Coarse-grained reliability

Paweł Kawalec[1]

Center for Philosophy of Science, University of Pittsburgh

Wydział Filozofii, Katolicki Uniwersytet Lubelski

1. Introduction

Despite the growing interest in the now widely applied graphical models of causality few realize that it stems from Andrzej Mostowski's work on recursively enumerable models. He raised the following problem in (1955):

> The formula to be given below was found in the course of unsuccessful attempts to construct a formula no model of which would belong to the smallest field of sets generated by the classes $P_1^{(n)}$ and $Q_1^{(n)}$. (Mostowski 1955, 125)

In order to come up with an answer to this problem Hilary Putnam developed a framework for is now acknowledged as the logical norms for reliable knowledge acquisition.

In this section I elucidate how the reliabilistic framework enabled Putnam to come up with a negative answer (1965) to Mostowski's problem. Elaborating Hans Reichenbach's straight rule Putnam proposed the following modification of a decision procedure for the recursive sets:

> (1) allowing the procedure to "change its mind" any finite number of times …

(2) we give up the requirement that it be possible to tell (effectively) if the computation has terminated. (Putnam 1965, 49)

Such sets, following Mark E. Gold (1965), are called limiting recursive and are extensions of Putnam's trial and error predicates (Glymour 1996, 272).

Definition 1. (*k*-trial predicates, Putnam 1965)

P is a k-*trial error predicate* iff there is a general recursive function *f* such that for all $x_1, x_2, \ldots, x_n$

(1) $P(x_1, x_2, \ldots, x_n) \equiv \lim_{y\to\infty} f(x_1, x_2, ..., x_n, y) = 1$ and

(2) there are *at most* *k* integers *y* such that $\lim_{y\to\infty} f(x_1, x_2, ..., x_n, y) \neq \lim_{y\to\infty} f(x_1, x_2, ..., x_n, y+1)$;

where $\lim_{y\to\infty} f(x_1, x_2, ..., x_n, y) = k \Leftrightarrow \exists_y \forall_z (z \geq y \supset f(x_1, x_2, ..., x_n, z) = k)$.

In this framework Mostowski's problem becomes the search for "a formula with no model in which (1) the universe of discourse is the natural numbers; and (2) the predicate letters are all interpreted as recursively enumerable predicates, or truth functions (i.e. Boolean combinations) of predicates" (Putnam 1965, 50).

Theorem 1. P *is a* k-*trial predicate iff* $P \in \Delta_2 = \Sigma_2 \cap \Pi_2$ (a set is limiting recursive iff it is in Δ_2)

Proof. Analogously to (Putnam 1965, 51-52)

---

[1] I am grateful to the Foundation for Polish Science for support of my research while preparing this text.

Theorem 2. *There exists a k such that* P *is a k-trial predicate iff* P *belongs to* $\sum_1^*$ (*i.e. the smallest class containing the recursively enumerable predicates and closed under truth functions*).

Proof. (Putnam 1965, 52-53)

Putnam's answer to Mostowski's problem (1965):

Theorem 3. *Every consistent formula of quantification theory without identity has a model in* $\sum_1^*$.

Proof. (Putnam 1965, 53-55)

2. Reliability and scientific discovery

Clark Glymour succinctly words a question that one should ask at this point:

> How does one get from the characterization of the limiting recursive sets of numbers to an understanding of empirical questions for which discovery methods do and do not exist? (Glymour 1996, 274)

If the information contained in the upcoming evidence is coded numerically, we obtain a data stream ε, a finite portion of which has been observed. The subsequent work removed Putnam's constraint that the ordering of the data could not somehow restricted by the background knowledge and that there could be alternative ways for a hypothesis to be correct with regard to the recorded data.

