notion or probabilistic causation is commonly used by scientists and laymen who arguably based their use on data based inferences. Interestingly, causality statements are usually avoided by statisticians who prefer to use association notions. For example, Lindley and Novick (1981) avoid its use because "causality, although widely used does not seem to be well-defined". Speed (1990) suggests that "considerations of causality should be treated as they have always been treated ins statistics: preferable not at all but, if necessary, with very great care." The concept of probabilistic causal inference is perhaps one of the most difficult and important issues in probability and statistics. It has deep consequences for scientific analysis, and for the manipulation of cause factors. This paper presents a valid approach to deal with this topic.

Science is build for the sake of curiosity, as humans usually like to learn about causal relationships. But science is also developed with operational aims, and particularly to obtain optimal results indirectly through the use of causal relationships. Some philosophers even define causality as relationships that are potentially exploitable for purposes of manipulation and control, i.e. A causes B if we can manipulate A to change B. This idea is the cornerstone of manipulability theories of causation. It is also advocated in experimental design, where causal relationships are associated to situations where an effect is manipulated varying the cause factor.

Causality operations can be designed optimally. Consider a family of causal relationships $\{(A_z, B_z)\}_{z\in\mathcal{Z}}$ where $A_z \Rightarrow B_z$ for all $z \in \mathcal{Z}$, and we have the possibility of setting $z \in \mathcal{Z}$. Then we can consider the optimal decision

$$\max\{U(B_z) : z \in \mathcal{Z}\}.$$

where $U(\cdot)$ is a utility function. In the context of probabilistic causation, if we assume that A_z causes B_z almost surely for all $z \in \mathcal{Z}$, and we have the possibility of setting $z \in \mathcal{Z}$, then we can consider the stochastic optimization problem

$$\max\{U(B_z) : z \in \mathcal{Z}, P(A_z \cap B_z^c) = 0\},$$

where $U(\cdot)$ is a utility function for uncertain results, e.g. we can consider a expected utility approach

with $U(B_z) = \int_{B_z} udP$. In a similar way we can consider ε-causality,

$$\max\left\{U(B_z) : z \in \mathcal{Z}, P(A_z \cap B_z^c) \leq \varepsilon\right\}.$$

The numerical solution of these problems can be addressed using stochastic optimization based on scenario approaches.

For example, consider a probability space $(\Omega, \mathcal{F}, P)$ where P has positive measure at a finite number of scenarios $\{(\omega_s, P_s), s = 1, ..., m\}$, then

$$U(B_z) = \sum_{s=1}^{m} u(\omega_s) \cdot \delta_s^{B_z} \cdot P_s.$$

where $\delta_s^B = I(\omega_s \in B)$, and the problem is

$$\max\left\{\sum_{s=1}^{m} u(\omega_s) \cdot \delta_s^{B_z} \cdot P_s : z \in \mathcal{Z}, \sum_{s=1}^{m} \delta_s^{A_z}\left(1 - \delta_s^{B_z}\right) \cdot P_s = 0\right\}.$$

with $I(\omega_s \in A_z \cap B_z^c) = \delta_s^{A_z}\left(1 - \delta_s^{B_z}\right)$. This problem can be solved using numerical algorithms for mixed integer programming.

APPENDIX A: CONDITIONAL PROBABILISTIC CAUSATION

Physical laws can be described at different coarse levels of detail.There might be causal relationships that are valid relatively to a specific level of analytical detail, but not in general. To define finer structures, we can use a conditional probability with respect to some event C with $P(C) > 0$. Therefore, I define:

Definition Conditional causality. Given a probability space $(\Omega, \mathcal{F}, P)$, for any sets $A, B \in \mathcal{F}$ I say A causes B almost surely conditioned to $C \in \mathcal{F}$ with $P(C) > 0$ if, $P(A \cap B^c|C) = 0$ a.s. (or equivalently if $E[I(A) \cdot I(B^c)|C] = 0$ a.s.), and I denote it by $A \overset{a.s.[P|C]}{\Rightarrow} B$.

The condition $P(A \cap B^c|C) = 0$, is equivalent to $P((A \cap C) \cap B^c) = 0$, and whenever $P(A) > 0$ it is equivalent to $P(C \cap B^c|A) = 0$. In other words, if A, C have non null probability, then $A \overset{a.s.[P|C]}{\Rightarrow} B$ if and only if $C \overset{a.s.[P|A]}{\Rightarrow} B$ if and only if $(A \cap C) \overset{a.s.[P]}{\Rightarrow} B$. Therefore conditional probabilistic causality is just a form of unconditional causality, and the use of a conditioning event C simply imposes a restriction $A \cap C$ over the causal event A. However, for testing purposes $E[I(A) \cdot (1 - I(B))|C] = 0$ the use of conditional causality can lead to alternative statistical procedures, particularly when random variables and/or stochastic processes are considered.

