Weak Predictivism, Ad hoc Modification, . . . , and Intervention

Christian J. Feldbacher-Escamilla

Spring 2015

Project Information

Talk(s):

- Feldbacher-Escamilla, Christian J. (2015-04-20/2015-04-24). *Weak Predictivism, Ad Hoc-Modification, . . . , and Intervention*. Workshop. Presentation (invited). Workshop with Christopher Hitchcock. University of Düsseldorf: Düsseldorf Center for Logic and Philosophy of Science (DCLPS).

Motivation

A (not only) practical point of view: Complexity $\Downarrow$ $\Rightarrow$ *Predictability* $\Uparrow$

A first-sight tension between intervention and predictability:

- On the one hand, surplus knowledge on causal manipulability (intervention) allows for extra predictability.
- On the other hand, modifying a causal system by implementing intervention seems to increase its complexity ...
- ... and by this predictability seems to be decreased.

Aim of this talk: Provide an analysis of this tension in an information theoretic framework.

Contents

Predictivism

The Problem of Novel Predictions

A paradox of confirmation (cf. Menke 2009, p.7):

- (es) Some hypotheses can be confirmed by novel and well-established facts.
- (es) From a practical point of view hypotheses are more confirmed by novel facts than by well-established facts. That is: if $ass_G(T, D) = x$, given $Novel(D)$, and $ass_G(T, D) = y$, given $\sim Novel(D)$, then $x > y$.
- (es) From a logical point of view novelity is only of historical interest and does not affect confirmation. That is: if $ass_G(T, D) = x$, given $Novel(D)$, and $ass_G(T, D) = y$, given $\sim Novel(D)$, then $x = y$.

Problem: 1–3 are incompatible.

The Problem of Novel Predictions

There is also the following possibility:

- If $ass_G(T, D) = x$, given $Novel(D)$, and $ass_G(T, D) = y$, given $\sim Novel(D)$, then $x < y$.

The main positions in the debate:

Accommodationism:	4
Neutralism:	3
Predictivism:	2

A Taxonomy

Also relevant are:

1. the quantification over T: strong ($\forall T \ldots$), weak ($\exists T \ldots$ & $\sim\forall T \ldots$)
2. the interpretation of G: epistemic or instrumentalistic
3. the Interpretation of *Novel*:

temporal	Fact D was un-/known at the formulation of T.
heuristic	Fact D was un-/used for the formulation of T.
theoretical	Fact D was un-/explained at the formulation of T by every/a rival of T.

The categories in 1 und 2 are exclusive.

In 3 no category is a subcategory of any other. (But they are compatible.)

In toto there are 36 possible positions according to this taxonomy.

(Hitchcock and Sober 2004) argue for a: Heuristic instrumentalistic weak predictivism

Heuristic Instrumentalistic Weak Presdictivism

$\exists G \in \mathit{instrval}\ \exists T, D \ldots \forall x, y$: If $\mathit{ass}_G(T, D) = x$, given $\mathit{Novel}(D)$, and $\mathit{ass}_G(T, D) = y$, given $\sim \mathit{Novel}(D)$, then $x > y$.

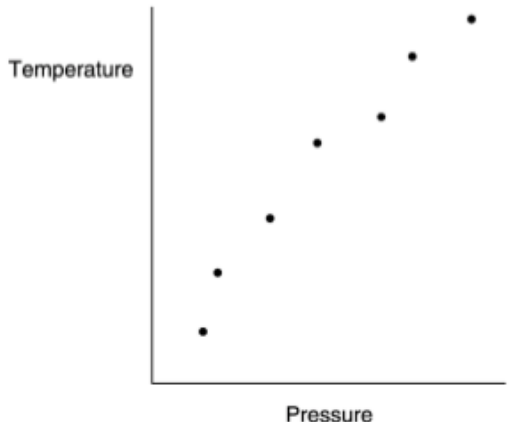
(Hitchcock and Sober 2004) interpret *Novel* here heuristically.

Technically this can be expressed by splitting up the available evidence or data before formulating a hypothesis or theory $D = D_1 \cup D_2$ (where $D_1 \cap D_2 = \emptyset$) and using only D_1 for the formulation.

The predictivistic problem can be reduced then to a question of curve-fitting: Why is it not always adequate to fit a hypothesis or theory exactly to the data ($D = D_1 \cup D_2$), but only approximately (or partially: D_1)?

