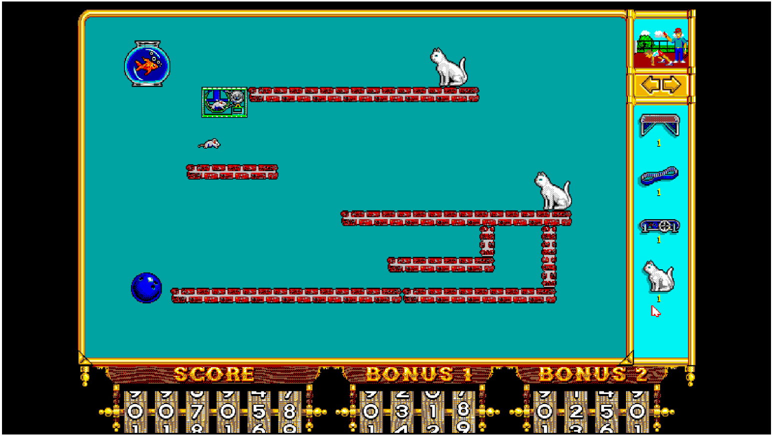
The Inflation Technique for Causal Inference

Elie Wolfe, Robert W. Spekkens, Tobias Fritz, arXiv:1609.00672

September 2017



Introduction

Given some correlations between the grammar or vocabulary of some languages, what can we say about common ancestor languages?

$\Rightarrow$ *Reichenbach's principle*: Two random variables A and B that are not independent must have a common ancestor.

Given three languages or three random variables A, B, C, how can we distinguish:

- There is a three-way common ancestor

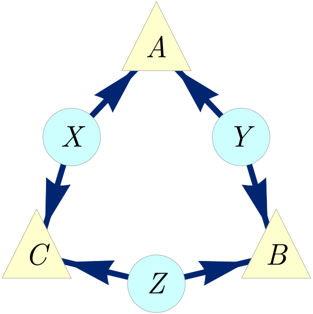
from the null hypothesis

- Any two of them have a pairwise common ancestors?[1]

---

[1] Bastian Steudel and Nihat Ay. "Information-theoretic inference of common ancestors". In: *Entropy* 17 (2015), pp. 2304–2327; Tobias Fritz. "Beyond Bell's theorem: correlation scenarios". In: *New J. Phys.* 14.10 (2012), p. 103001.

Or: given a joint distribution P_{ABC}, can it be obtained by marginalization from a Bayesian network of the "triangle" shape:

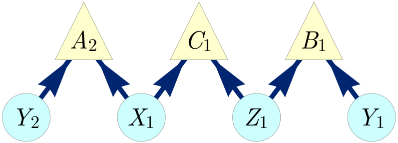


Example

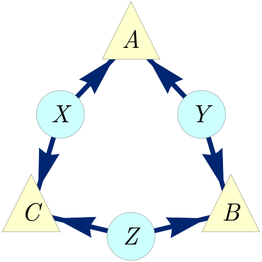
Binary variables with perfect correlation:

$$P_{ABC} = \frac{[000] + [111]}{2}$$

To solve the example problem, let's consider a slightly different graph:



Assuming that the causal dependences are as in the original



we conclude that some marginals are the same:

$$P_{A_2 C_1} = P_{AC}, \qquad P_{C_1 B_1} = P_{CB}.$$

…we conclude that some marginals are the same:

$$P_{A_2 C_1} = P_{AC}, \qquad P_{C_1 B_1} = P_{CB}.$$

We can now infer:

- A_2 and C_1 are perfectly correlated, as are C_1 and B_1.
- Hence A_2 and B_1 are perfectly correlated as well.
- But also: A_2 and B_1 are independent, since they do not hae a common ancestor! Thus

  $$P_{A_2 B_1} = P_A P_B,$$

  in contradiction with the assumption.

Theorem

P_{ABC} is incompatible with the Triangle graph.

We can make this more quantitative by deriving a *causal compatibility inequality*.

In terms of $\{\pm 1\}$-valued variables: The existence of a joint distribution $P_{A_2 B_1 C_1}$ implies a constraint on marginal distributions,

$$\langle A_2 C_1 \rangle + \langle C_1 B_1 \rangle \leq 1 + \langle A_2 B_1 \rangle. \tag{1}$$

These expectations values can be expressed in terms of the original ones on the triangle,

$$\langle A_2 C_1 \rangle = \langle AC \rangle, \qquad \langle C_1 B_1 \rangle = \langle CB \rangle, \qquad \langle A_2 B_1 \rangle = \langle A \rangle \langle B \rangle.$$

Substituting into (1) gives:

Theorem

Every P_{ABC} with $\{\pm 1\}$-valued variables compatible with the Triangle graph satisfies