Kevin Kelly's book *The logic of reliability* (1996) is the climax of subsequent generalizations of Putnam's work. An inductive problem ($\mathcal{C}, \mathcal{K}, \mathcal{M}, \mathcal{H}$) specifies $\mathcal{C}$ – a correctness relation between data and hypotheses h $\in \mathcal{H}$, $\mathcal{K}$ – scientist's background knowledge (a set of possible data streams ε) and $\mathcal{M}$ – a set of assessment methods admissible for a scientists.

There is a number of ways to learn – if possible – the correct answer to a given problem.

Definition 2. $\mathcal{H}$ is verifiable/refutable/decidable by *n* given $\mathcal{K}$ by a method in $\mathcal{M}$ iff there is an $\alpha \in \mathcal{M}$ such that α verifies/refutes/decides $\mathcal{H}$ by *n* given $\mathcal{K}$.

Definition 3. α verifies $\mathcal{H}$ by *n* given $\mathcal{K}$ iff for all h $\in \mathcal{H}$ α produces 1 at *n* on h, $\varepsilon \equiv \mathcal{C}(\varepsilon, h)$.

Definition 4. α refutes $\mathcal{H}$ by *n* given $\mathcal{K}$ iff for all h $\in \mathcal{H}$ α produces 0 at *n* on h, $\varepsilon \equiv \neg\mathcal{C}(\varepsilon, h)$.

Definition 5. α decides $\mathcal{H}$ by *n* given $\mathcal{K}$ iff for all h $\in \mathcal{H}$ α both verifies and refutes h by *n* on ε.

A particular application of Kelly's work is to the inference in the empirical sciences.

Proposition 2.

If h is structural and $\mathcal{K}$ entails plenitude and satisfies D1-D5,[2] then

h is computably verifiable/refutable/decidable in the limit given $\mathcal{K}$ iff

$\mathcal{D}ata_{\mathcal{K}, C, \zeta}(\text{h}) \in \Sigma[\mathcal{D}ata_{\zeta}(\mathcal{K})]^{A}_{2} / \Pi[\mathcal{D}ata_{\zeta}(\mathcal{K})]^{A}_{2} / \Delta[\mathcal{D}ata_{\zeta}(\mathcal{K})]^{A}_{2}$.

3. Reliability and causal inference.

Starting with the work of Glymour and his co-workers (Glymour et al. 1987) the reliabilistic framework has been successfully applied to the formalization of causal learning and causal discovery from data (Spirtes, Glymour and Scheines 2000; Pearl 2000). There are two major assumptions underlying reliabilist approach to causal inference: the Causal Markov Assumption and the Faithfulness Assumption.

Let random variable W be a vertex in an acyclic directed graph G. V is a parent of W if an edge $V \rightarrow W$ in G. Z is a descendant of W if there is a directed path from W to Z in G. Let $\boldsymbol{PA}(W)$ be the set of parents of W and $\boldsymbol{DE}(W)$ be the set of descendants of W.

Definition 6. (Markov Assumption)

A directed graph G over $\boldsymbol{V}$ and a probability distribution $P(\mathbf{V})$ satisfy the Markov Assumption iff for every W in $\boldsymbol{V}$, $W \perp\!\!\!\perp \boldsymbol{V} \backslash (\boldsymbol{DA}(W) \cup \boldsymbol{PA}(W)) \mid \boldsymbol{PA}(W)$.

Let $\mathcal{P}(G)$ denote all distributions that satisfy the Markov Assumption with G and $\mathcal{F}(P)$ represent all independence and conditional independence relations for the variables in $\boldsymbol{V}$ under P. Let $\mathcal{F}_G = \cap_{Q \in \mathcal{P}(G)} \mathcal{F}(Q)$ be all independence relationships common to all distributions in $\mathcal{P}(G)$.

Definition 7. (Faithfulness Assumption)

A probability distribution $P(\mathbf{V})$ satisfies the Faithfulness Condition with regard to G iff $\mathcal{F}_G = \mathcal{F}(P)$.