Heuristic Instrumentalistic Weak Presdictivism

Closely related: Why, e.g., to use only a linear model ($y = a_1 \cdot x^1 + a_0$) for fitting (cf. Sober 2008) ...



... instead of, e.g., a degree 6 polynomial one fitting exactly:

$$y = a_6 \cdot x^6 + \cdots + a_1 \cdot x^1 + a_0$$

($a_0, \ldots, a_6$ ... parameters of model ... polynomial of degree 6)

There is an information theoretical answer: Akaike framework

Akaike Information Framework

The Framework

General intuition: We do not only aim at true (accurate) models, but also at informative (explanatory/predictive successful) ones.

Take, e.g., one of Popper's critique of confirmation theory: The most probable theories are the ones with least (empirical) content.

Therefore the choice of "his" information measure: $inf(T) = 1 - Pr(T)$

E.g.: $inf(\top) = 0$ (explains nothing), whereas $inf(\bot) = 1$ (explains everything)

This tension between truth (accuracy) and informativity (explanatory/ predictive success) is investigated/explained in the Akaike framework by reference to the problem of overfitting.

The Framework

Idea (cf. Forster and Sober 1994):

- Data is noisy and involves error.
- An accurate fit to the data fits also error (overfits).
- Whereas a less accurate fit may depart from the error.
  ("Closeness to the truth is different from closeness to the data.")
- Fact: The more parameters a model has, the more prone it is to overfit.
- "Hence": Simplicity may (w.r.t. truth) account for inaccuracy (w.r.t data)

Theorem (Akaike's Theorem, (cf. Forster and Sober 1994))

Estimated predictive accuracy of model M given data D $(AIC(M, D)) = \frac{1}{N} \cdot (log(Pr(D|L(M))) - k(M))$

Where: N ... sample size ($|D|$), $L(M, D)$ ... most accurate parametrisation of M w.r.t. D, $k(M)$ ... number of parameters of M.

Application to Weak Predictivism

We are now ready to apply this result to weak predictivism:

Consider the following cases (cf. Hitchcock and Sober 2004):

1. We know $M_a = M_p$: $ass_G(M_a, D) = ass_G(M_p, D)$
2. We know the exact form of M_a, M_p: $ass_G(M_a, D) \propto AIC(M_a, D)$, $ass_G(M_p, D) \propto AIC(M_p, D)$
3. ...
4. ...
5. We know that $Pr(D|M_p) \uparrow$, $Pr(D|M_a) \approx 1$, and M_p was formulated on basis of $D_1 \subset D$ only, whereas M_a was formulated on basis of D: $ass_G(M_a, D) < ass_G(M_p, D)$

Application to Weak Predictivism

Hitchcock and Sober's argument for the assessment in case 5:

1. $Pr(D|M_p) \uparrow, Pr(D|M_a) \approx 1$ and M_p formulated on $D_1 \subset D$, M_a formulated on D (assumption of the case)
2. $Pr(\textit{accommodation of } D \textit{ by } M|M \textit{ is balancing}) < Pr(\textit{accommodation of } D \textit{ by } M|M \textit{ is fitting})$ (assumption)
3. $Pr(\textit{prediction of } D_2 \textit{ by } M|M \textit{ is balancing}) > Pr(\textit{prediction of } D_2 \textit{ by } M|M \textit{ is fitting})$ (assumption)
4. $Pr(M_p \textit{ is balancing}) > Pr(M_p \textit{ is fitting})$ (1,3,...)
5. $Pr(M_a \textit{ is fitting}) > Pr(M_a \textit{ is balancing})$ (1,2,...)
6. Probably: $k(M_p) < k(M_a)$ (4,5)
7. Probably: $AIC(M_p, D) > AIC(M_a, D)$ (1,6, Akaike's Theorem)
8. $ass_G(M_p, D) > ass_G(M_a, D)$ (7)

Application to Weak Predictivism

Note that in the argument *balancing / fitting* serves as an indicator (instrument) for figuring out *AIC* comparatively.

And by this also as an instrument for the epistemic goal G in $ass_G(M_p, D) > ass_G(M_a, D)$.

So, there seems to be a reasonable context where prediction instrumentally exceeds accommodation in theory assessment.

Hence: weak (instrumentalistic and heuristic) predictivism

Application to Other Problems: Ad Hoc Modifications

Within the Akaike framework also other questions can be easily addressed (cf. Forster and Sober 1994):

E.g.: The problem of characterising ad hoc modifications

Popper's proposal: ad hoc modifications are those whose empirical content decreases in reaction to a falsification (for problems cf. u.a. (Grünbaum 1976)).