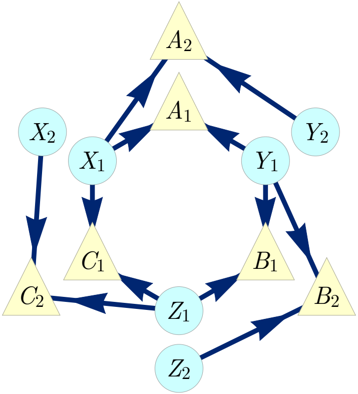
$$\langle AC \rangle + \langle BC \rangle \leq 1 + \langle A \rangle \langle B \rangle.$$

Another example

Is

$$P_{ABC} = \frac{[001] + [010] + [100]}{3}$$

compatible with the Triangle graph? Let's consider the *Spiral inflation*, where having the same causal dependences implies:



$$\begin{aligned}
P_{A_1 B_1 C_1} &= P_{ABC} \\
P_{A_1 B_2 C_2} &= P_{AB} P_C \\
P_{A_2 B_1 C_2} &= P_{BC} P_A \\
P_{A_2 B_2 C_1} &= P_{AC} P_B \\
P_{A_2 B_2 C_2} &= P_A P_B P_C .
\end{aligned}$$

These marginals are such that $A_2 = B_2 = C_2 = 1$ has positive probability.

Whenever this event happens, also one of the following must happen:

- $A_1 = B_2 = C_2 = 1$,
- $A_2 = B_1 = C_2 = 1$,
- $A_2 = B_2 = C_1 = 1$,
- $A_1 = B_1 = C_1 = 0$.

However, all of these have probability zero!

$\Rightarrow$ There is no joint distribution for all six variables that reproduces these marginals.

$\Rightarrow$ The original distribution P_{ABC} is not compatible with the Triangle graph.

Again one can make this inference quantitative by deriving an inequality.

At the level of the Spiral inflation, the union bound implies that

$$\begin{aligned} P_{A_2 B_2 C_2}(111) \leq P_{A_1 B_2 C_2}(111) + P_{A_2 B_1 C_2}(111) \\ + P_{A_2 B_2 C_1}(111) + P_{A_1 B_1 C_1}(000) \end{aligned}$$

in every joint distribution.

The above equations for the marginals translate this into

$$\begin{aligned} P_A(1) P_B(1) P_C(1) \leq P_{AB}(11) P_C(1) + P_{BC}(11) P_A(1) \\ + P_{AC}(11) P_B(1) + P_{ABC}(000), \end{aligned}$$

which is another causal compatibility inequality for the Triangle graph.

The inflation technique

So what is the general method?

Instead of the triangle, we may start with an arbitrary causal structure G, which is a *directed acyclic graph* with a distinction between observed and hidden nodes.

Definition

Given a graph G, an *inflation graph* is a graph G' together with a graph map $\pi : G' \to G$ such that its restriction to the ancestry of any node is an isomorphism.

In particular, $\pi : G' \to G$ must be a *fibration*: every edge in G with a lift of its source to G' lifts uniquely to G'.

In our figures, we specify the map by labelling each node in G' by the label of its image in G and a "copy index".

Every causal model on G *inflates* to a causal model on G' by using the same causal dependencies.

Definition

A set of nodes $U \subseteq G'$ is *injectable* if $\pi|_{\mathbf{U}}$ is bijective.

The distribution on an injectable set in an inflation model is specified by the corresponding marginal distribution on G.

Lemma

If a distribution on observable nodes of G is compatible with G, then the associated family of distributions on injectable sets is compatible with G'.

This is the central observation that makes the inflation technique work. It lets us apply any method for causal inference on G' and translate it to causal inference on G.

This can amplify the power of causal inference methods significantly. In the earlier examples, we have only used two things at the level of G',

- The existence of a joint distribution,
- Sets of nodes with disjoint ancestry are independent.

For G, we thereby obtain inequalities that do not just follow from the same requirements at the level of G!

Some of the sets that are relevant for applying the inflation technique in conjuction with the above two constraints on G' are:

Definition

A set of nodes $U \subseteq G'$ is *ai-expressible* if it is the union of injectable sets with disjoint ancestry.

(ai = ancestrally independent)

The most general class of sets of nodes on G' for which we can infer the joint distribution:

Definition

The collection of *expressible sets* is the smallest collection of sets of nodes such that:

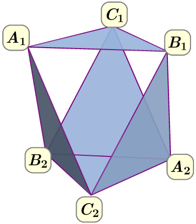
1. Every injectable set is expressible.
2. If $A \subseteq G'$ is expressible, then so is every subset of A.
3. For $A, B, C \subseteq G'$, if
   - C d-separates A from B,
   - $A \cup C$ and $B \cup C$ are expressible,

   then $A \cup B \cup C$ is also expressible.

Not every expressible set is ai-expressible.

There are many possible techniques that one can apply at the level of the inflation graph.

The simplest is to use merely *the existence of a joint distribution*. Thus we need to solve the *marginal problem*: when does the family of marginal distributions on expressible sets permit a joint distribution?



The families of distributions that arise in this way form the *marginal polytope*. Its linear facet inequalities become polynomial causal compatibility inequalities for G.

Thus the computational problems are those of *linear programming* and *facet enumeration* for the marginal polytope.

