Causal Inference Theory with Information Dependency Models

When Causal Inference meets Statistical Analysis, CNAM, Paris, 17-21 avril 2023

Benjamin Heymann, Michel De Lara, Jean-Philippe Chancelier

April 17, 2023

BH: Criteo AI Lab, Paris, France
MDL and JPC : Cermics, École des Ponts, Marne-la-Vallée, France

Information Dependency Models and Information Fields

- **Information dependency models**: causality with **information fields**
- **Information fields**: Witsenhausen's 1971 paper [1]
- Witsenhausen's motivation: control of multi-agent systems
- but in fact, it is a very generic tool
  - Used to revisit the foundations of game theory[2]
  - Theoretical toolbox for causality: **the Information Dependency Model** (IDM)

[1] *On information structures, feedback and causality.*
[2] *Kuhn's equivalence theorem for games in product form*

Making the case for Information Dependency Model (IDM)

- Unlock **mathematical toolboxes**
- **Unifying and generalizing** framework for causality[3]
- Elegant style of expression and proof : **equational reasoning**
- Potential to **bridge** causality, game theory, control and Reinforcement Learning

[3] can deal with spurious edges, cycles

What is the common denominator to those areas?

In some sense:

"To depend on" = "observing" = "knowing" = "playing after"

The three main ideas

- **IDM**, as a generalization of causal graphs/an alternative language to describe causal dependencies
- **Binary relations**, as a way to encode causal influence
- **Topological separation**, as an alternative definition of d-separation

"Alice, Bob and a coin tossing" configuration space

Example

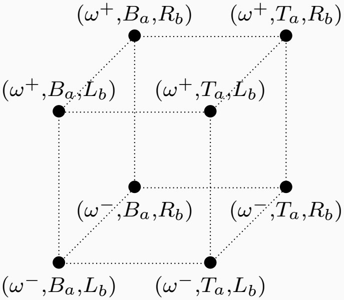
- two states of Nature $\Omega = \{\omega^+, \omega^-\}$ (heads/tails)
- two agents a and b
- two possible actions each: $\mathbb{U}_a = \{T_a, B_a\}$, $\mathbb{U}_b = \{R_b, L_b\}$

"Alice, Bob and a coin tossing" configuration space

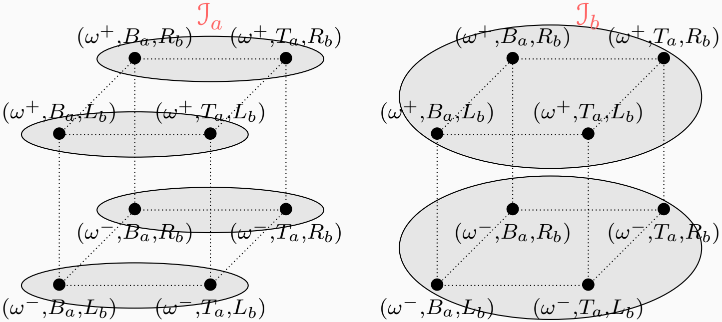
Example

- two states of Nature $\Omega = \{\omega^+, \omega^-\}$ (heads/tails)
- two agents *a* and *b*
- two possible actions each: $\mathbb{U}_a = \{T_a, B_a\}$, $\mathbb{U}_b = \{R_b, L_b\}$
- product configuration space (8 elements)

$$\mathbb{H} = \{\omega^+, \omega^-\} \times \{T_a, B_a\} \times \{R_b, L_b\}$$



"Alice, Bob and a coin tossing" information partitions



$$\mathcal{I}_b = \overbrace{\{\emptyset, \{\omega^+\}, \{\omega^-\}, \{\omega^+, \omega^-\}\}}^{\text{Bob knows Nature's move}} \otimes \overbrace{\{\emptyset, \{T_a, B_a\}\}}^{\text{Bob does not know what Alice does}} \otimes \{\emptyset, \mathbb{U}_b\}$$

$$\mathcal{I}_a = \underbrace{\{\emptyset, \{\omega^+\}, \{\omega^-\}, \{\omega^+, \omega^-\}\}}_{\text{Alice knows Nature's move}} \otimes \{\emptyset, \mathbb{U}_a\} \otimes \underbrace{\{\emptyset, \{R_b\}, \{L_b\}, \{R_b, L_b\}\}}_{\text{Alice knows what Bob does}}$$