Forster and Sober's explication of the follow up proposal of Lakatos (innovative vs. degenerative research programmes): *AIC*-balancing:

Definition

A research programme is degenerative iff a loss in simplicity of the programme's core is not compensated by a sufficient gain in fit with data according to AIC (negative AIC development).

Application to Other Problems: Unification

A similar application may be performed in the debate about unification (cf. Forster and Sober 1994, sect.3).

The problem put in the Akaike framework:

- Given two domains D_1, D_2 ...
- ... why is it sometimes better to provide a unified (about domain $D = D_1 \cup D_2$), but less accurate model M_u instead of two separate models M_1, M_2, each one for one domain?
- So, why choose M_u, although $Pr(D|M_u) < Pr(D|M_1\&M_2)$?
- Answer: This inaccuracy may be compensated by simplicity (which is relevant for not overfitting) of M_u such that $AIC(M_u, D) > AIC(M_1\&M_2, D)$.

Application to Other Problems: Causal Modelling

And also such an application may be performed for rationalising several strategies of causal modelling (cf. Forster and Sober 1994, sect.4).

Again, the problem put in the Akaike framework:

- Given an effect E, why explanations that postulate fewer causes should be preferred over explanations that postulate more?
- Example (let's assume on/off causes: $val(C_1), val(C_2) \in \{0, 1\}$):

$Pr(E\|\dots)$	C_1	$\sim C_1$
C_2	$c_0, c_1, c_2, i_{c_1,c_2}$	c_0, c_2
$\sim C_2$	c_0, c_1	c_0

- We can formulate the following models:
  - $Pr(E|C_1, C_2) = c_0 + c_1 \cdot val(C_1)$
  - $Pr(E|C_1, C_2) = c_0 + c_1 \cdot val(C_1) + c_2 \cdot val(C_2)$
  - $Pr(E|C_1, C_2) = c_0 + c_1 \cdot val(C_1) + c_2 \cdot val(C_2) + i_{c_1,c_2} \cdot val(C_1) \cdot val(C_2)$

Application to Other Problems: Causal Modelling

Again, by similar reasoning as before one can argue for ...

$$ass_G(1, D) > ass_G(2, D) > ass_G(3, D)$$

... given an equal accurate description of the data D.

But note, here our seemingly tension between increase of predictive accuracy via implementation of intervention and decrease due to raised complexity appears.

Let's analyse it in the Akaike framework!

Causal Modelling

Interventions

Interventions play an important role in causal modelling/reasoning:

- (Woodward 2003): They are the foundation of causal discovery (not only an implementation in causal modelling).
- Causal Decision Theory: Our actions should be based on interventional knowledge (vs. knowledge about conditioning of classical decision theory).

General idea: By an intervention one "forces the system" to provide more specific causal information.

Implementation: Add a further variable that allow control of another one and screens more or less its ancestors off.

Depending on the "degree" to which screening off happens one may distinguish between structural (hard, arrow-braking) and parametric (soft) interventions.

The Causal Markov Condition

The usual preliminaries ...

Definition (Causal Markov Contition)

A causal bayes net (CBN=$\langle V, G, Pr\rangle$ (with $V = \{X_1, \ldots, X_n\}$,$G \subseteq V^2$,$parents(X) = \{Y : \langle Y, X\rangle \in G\}$, Pr is a prob. distr. over V) satisfies the causal Markov condition iff

$$Pr(X_1, \ldots, X_n) = \prod_{1 \leq i \leq n} Pr(X_i | parents(X_i))$$

Structural Interventions

Cf. (Eberhardt and Scheines 2007):

Definition (Structural Intervention)

I_s structurally intervenes on $X \in V$ of a CBN$=\langle V, G, Pr\rangle$ iff

1. I_s is a variable with two states (1/0, on/off).
2. When I_s is off, the passive observational distribution over V obtains.
3. I_s is a direct cause of X and only X.
4. I_s is exogenous, that is, uncaused.
5. When I_s is on, I_s makes X independent of its causes in V and determines the distribution of X; that is, in the factored joint distribution $Pr(X_1, \ldots, X_n)$, the term $Pr(X|parents(X))$ is replaced with the term $Pr(X|I_s)$, all other terms in the factorized joint distribution are unchanged.

Parametric Interventions

So-called 'parametric interventions' I_p are defined similarly (1–4) without demanding screening off.