The almost surely conditional causality can be relaxed. Given a probability space $(\Omega, \mathcal{F}, P)$, for any sets $A, B \in \mathcal{F}$ and any $\epsilon \in (0,1)$, I say that A causes B in terms of $\epsilon-$probabilistic causality conditioned to $C \in \mathcal{F}$ with $P(C) > 0$, if $P(A \cap B^c|C) \leq \epsilon$. The minimum value ϵ is the C-conditional ε-tolerance. Notice that if A, C have non null probability, then

$$P(A \cap B^c|C) = \frac{P(A \cap C \cap B^c)}{P(C)} = \frac{P(A)}{P(C)} \frac{P(A \cap C \cap B^c)}{P(A)} = \frac{P(A)}{P(C)} P(C \cap B^c|A);$$

so that $P(A \cap B^c|C) \leq \epsilon$ if and only if $P(A \cap C \cap B^c) \leq \epsilon P(C)$, or equivalently if

$$P(C \cap B^c|A) \leq \epsilon P(C) / P(A).$$

Therefore, if $P(A \cap B^c|C) \leq \epsilon$ with $0 < P(C) \leq P(A)$, I conclude that $P(C \cap B^c|A) \leq \epsilon$.

In many instances, ε-probabilistic causality can be only obtained conditional to certain levels of detail. But, since conditional probabilistic causality is just a form of probabilistic unconditional causality from $A \cap C$ to B, I will not stress this case in this paper.

APPENDIX B: ROBUSTNESS

Next we discus the robustness of probabilistic and sequential causality against changes probability measurement. Given a measurable space $(\Omega, \mathcal{F})$, consider two probability measures P, Q. Assume that $A \overset{a.s.[P]}{\Rightarrow} B$, under which conditions is it straightforwardly satisfied that $A \overset{a.s.[Q]}{\Rightarrow} B$ without the need of further checking? The answer lies on the notion of absolutely continuity.

We say that a probability measure Q is absolutely continuous with respect to the probability measure P, both probability distributions defined on $(\Omega, \mathcal{F})$, if $P(A) = 0$ implies that $Q(A) = 0$ for all set $A \in \mathcal{F}$. This is denoted by $Q << P$. Furthermore P and Q are orthogonal if Ω can be partitioned in two disjoint sets Ω_P, Ω_Q such that $P(\Omega_Q) = 0 = Q(\Omega_P)$, and denoted by $P \perp Q$. For any $A, B \in \mathcal{F}$,

- if $Q << P$ then $A \overset{a.s.[P]}{\Rightarrow} B$ implies that $A \overset{a.s.[Q]}{\Rightarrow} B$,
- if $Q << P$ and $P << Q$ then $A \overset{a.s.[P]}{\Rightarrow} B$ if and only if $A \overset{a.s.[Q]}{\Rightarrow} B$

In general P and Q have no need to be neither absolutely continuous nor orthogonal. Assume that Q, P are absolutely continuous with respect to a measure μ defined on $(\Omega, \mathcal{F})$, then the Radon-Nikodym derivatives with respect to μ (densities) p, q exist (where p and q are measurable functions defined from $\Omega \to [0, \infty)$, such that $P(A) = \int_A p d\mu$ and $Q(A) = \int_A q d\mu$ for any set $A \in \mathcal{F}$). The

Lebesgue decomposition states that we can decompose $Q = Q^{ac} + Q^{\perp}$ where $Q^{ac} << P$ and $Q^{\perp}$ is orthogonal with respect to P. It is given by

$$Q^{ac}(A) = Q(A \cap \{p > 0\}), \qquad Q^{\perp}(A) = Q(A \cap \{p = 0\}),$$

for all measurable set A, and the orthogonality is satisfied setting $\Omega_P = \{p > 0\}$ and $\Omega_Q = \{q > 0\}$. It can be proved that