The Faithfulness Assumption is also called 'stability' (Pearl 2000) as it may be interpreted as representing the fact that conditional independencies in P are due to causal structure and not casual configurations of the values of parameters in G.

We now turn to PC and FCI algorithms for causal discovery which implement the assumptions.

Let $\boldsymbol{ADJ}(C, A)$ be the set of vertices adjacent to A in directed acyclic graph C.

PC algorithm

i) Form the complete undirected graph C on the vertex set $\boldsymbol{V}$.

ii)

$n = 0$

repeat

repeat

[2] For articulation see (Kelly 1996, 366).

select an ordered pair of variables X and Y that are adjacent in C such that $\boldsymbol{ADJ}(C, X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset $\boldsymbol{S}$ of $\boldsymbol{ADJ}(C, X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given **S** delete edge between X and Y from C and record $\boldsymbol{S}$ in $\boldsymbol{Sepset}(X, Y)$ and $\boldsymbol{Sepset}(Y, X)$;

until all ordered pairs of adjacent variables X and Y such that $\boldsymbol{ADJ}(C, X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets $\boldsymbol{S}$ of $\boldsymbol{ADJ}(C, X)\backslash\{Y\}$ of cardinality n have been tested for d-separation;

$n = n + 1$;

until for each ordered pair of adjacent vertices X, Y, $\boldsymbol{ADJ}(C, X)\backslash\{Y\}$ is of cardinality less than n.

iii) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient $X – Y – Z$ as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in $\boldsymbol{Sepset}(X, Z)$.

iv) repeat

If $X \rightarrow Y$, Y and Z are adjacent, A and C are not adjacent, and there is no arrowhead at Y, then orient $Y – Z$ as $Y \rightarrow Z$.

If there is a *directed* path from X to Y, and an edge between X and Y, then orient $X – Y$ as $X \rightarrow Y$;

until no more edges can be oriented.

The PC algorithm requires that *the causal sufficiency* be satisfied: $\boldsymbol{V}$ is causally sufficient for a given population iff in the population any common cause of two variables in $\boldsymbol{V}$ is also in $\boldsymbol{V}$ or has the same value for all units in the population. Thus, no latent variables are allowed.

This requirement is alleviated in the FCI algorithm.

FCI algorithm

Step (i) and (ii) are as in the PC algorithm.

iii) Let F be the undirected graph resulting from step (ii). Orient each edge as o–o. For each triple of vertices X, Y, Z such that the pair A, B and the pair B, C are each adjacent in F but the pair A, C are not adjacent in F, orient $A \ast\!-\!\ast B \ast\!-\!\ast C$ as $A \ast\!\!\rightarrow B \leftarrow\!\!\ast C$ if and only if B is not in ***Sepset***(A, C).

iv) For each pair of variables A and B adjacent in F, if A and B are d-separated given any subset $\boldsymbol{S}$ of ***Possible-d-sep***(A, B)\{A, B} or any subset $\boldsymbol{S}$ of ***Possible-d-sep***(B, A)\{A, B} in F remove the edge between A and B, and record $\boldsymbol{S}$ in ***Sepset***(A, B) and ***Sepset***(B, A).

v) repeat

If there is a directed path from A to B, and an edge $A \ast\!-\!\ast B$, orient $A \ast\!-\!\ast B$ as $A \ast\!\!\rightarrow B$, else if B is a collider along $\langle A, B, C \rangle$, B is adjacent to D, and D is in ***Sepset***(A, C), then orient $B \ast\!-\!\ast D$ as $B \leftarrow\!\!\ast D$,

else if U is a definite discriminating path between A and B for M, and P and R are adjacent to M on U, and $P - M - R$ is a triangle, then

if M is in ***Sepset***(A, B) then M is marked as a noncollider on subpath $P \ast\!-\!\ast M \ast\!-\!\ast R$, else $P \ast\!-\!\ast M \ast\!-\!\ast R$ is oriented as $P \ast\!\!\rightarrow M \leftarrow\!\!\ast R$.

else if $P \ast\!\!\rightarrow M \ast\!-\!\ast R$ then orient as $P \ast\!\!\rightarrow M \rightarrow R$.

until no more edges can be oriented.