I.e. (cf. Eberhardt and Scheines 2007, p.988): 5 changes to: $Pr(X|parents(X))$ is replaced with the term $Pr(X|parents(X), val(I_p) = 1)$

There are some interesting relations:

- Parametric interventions allow easily for a combination of several interventions, whereas structural interventions are due to their arrow-breaking property hardly simultanuously performable.
- There is also a difference in the number of interventions needed in order to figure out causal relations.

What we are interested in is their "performance" in the Akaike framework.

Application to Interventions

Since CBNs with common causes are mainly relevant in, e.g., causal decision theory, we will consider such a case:

$K \to H$, $K \to B$
(K ... potassium deficiency, B ... eating a banana, H ... having headache—(cf. Hitchcock 2015))

Interested in headache we may ask for:

$Pr(H\|\dots)$	B	$\sim B$
K	$s, b, k, i_{b,k}$	s, k
$\sim K$	s, b	s

And model: $Pr(H|B,K) = s + b \cdot val(B) + k \cdot val(K) + i_{b,k} \cdot val(B) \cdot val(K)$

Application to Interventions

What happens, if we intervene structurally (I_s)?

According to definition I_s,3 we expand our system by I_s.

So, seemingly the causal influence on headache is described more complex, although the interaction term between b, k vanishes:

$$Pr(H|B,K,I) = s + b \cdot val(B) + k \cdot val(K) + i_s \cdot val(I_s) + i_{i_s,b} \cdot val(B) \cdot val(I_s)$$

But one has to note that according to definition I_s,3,5 b as well as $i_{i_s,b}$ are determined by i_s: $b = f_1(i_s)$, $i_{i_s,b} = f_2(i_s)$.

By this in structurally intervening (in a common cause CBN) we even end up with a simpler model.

Of course it's different with parametric interventions where the model's complexity increases (due to the intervention and interaction parameters); a positive AIC-balance is gained only by sufficient causal information.

Outlook: Akaike and Causal Minimality

The so-called 'causal minimality condition' states:

Definition (Causal Minimality)

If CBN$=\langle V, G, Pr\rangle$ satisfies the causal Markov condition then it is causally minimal iff for all $G' \subset G$: $\langle V, G', Pr\rangle$ does not satisfy the causal Markov condition.

In the Akaike framework one may show that if a causal model (CBN with a model of Pr) has maximal AIC, then it is also causally minimal.

Summary

- We have mentioned a first-sight tension between predictability via interventional knowledge and complexity ...
- Then we presented Hitchcock and Sober's formulation of the predictivism vs. accommodationism problem ...
- ... and indicated how they try to solve it by help of Akaike's information framework.
- We have seen that this framework allows also for the discussion of nearby problems: ad-hoc modification, unification, minimal causal modelling.
- By applying it to interventions we saw that (strong) interventions even decrease complexity ...
- ... and by this the first-sight tension vanishes ...

References I

Eberhardt, Frederick and Scheines, Richard (2007). "Interventions and Causal Inference". English. In: *Philosophy of Science* 74.5, pp. 981–995. DOI: 10.1086/525638.

Forster, Malcolm R. and Sober, Elliott (1994). "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions". In: *The British Journal for the Philosophy of Science* 45.1, pp. 1–35. DOI: 10.1093/bjps/45.1.1.

Grünbaum, Adolf (1976). "Ad Hoc Auxiliary Hypotheses and Falsificationism". In: *The British Journal for the Philosophy of Science* 27.4, pp. 329–362. URL: http://www.jstor.org/stable/686862.

Hitchcock, Christopher (2015). "Conditioning, Intervening, and Decision". English. In: *Synthese*, pp. 1–20. DOI: 10.1007/s11229-015-0710-8.

Hitchcock, Christopher and Sober, Elliott (2004). "Prediction Versus Accommodation and the Risk of Overfitting". In: *The British Journal for the Philosophy of Science* 55.1, pp. 1–34. DOI: 10.1093/bjps/55.1.1.

Menke, Cornelis (2009). *Zum methodologischen Wert von Vorhersagen*. Paderborn: Mentis.

Sober, Elliott (2008). "Parsimony". In: *The Philosophy of Science. An Encyclopedia*. Ed. by Sarkar, Sahorta and Pfeifer, Jessica. London: Routledge, pp. 531–538.

Woodward, James (2003). *Making Things Happen. A Theory of Causal Explanation*. Oxford: Oxford University Press.