$$Q^{ac}(A) = \int_A \frac{q}{p} dP,$$

where we set $q/p = 1$ if $p = q = 0$ and $p/q = \infty$ if $p > q = 0$. The expression q/p is the likelihood ratio, and it is equal to the Radon-Nikodym derivative dQ^{ac}/dP (Q- a.s.). Clearly, $Q^{\perp}$ is the component that generates incongruences between causality relationships based on Q and P. If $Q << P$ then $Q^{\perp}$ is the null measure and the a.s. causality relationships are invariant to this change of probabilities. Otherwise the maximum distortion is bounded by $Q^{\perp}(\Omega)$. Note that $Q << P$ if and only if $Q(p = 0) = 0$ if and only if $\int (q/p) dP = 1$. Therefore a measurement of relative consistency of P causation with respect to the measure Q is given by

$$T(\Omega) = 1 - \int_{\Omega} (q/p) dP = \int_{\Omega} (1 - (q/p)) dP \geq 0.$$

In particular, assume that $P(A \cap B^c) = 0$, if $T(A \cap B^c) = \int_{A \cap B^c} (1 - (q/p)) dP = 0$ the result is also true under Q.

Next we discus the invariance of sequential causality relationships under changes of probability measurement. Consider a sequence $\{(\Omega, \mathcal{F}_n, P_n)\}_{n=1}^{\infty}$, and $\{A_n\}$ causes B sequentially for $\{P_n\}$. Assume that an alternative researcher develops an alternative theoretical and experimental procedure that generates the sequence $\{(\Omega, \mathcal{F}_n, Q_n)\}_{n=1}^{\infty}$. When can we assure that A causes B sequentially for $\{Q_n\}$, provided that this happens for $\{P_n\}$? The key idea is the notion of contiguity, developed by Le Cam (1960, 1985). A sequence of probability measures $\{Q_n\}$ is contiguous with respect to other sequence $\{P_n\}$ if $\lim_{n\to\infty} P_n(A_n) = 0$ implies that $\lim_{n\to\infty} Q_n(A_n) = 0$, for every sequence of measurable events A_n, and this is denoted by $Q_n \lhd P_n$. I say that P_n and Q_n are mutually contiguous if $Q_n \lhd P_n$ and $P_n \lhd Q_n$, and this is denoted by $Q_n \lhd\rhd P_n$. For any $\{A_n\}$ with $A_n \in \mathcal{F}_n$, and $B_n \in \mathcal{F}$,

- if $Q_n \lhd P_n$ then $\{A_n\}$ asymptotically causes B_n sequentially respect to $\{P_n\}$, implies the same with respect to $\{Q_n\}$
- if $Q_n \lhd\rhd P_n$. then $\{A_n\}$ asymptotically causes B_n sequentially respect to $\{P_n\}$, if and only if the same is true with respect to $\{Q_n\}$.

We can characterize contiguity when there is a measure μ such that P_n and Q_n are absolutely continuous with respect to μ for all n, with $p_n = dP_n/d\mu$ and $q_n = dQ_n/d\mu$. We define $dQ_n/dP_n = q_n/p_n$ if $p_n > 0$, $dQ_n/dP_n = 1$ if $p_n = q_n = 0$ and equal to infinite if $p_n > q_n = 0$. Given a sequence of measurable functions $\{f_n\}, f$ all of them defined from $\Omega \to \mathbb{R}^k$, we denote the weak convergence with respect to the probability measures $\{P_n\}$ by $f_n \longrightarrow_{w.[P_n]} f$. The first Lemma of Le Cam states that the following statements are equivalent, (1) $Q_n \lhd P_n$, (2) If $dP_n/dQ_n \longrightarrow_{w.[Q_n]} U$ along a subsequence, then $P(U > 0) = 1$, (3) If $dQ_n/dP_n \longrightarrow_{w.[P_n]} V$ along a subsequence, then $E[V] = 1$, (4) for any statistics $T_n : \Omega \to \mathbb{R}^k$, if $T_n \xrightarrow{P_n} 0$ then $T_n \xrightarrow{Q_n} 0$. The proof can be found in van der Vaart (1998, Chap. 6).

We define the expression $T_n(A) = \int_A (1 - (q_n/p_n))\, dP_n$. Then $T_n(A \cap B^c)$ is a measurement of relative consistency of $\{P_n\}$ sequential causation with respect to the measures Q_n. The sequence $T_n(\Omega) \longrightarrow_{w.[P_n]} T = 1 - E[V] = 0$ iff $Q_n \lhd P_n$.

REFERENCES

Arntzenius, F. (1993) "The Common Cause Principle," in: Hull, D., M. Forbes, and K. Okruhlik eds., 1993, PSA 1992, Volume Two. East Lansing: Philosophy of Science Association, pp. 227 - 237.