4. Causal discovery and intervention data

In (2004) Rubin comes up with an example which is intended to demonstrate that any causal discovery method which relies on conditional independencies is not reliable as it is guaranteed not to arrive at the true causal structure.[3] A method of operation W (either new ($W = 1$) or old ($W = 0$)) is applied to patients with different age X as the only measured covariate and the success is the method is evaluated in terms of years of survival Y after the operation is performed. Rubin's data for this example are displayed in Table 1.

Covariate X	W	Potential outcomes $Y(0)$	Potential outcomes $Y(1)$	Individual causal effects $Y(1) - Y(0)$
68	1	13	14*	+1
76	0	6*	0	−6
66	0	4*	1	−3
81	0	5*	2	−3
70	0	6*	3	−3
72	0	6*	1	−5
81	1	8	10*	+2
72	1	8	9*	+1
True averages		7	5	−2
Observed averages		5.4*	11*	

*Observed values. (Y, years lived postoperation; X, age at start of study).

Table 1. (Rubin 2004)

[3] For general results on reliability of Bayes nets approach to causal inference see (Robins et al. 2003).

Potential outcomes[4] data are the years of survival given a different method of operation would be applied to the same patient. Potential outcomes cannot be observed, so they are averages from a subclass of the general population with regard to age. The average difference between (counterfactual) effects of new and old operation $Y(1) - Y(0)$ is negative –2. PC algorithm for the observed values estimates the effect of new operation $W = 1$ on survival Y to be positive with coefficient 5.9.

For data in Table 1 PC finds coefficients with the opposite signs, while FCI fails to orient the links between the variables.

One way out would be to take advantage of the background knowledge and claim that with *U* being a common cause of *W* and *Y* has different values for the sample and the general population. This will, however, violate the causal sufficiency.

I contend to Rubin's conclusion to the extent that a constraint needs to be imposed on the kind of data for a causal discovery algorithm.

In the remainder I suggest a different solution which turns on the kind of data.

Definition 8. (Active interference)

A researcher's act A (= 1, 0) interferes actively *with* (*a phenomenon represented by*) *a variable* X (*with states* $x_1, ..., x_n$ *in the probability space* $\mathcal{X}$) *iff* $P(X = x_i | A = 1) - P(X = x_i | A = 0) = 1$ (*where* $x_i \in \mathcal{X}_i \subset \mathcal{X}$ *for* $1 \leq i \leq n$).

If there is an active interference *A* on a variable *X* then *X* is called *manipulable variable*.

[4] The potential variables were first developed by Jerzy Neyman see (Neyman 1923/1990).

Definition 9. (Intervention data)

A set $\mathcal{I}_{\mathrm{X}}^{\mathrm{V}} = \times_{i \in \mathrm{X}} \mathcal{V}_i \times_{j \in \mathrm{V} \setminus \mathrm{X}} \mathcal{V}_j$ *of vectors of the values of variables* $\mathrm{x}_i \in \mathbf{X} \subset \mathbf{V}$ *and* $\mathrm{v}_j \in (\mathbf{V} \setminus \mathbf{X}) \subset \mathbf{V}$ *from the domain* D *recorded when A actively interferes on the variables in* **X** *is called* intervention data on **X**.

Definition 10. (Intervention field)

The intervention field *in a domain* D *for a variable* Y *is a set* PI = (<(X_1, ..., X_n), Y>) *such that for variables* $(\mathrm{X}_1, ..., \mathrm{X}_n) \subseteq \mathrm{X}$ *with intervention data* $\mathcal{I}_{\mathrm{X}}^{\mathrm{V}}$ *it holds that* $f(\mathrm{Y} \mid \mathrm{V}) = f(\mathrm{Y} \mid (\mathrm{X}_1, ..., \mathrm{X}_n))$ *where* $f(\mathrm{X}_i)$ *is a density function*.