Let's see some results!

Facet enumeration for the marginal polytope of the Spiral inflation with $\{\pm 1\}$-valued variables results in 4 symmetry classes of nontrivial irredundant inequalities:

$$0 \leq 1 - \langle AC \rangle - \langle BC \rangle + \langle A \rangle \langle B \rangle$$

$$\begin{aligned} 0 \leq 3 &- \langle A \rangle - \langle B \rangle - \langle C \rangle + 2\langle AB \rangle + 2\langle AC \rangle + 2\langle BC \rangle + \langle ABC \rangle \\ &+ \langle A \rangle \langle B \rangle + \langle A \rangle \langle C \rangle + \langle B \rangle \langle C \rangle - \langle A \rangle \langle BC \rangle - \langle B \rangle \langle AC \rangle - \langle C \rangle \langle AB \rangle \\ &+ \langle A \rangle \langle B \rangle \langle C \rangle \end{aligned}$$

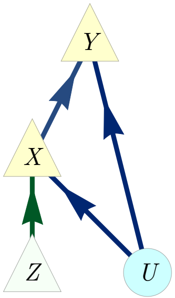
$$\begin{aligned} 0 \leq 4 &+ 2\langle C \rangle - 2\langle AB \rangle - 3\langle AC \rangle - 2\langle BC \rangle - \langle ABC \rangle \\ &+ 2\langle A \rangle \langle B \rangle + \langle A \rangle \langle C \rangle - \langle A \rangle \langle BC \rangle - \langle C \rangle \langle AB \rangle \\ &+ \langle A \rangle \langle B \rangle \langle C \rangle \end{aligned}$$

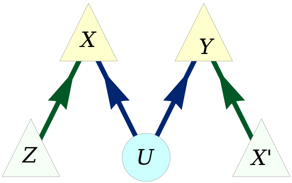
$$\begin{aligned} 0 \leq 4 &- 2\langle AB \rangle - 2\langle AC \rangle - 2\langle BC \rangle - \langle ABC \rangle \\ &+ 2\langle A \rangle \langle B \rangle + 2\langle A \rangle \langle C \rangle + 2\langle B \rangle \langle C \rangle - \langle A \rangle \langle BC \rangle - \langle B \rangle \langle AC \rangle - \langle C \rangle \langle AB \rangle \end{aligned}$$

Towards Completeness

This outlines the simplest way of applying the inflation technique. One can strengthen it in two ways:

- We can assume that isomorphic subgraphs carry the same distribution, e.g. $P_{A_1 Y_1} = P_{A_2 Y_2}$ in the Spiral inflation.
- Some causal structures do not have interesting inflations. This can be fixed by Introducing counterfactual variables:





It has recently been proven:[2]

Theorem

Supplemented with these two techniques, **the inflation technique solves causal inference with hidden variables completely**: a given distribution P is compatible with a given causal structure if and only if it survives all inflation tests.

Main ingredient of proof:

- A de Finetti theorem for Bayesian networks with one hidden layer and one observable layer (*restricted Boltzmann machines*).

[2] Miguel Navascués and Elie Wolfe. *The inflation technique solves completely the classical inference problem*. arXiv:1707.06476.

Entropic inequalities

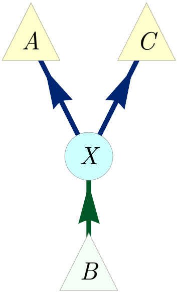
The *laws of information theory* (Yeung):

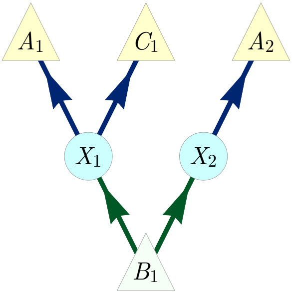
- Submodularity, $H(AB) + H(BC) \geq H(B) + H(ABC)$.
- *Non-Shannon-type* inequalities, such as the Zhang–Yeung inequality,

  $$\begin{aligned} &3H(AC) + 3H(AD) + H(BC) + H(BD) + 3H(CD) \\ &\qquad \geq 4H(ACD) + H(BCD) + H(AB) + H(A) + 2H(C) + 2H(D), \end{aligned}$$

  which are not consequences of submodularity.
- These are the inequalities that bound the entropy cone.
- Finding a complete list of non-Shannon-type inequalities is an open problem.

The derivation of the known non-Shannon-type inequalities relies on the *copy lemma*, which secretly is an application of the inflation technique:





Problem

Is it possible to derive new non-Shannon-type inequalities by considering other inflation graphs?

Other types of networks

How general is the basic idea? Can one apply it to networks like the following:

- ~~Circuit diagrams and deterministic computation, as e.g. in neural networks~~
- Restricted Boltzmann machines?
- Phylogenetic trees? (Three languages example!)

There are two variants of the general technique, applying to networks in monoidal categories where

- Systems can be discarded: the unit object is terminal (weaker variant).
- Systems can be copied: in addition, every object comes equipped with a comonoid structure (stronger variant).