Witsenhausen's philosophy

- $\mathbb{H}$ is the **domain** of every function
- for any variable *a* **encode** the "dependence" by asking for **measurability** w.r.t. information field[4] $\mathcal{I}_a$, that is,

$$\lambda_a : (\mathbb{H}, \mathcal{H}) \to (\mathbb{U}_a, \mathcal{U}_a)$$

$$\lambda_a^{-1}(\mathcal{U}_a) \subset \mathcal{I}_a$$

4

- A *σ-field* over a set $\mathbb{D}$ is a subset $\mathcal{D} \subset 2^{\mathbb{D}}$, containing $\mathbb{D}$, and which is stable under complement and countable union. (The trivial σ-field over the set $\mathbb{D}$ is $\{\emptyset, \mathbb{D}\}$)
- Probability theory defines a *random variable* as a measurable mapping from $(\Omega, \mathcal{F})$ to $(\mathbb{U}, \mathcal{U})$.

Structural Causal Model (SCM)

$$U_a(\omega) = \lambda_a(U_{P(a)}(\omega), \omega_a) \quad \forall \omega \in \Omega \quad \forall a \in \mathbb{A}$$

- $(\lambda_a)_{a \in \mathbb{A}}$: assignments
- $P : \mathbb{A} \to 2^{\mathbb{A}}$: parental mapping

In the example:

- $\lambda_{Bob} = \lambda_{Bob}(U_{Coin}, \omega_{Bob})$
- $\lambda_{Alice} = \lambda_{Alice}(U_{Coin}, U_{Bob}, \omega_{Alice})$

Information Dependency Model (IDM)

1. The configuration space is the product space

$$\mathbb{H} = \prod_{a \in \mathbb{A}} \mathbb{U}_a \times \Omega$$

2. $\mathcal{H}$ is the product field of $\mathbb{H}$
3. An Information Dependency Model is a collection $(\mathcal{I}_a)_{a \in \mathbb{A}}$ of subfields of $\mathcal{H}$ such that, for $a \in \mathbb{A}$,

$$\mathcal{I}_a \subset \bigotimes_{b \in \mathbb{A}} \mathcal{U}_b \otimes \mathcal{F}_a$$

   The subfield $\mathcal{I}_a$ is called the **information field** of a.
4. SCM now defined by the field inclusion

$$\lambda_a^{-1}(\mathcal{U}_a) \subset \mathcal{I}_a \quad \forall a \in \mathbb{A}$$

From SCM to IDM, an illustration

$$U_a(\omega) = \lambda_a(U_{P(a)}(\omega), \omega_a) \quad \forall \omega \in \Omega \quad \forall a \in \mathbb{A}$$

In the example,

- $\lambda_{Bob} = \lambda_{Bob}(U_{Coin}, \omega_{Bob})$ becomes $\lambda_{Bob}^{-1}(\mathcal{U}_{Bob}) \subset \mathcal{I}_{Bob}$
- $\lambda_{Alice} = \lambda_{Alice}(U_{Coin}, U_{Bob}, \omega_{Alice})$ becomes $\lambda_{Alice}^{-1}(\mathcal{U}_{Alice}) \subset \mathcal{I}_{Alice}$,

where

$$\mathcal{I}_{Bob} = \overbrace{\{\emptyset, \{\omega^+\}, \{\omega^-\}, \{\omega^+, \omega^-\}\}}^{\text{Bob knows Nature's move}} \otimes \overbrace{\{\emptyset, \{T_a, B_a\}\}}^{\text{Bob does not know what Alice does}} \otimes \{\emptyset, \mathbb{U}_b\}$$

$$\mathcal{I}_{Alice} = \underbrace{\{\emptyset, \{\omega^+\}, \{\omega^-\}, \{\omega^+, \omega^-\}\}}_{\text{Alice knows Nature's move}} \otimes \{\emptyset, \mathbb{U}_a\} \otimes \underbrace{\{\emptyset, \{R_b\}, \{L_b\}, \{R_b, L_b\}\}}_{\text{Alice knows what Bob does}}$$

DAGs v.s. information fields

	Pearl	*Witsenhausen*
Structure	DAG	binary relations[5]
Dependence	SCM	information fields
	functional relation	measurable policy profiles
Resolution	induction	solution map[6]
Intervention	do operator	encoded with information fields
Causal ordering	fixed	not fixed (might not exist)