The minimal intervention field for a variable *Y* with regard to variable *X* is thus the intervention field PI = (<X, Y>).

Proposition 3.

For a manipulable variable X *the edge* X → Y *is invariant between intervention field and the minimal intervention field.*

Definition 11. (Expansion of the minimal intervention field)

The expansion of the minimal intervention field PI = <X, Y> *is an intervention field* PI′ = <X′, Y′> *such that* R(X′, X) *and* R(Y′, Y), *where* R *is a part-whole relation or its converse and there is intervention data* $\mathcal{I}_{\mathrm{X}'}$ *for the set* X′ ⊇ (X′, Y′).

Definition 12. (Procedural criterion of causal dependence in a domain)

The dependence X → Y *is* causal in domain D *iff for intervention data interwencyjnych* ι_{X}^{V} *in the minimal intervention field* Y *for* X *and values of variables in* **V** *in D it holds that* f(Y | V) = f(Y | X) *or for intervention data* X´ *in the expanded minimal intervention field* Y´ *for* X´ *it holds that* f(Y´ | X´, V) = f(Y´ | X´), *when* X *is not manipulable variable*.

The procedural criterion leads us to transform Rubin's data displayed in Table 1 into two different sets of data for possibly two different populations with regard to an unknown variable *U* (or set of variables) as displayed in Table 2.

a)

X	W	Y
66	0	4
68	1	14
70	0	6
72	0	6
72	1	9
76	0	6
81	0	5
81	1	10

b)

X	W	Y
66	1	1
68	0	13
70	1	3
72	1	1
72	0	8
76	1	0
81	1	2
81	0	8

Table 2. a) intervention data for 8 persons submitted to the medical trial with *W* = 0, 1; b) data for the population.

PC algorithm for the intervention data yields the correct answer of there being a negative effect of the new operation on survival with coefficient –2.

REFERENCES

Glymour C., 1996, The hierarchies of knowledge and the mathematics of discovery, in P. Millikan and A. Clark (ed.), *The legacy of Alan Turing*, Oxford: Clarendon Press, p. 265-91.

Glymour C., Scheines R., Spirtes P., Kelly K., 1987, *Discovering causal structure: artificial intelligence, philosophy of science, and statistical modeling,* New York: Academic Press.

Jain S., Osherson D., Royer J., Sharma A., 1999, *Systems that learn: an introduction to learning theory,* A Bradford Book/Cambridge, MA.: MIT.

Juhl C., 1993, Bayesianism and reliable scientific inquiry, *Philosophy of Science,* 60, p. 302-19.

Kelly K. T., 1996, *The logic of reliable inquiry,* New York: Oxford University Press.

Mostowski A., 1955, A formula with no recursively enumerable model, *Fundamentae Mathematicae*, 43, p. 125-40.

Neyman J., 1923/1990, *On the application of probability theory to agricultural experiments. Essay on principles. Section 9.,* tłum. M. Dabrowska, T. P. Speed, 5, p. 465-72.

Pearl J., 2000, *Causality: models, reasoning, and inference,* Cambridge: Cambridge University Press.

Putnam H., 1965, Trial and error predicates and the solution to a problem of Mostowski, *Journal of Symbolic Logic*, 30, p. 49-57.

Robins J. M., Scheines R., Spirtes P., Wasserman L., 2003, Uniform consistency in causal inference, *Biometrika* 90, p. 491-515.

Rubin D. B., 2004, Direct and indirect causal effects via potential outcomes, *Scandinavian Journal of Statistics*, 31, p. 161-70.

Spirtes P., Glymour C., Scheines R., 2000 2nd ed., *Causation, prediction, and search*, Cambridge: MIT.