Bollen, K. A. (1989): Structural Equations with Latent Variables, J. Wiley and Sons, New York.

Boucheron, S., O. Bousquet, and G. Lugosi, (2005): "Theory of Classification: a Survey of Recent Advances", ESAIM: Probability and Statistics, 9, 323–375.

Cartwright, N. (1979): "Causal Laws and Effective Strategies," Noûs, 13, 419-437.

Cliff, N. (1983): "Some cautions concerning the application of causal modelling," Multivariate Behavioural Research, 18, 115-126.

Cochran, W. and G.M. Cox (1957): Experimental Designs (2nd ed.). Wiley, New York.

Collingwood, R.(1940): An Essay on Metaphysics. Oxford: Clarendon Press.

Dawid, A. P. (2004): "Probability, Causality and the empirical world: A Bayes-de Finetti-Popper-Borel sinthesis", Statistical Science, 19, 1, 44-57.

Deroye, L., L. Györfi, and G. Lugosi (1996): A probabilistic theory of pattern recognition, Springer Verlag, New York.

Deroye, L. and G. Lugosi (2001): Combinatorial methods in density estimation, Springer Verlag, Berlin.

Dudley, R. M. (1978): "Central limit theorems for empirical measures," Annals of Probability, 6,

899-929.

Duncan, O. D. (1966): "Path Analysis: Sociological Examples," American Journal of Sociology, 72, 1-16.

Earman, John. (1986): A Primer on Determinism. Dordrecht: Reidel.

Eells, E. (1991): Probabilistic Causality. Cambridge, U.K.: Cambridge University Press.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983): "Exogeneity". Econometrica, 51, 277–304

Gasking, D. (1955): "Causation and Recipes", Mind, 64, pp. 479-487.

Geweke, J. (1984): "Inference and Causality", In: Handbook of Econometrics (ed. Z Griliches and M. D. Intrilligator), Chap. 19, Vol. 2, Elsevier, Amsterdam.

Granger, C.W.J. (1969): "Investigating causal relation by econometric and cross-sectional method", Econometrica, 37, 424–438.

Haussler, D. (1995): "Sphere packing nmbers for subsets of the Bolean n-cube with bounded Vapnik-Chervonenkis dimension," Journal of Combinatorial Theory, Series A, 69, 217-232.

Hausman, D. (1998): Causal Asymmetries. Cambridge: Cambridge University Press.

Hausman, D. and Woodward, J. (1999): "Independence, Invariance, and the Causal Markov Condition," British Journal for the Philosophy of Science 50: 1 - 63.

Hitchcock, C. (1993): "A Generalized Probabilistic Theory of Causal Relevance," Synthese 97: 335-364.

Holland (1988): "Causal inference, Path analysis and Recursive Structural Equation Models," in C. C. Clogg (Ed.): Sociological Methodology, Americal Sociological Association, Washighton D. C., pp. 449-493.

Hume, D. (1748): An Enquiry Concerning Human Understanding.

Humphreys, P. (1989): The Chances of Explanation: Causal Explanations in the Social, Medical, and Physical Sciences, Princeton: Princeton University Press.

Jöreskog, K. G. (1973): "A General Method for Estimating a Linear Structural Equation System", in: Structural Equation Models in Mathematical Psychology, Vol. 2, Goldberger, A. S. and O. D. Duncan eds, Seminar, New York, pp. 85-112.

Jöreskog, K. G. (1978): "Statistical Analysis of Covariance and Correlation Matrices", Psychometrica 43, 443-447.

Kvart, I. (1997): "Cause and Some Positive Causal Impact," Philosophical Perspectives, 11, 401-432.

Koltchinskii, V. (2006): "Local rademacher complexities and oracle inequalities in risk minimiza-

tion", The Annals of Statistics, 34, 2593-2656.

Le Cam, L. (1960): Local asymptotically normal families of distributions. University of California Publications in Statistics, 3, 37-98.

Le Cam, L. (1985): Asymptotic methods in statistical decision theory, Springer Verlag, New York.

Lindley, D. V. and M. R. Novick (1981): "The role of excheangeability in inference", Annals of Statistics, 9, 45-58.

Lugosi, G. (2002): "Pattern classification and learning theory," In: Principles of Nonparametric Learning, (L. Györfi ed.), Springer Verlag, Viena, pp. 1-56.

Lewis, D. (1986): Philosophical Papers, Volume II. Oxford: Oxford University Press.