Table 1: Correspondences between Pearl's DAG and Witsenhausen's intrinsic model

[5] minimality for free
[6] allows for compositional arguments

Well-posedness

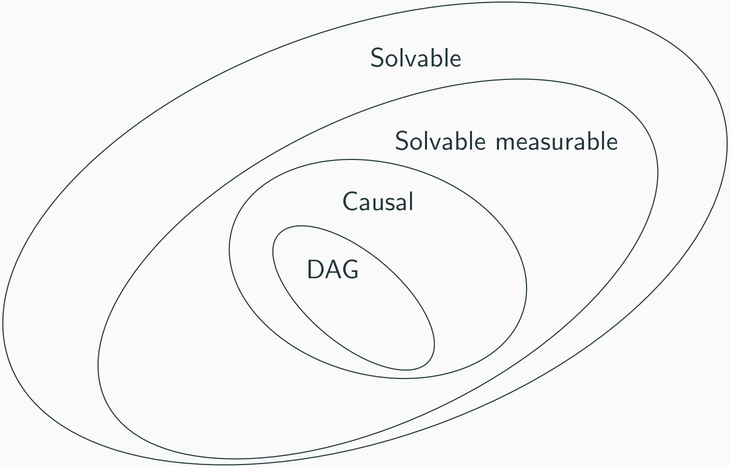


Figure 1: Hierarchy of systems

(W, H)-Conditional Precedence

Definition

The conditional predecessor set $\mathcal{E}^{W,H}a$ is the smallest subset $B \subset \mathbb{A}$ such that

$$\mathcal{I}_a \cap H \subset \mathcal{H}_{B \cup W} \cap H$$

(for $W \subset \mathbb{A}$, $H \subset \mathbb{H}$ and $a \in \mathbb{A}$).

We denote by $\bar{B}$ (or $\bar{B}^{W,H}$) the topological closure of B, which is the smallest subset of $\mathbb{A}$ that contains B and its own predecessors under $\mathcal{E}^{W,H}$.

Topological separation and Do-calculus

Definition (Topological Separation)

We say that B and C are (conditionally) *topologically separated* (wrt (W, H)), and write

$$B \perp\!\!\!\perp_t C \mid (W, H),$$

if there exists $W_B, W_C \subset W$ such that

$$W_B \sqcup W_C = W \text{ and } \overline{B \cup W_B} \cap \overline{C \cup W_C} = \emptyset$$

Theorem (Do-calculus)

$$Y \perp\!\!\!\perp_t Z \mid (W, H) \Longrightarrow \Pr(U_Y \mid U_W, U_{\bar{Z}}, H) = \Pr(U_Y \mid U_W, H)$$

Topological separation and d-separation are equivalent

Theorem

Let $(\mathcal{V}, \mathcal{E})$ be a graph, that is, $\mathcal{V}$ is a set and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, and let $W \subset \mathcal{V}$ be a subset of vertices, we have the equivalence

$$b \underset{t}{\perp\!\!\!\perp} c \mid W \iff b \underset{d}{\perp\!\!\!\perp} c \mid W \qquad \big(\forall b, c \in W^{\mathsf{c}}\big)$$

Topological separation: example 1

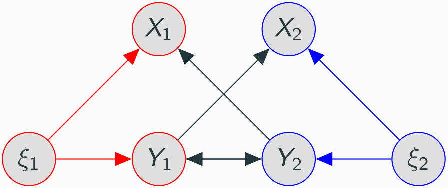


Figure 2: Let $W_{X_i} = Y_i$, for $i = 1, 2$. The closure of $X_1 \cup Y_1$ (resp. $X_2 \cup Y_2$), with the edges followed to build the closure, is in red (resp. blue).

Topological separation: example 2

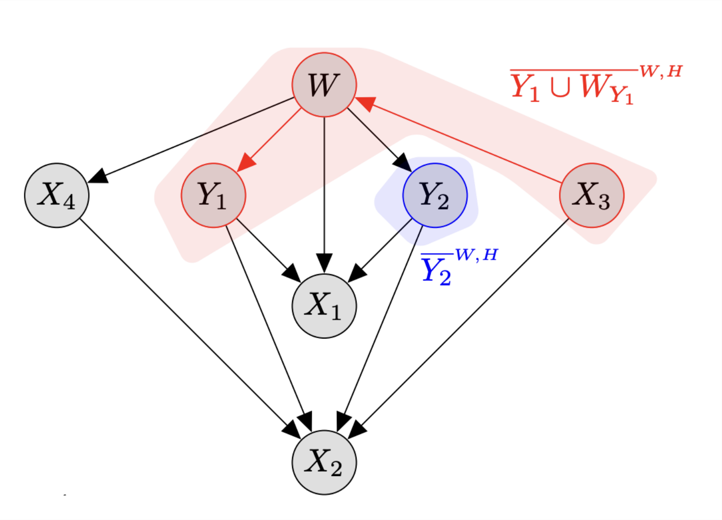


Figure 3: The split of W is a piece of information that can be insightful.