Mackie, J. (1974): The Cement of the Universe. Oxford: Clarendon Press.

McDiarmid, C. (1989): "On the method of bounded differences", In: Surveys in Combinatorics, Cambridge Univerity Press, Cambridge, pp. 148-188.

Massart, P. and É. Nédélec (2006): "Risk bounds for statistical learning", The Annals of Statistics, 34, 2326-2366.

Menzies, P. and Price, H. (1993): "Causation as a Secondary Quality," British Journal for the Philosophy of Science, 44, pp. 187-203.

Noordhof, P. (1999): "Probabilistic Causation, Preemption and Counterfactuals," Mind 108: 95 - 125.

Papineau, D. (1993): "Can We Reduce Causal Direction to Probabilities?" in in: Hull, D., M. Forbes, and K. Okruhlik eds., 1993, PSA 1992, Volume Two. East Lansing: Philosophy of Science Association, pp. 238-252.

Pearl, J. (1995): "Causal diagrams for empirical research", Biometrica, 669-688.

Pearl, J. (1999): "Reasoning with Cause and Effect," in Proceedings of the International Joint Conference on Artificial Intelligence (San Francisco: Morgan Kaufman), pp. 1437 - 1449.

Pearl, J. (2000): Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Pollard, D. (1984): Convergence of stochastic processes, Springer, News York.

Popper, K. R. (1959): The logic of scientific discovery, Hutchinson, London, (German original published in 1934).

Popper, K. R. (1983): Realism and the aim of science, Hutchinson, London.

Price, H. (1991): "Agency and Probabalistic Causality", British Journal for the Philosophy of Science 42: 157 -76.

Reichenbach, H. (1956): The Direction of Time. Berkeley and Los Angeles: University of California Press.

Russell, B. (1913): On the Notion of Cause. Proceedings of the Aristotelian Society 13: 1-26.

Russell, B. (1948): Human Knowledge. New York: Simon and Schuster.

Scheines, R. (1997): "An Introduction to Causal Inference," in: McKim, V., and S. Turner eds., 1997, Causality in Crisis?, Notre Dame: University of Notre Dame Press, pp. 185 - 199.

Skyrms, B. (1980): Causal Necessity. New Haven and London: Yale University Press.

Sobel, M. E. (1995): "Causal Inference in the Social and Behavioral Sciences," in G. Arminger, C. C. Clogg and M. E. Sobel (Eds.), Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York: Plenum Press, pp. 1-38.

Sobel, M. E. (1998): "Causal Inference in Statistical Models of the Process of Socioeconomic Achievement," Sociological Methods and Research, 27, 2, 318-348.

Speed, T. P. (1990): "Complexity, calibration and causality in influence diagrams". In: R. M. Oliver and J. Q. Smith (eds.), Influence Diagrams, Belief Nets and Decisions Analysis, Wiley, New York, pp. 49-63.

Spirtes, P., C. Glymour, and R. Scheines. (2000): Causation, Prediction and Search, Second edition. Cambridge, MA: M.I.T. Press.

Suppes, P. (1970): A Probabilistic Theory of Causality. Amsterdam: North-Holland Publishing Company.

Talagrand, M. (1996a): "A new look at independence", Annals of Probability, 24, 1-34.

Talagrand, M. (1996b): "New concentration inequalities in product spaces", Inventiones Mathematicae, 126, 505-563.

van der Vaart, A. W. (1998): Asymptotic Statistics, Cambridge University Press, Cambidge, UK.

van der Vaart, A. W. and J. A. Wellner (1996): Weak convergence and empirical processes, Springer Verlag, New York.

Vapnik, V. and A. Chervonenkis (1971): "On the uniform convergence of relative frequencies of events to their probabilities", Theory of Probability and its Applications, 16, 264-280.

Vapnik, V. and A. Chervonenkis (1974): Theory of Pattern Recognition, Nauka, Moscow (in Russian)

Vapnik, V. (1998): Statistical Learning Theort, Wiely, New York.

von Mises, R. (1939): Probability, Statistics and Truth. Hodge, London, (German original published in 1928).

von Wright, G.(1971): Explanation and Understanding. Ithica, New York: Cornell University Press.

Whittaker, J. (1990): Graphical models in Applied Multivariate Statistics, Wiley, New York.

Wiener, N. (1956): "The Theory of Prediction". In: E. F. Beckenback, e., Modern Mathematics for Engineers.

Wold, H (1954): "Causality and Econometrics". Econometrica, 22, (2), 162-77.