An illustration of equational reasoning

Proof We have that

$$
\begin{aligned}
&\Delta_{W^c}\big(\Delta \cup (\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W\big)\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\big(\Delta \cup \mathcal{C}^W(\mathcal{B}^{-W} \cup \mathcal{K}^W)\big)\Delta_{W^c} \\
&\quad = \Delta_{W^c}\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\Delta_{W^c} && \text{(by developing)} \\
&\quad\quad \cup \Delta_{W^c}\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\big(\mathcal{C}^W(\mathcal{B}^{-W} \cup \mathcal{K}^W)\big)\Delta_{W^c} \\
&\quad\quad \cup \Delta_{W^c}\big((\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W\big)\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\Delta_{W^c} \\
&\quad\quad \cup \Delta_{W^c}\big((\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W\big)\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\big(\mathcal{C}^W(\mathcal{B}^{-W} \cup \mathcal{K}^W)\big)\Delta_{W^c} \\
&\quad = \Delta_{W^c}\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\Delta_{W^c} \\
&\quad\quad \cup \Delta_{W^c}\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W(\mathcal{B}^{-W} \cup \mathcal{K}^W)\Delta_{W^c} && \text{(as } \mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\mathcal{C}^W = \mathcal{C}^W \text{ by (34c))} \\
&\quad\quad \cup \Delta_{W^c}(\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W\mathcal{E}^{-W*}\mathcal{E}^{W*}\Delta_{W^c} && \text{(also by (34c))} \\
&\quad\quad \cup \Delta_{W^c}(\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W(\mathcal{B}^{-W} \cup \mathcal{K}^W)\Delta_{W^c} && \text{(also by (34c) applied twice)} \\
&\quad = \Delta_{W^c}\big(\mathcal{B}^W \cup \mathcal{K}^W\big)\mathcal{C}^W\big(\mathcal{B}^{-W} \cup \mathcal{K}^W\big)\Delta_{W^c} && \text{(by (34d) and (34e))} \\
&\quad\quad \cup \Delta_{W^c}\big(\mathcal{B}^W \cup \mathcal{K}^W\big)(\mathcal{B}^{-W} \cup \mathcal{K}^W)\Delta_{W^c} && \text{(by (34e))} \\
&\quad\quad \cup \Delta_{W^c}(\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W\big(\mathcal{B}^{-W} \cup \mathcal{K}^W\big)\Delta_{W^c} && \text{(by (34d))} \\
&\quad\quad \cup \Delta_{W^c}(\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W(\mathcal{B}^{-W} \cup \mathcal{K}^W)\Delta_{W^c} \\
&\quad = \Delta_{W^c}(\mathcal{B}^W \cup \mathcal{K}^W)\mathcal{C}^W(\mathcal{B}^{-W} \cup \mathcal{K}^W)\Delta_{W^c} \,.
\end{aligned}
$$

This ends the proof. ■

Next steps

- Extend to continuous variables
- Relax the well-posedness assumption
- See how it goes for algorithm design

Conclusion

- Pearl's celebrated do-calculus provides a set of inference rules to derive an interventional probability from an observational one. The primitive causal relations are encoded as **functional dependencies**.
- In this paper, by contrast, we capture causality **without reference to functional dependencies**, but with **information fields**.
- The three rules of do-calculus reduce to a **unique sufficient condition for conditional independence**.
- We introduce the **topological separation**, a notion equivalent to d-separation, but that highlights other aspects.
- The proposed framework handles systems that cannot be represented with DAGs, for instance **'spurious' edges**.

→ A versatile, unifying foundational model

References

H. S. Witsenhausen.
On information structures, feedback and causality.
SIAM J. Control, 9(2):149–160, May 1971.

S. Tikka, A. Hyttinen, and J. Karvanen.
Identifying causal effects via context-specific independence relations.
In *Advances in Neural Information Processing Systems*, pages 2804–2814, 2019.

B. Heymann, M. De Lara, J. P. Chancelier.
Kuhn's equivalence theorem for games in product form,
In *Games and Economic Behavior*, Volume 135, 2022,

B. Heymann, M. De Lara, J. P. Chancelier.
Causal inference with information fields (long version), 2020.
preprint.