Abstraction Between Structural Causal Models and Measure of Abstraction Error

Fabio Massimo Zennaro

University of Warwick

September 5th, 2023

1. Introduction

Levels of Abstraction

Systems may be represented at different **levels of abstraction** (LoA).

Thermodynamics example:

Low-level / Base model:
Microscopic description $\mathbf{p}, \dot{\mathbf{p}}$.

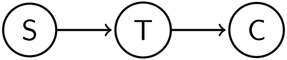
High-level / Abstracted model:
Macroscopic description P, T, V.

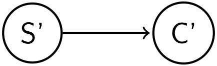
1. How do we *express relations* of abstraction?
2. How do we *measure correctness* of abstraction?
3. How do we *assess properties* at different LoAs?
4. How do we *take advantage* of multiple LoAs?
5. How do we *learn* LoAs?

Causal Models

We focus on **causal models** that can be expressed using graphical models.

Lung cancer scenario example:



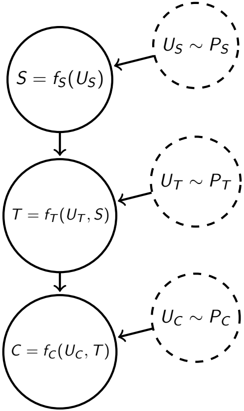


1. How do we *express relations* of abstraction among causal models?
2. How do we *measure correctness* of causal abstraction?
3. How do we *assess properties* at different LoAs? [3, 11, 2]
4. How do we *take advantage* of multiple LoAs? [12]
5. How do we *learn* LoAs? [12]

SCMs [6, 7]

We work with **structural causal models** (SCM) $\mathcal{M} = \langle \mathcal{X}, \mathcal{U}, \mathcal{F}, \mathcal{P} \rangle$:

- $\mathcal{X}$: set of *endogenous nodes* (S, T, C) representing variables of interest
- $\mathcal{U}$: Set of *exogenous nodes* (U_S, U_T, U_C) representing stochastic factors
- $\mathcal{F}$: Set of *structural functions* (f_S, f_T, f_C) describing the dynamics of each variable
- $\mathcal{P}$: Set of *distributions* (P_S, P_T, P_C) describing the random factors



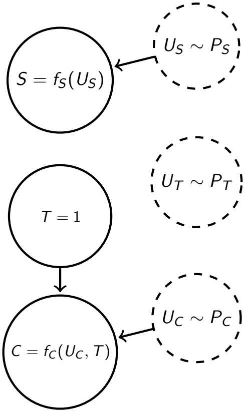
Every SCM $\mathcal{M}$ implies a (joint) **distribution** $P_{\mathcal{M}}$: $P_{\mathcal{M}}(S, T, C)$

Interventions

We can perform **interventions** on a causal model [6, 7]:

$do(T = 1)$

1. Remove incoming edges in the intervened node
2. Set the value of the intervened node

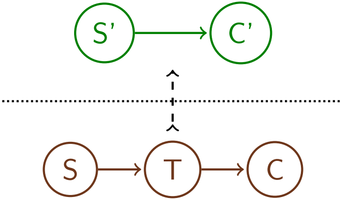


An intervention ι_1 effectively defines a new **intervened model** $\mathcal{M}_{\iota_1}$ such that $P_{\mathcal{M}}(S, T, C) \neq P_{\mathcal{M}_{\iota_1}}(S, T, C)$

2. Abstraction

Three approaches

Lung cancer scenario example:



$$\mathcal{M}'[S'] = \mathcal{M}'[C'] = \{0, 1\}$$

$$\mathcal{M}[S] = \mathcal{M}[T] = \mathcal{M}[C] = \{0, 1\}$$

- The **transformation** approach [10, 1]
- The **Φ-abstraction** approach [4, 5]
- The **α-abstraction** approach [9, 8]

The *transformation* approach: mapping [10]

Given two SCMs $\mathcal{M}$ and $\mathcal{M}'$, let us consider the **transformation**:

$$\tau : P_{\mathcal{M}} \mapsto P_{\mathcal{M}'}$$

Formally, τ is a function between variables implying a *pushforward* between distributions.

Under an assumption of *observational consistency*, this implies

$$\tau_{\#}(P_{\mathcal{M}}) = P_{\mathcal{M}'}$$

The *transformation* approach: consistency [10]

Let us consider a mapping between interventions:

$$\omega : \mathcal{I} \to \mathcal{I}'$$

$$\omega : \text{an intervention on } \mathcal{M} \mapsto \text{an intervention on } \mathcal{M}'.$$

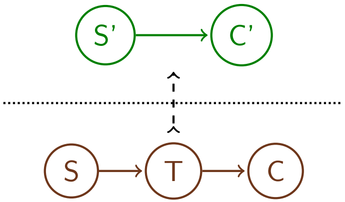
A transformation is an *exact transformation* if there exists a surjective order-preserving ω such that:

$$\begin{tikzcd} P_{\mathcal{M}} \arrow[r, "\tau"] \arrow[dd, "\iota"'] & \tau(P_{\mathcal{M}}) = P_{\mathcal{M}'} \arrow[d, "\omega(\iota)"] \\ & P_{\mathcal{M}_{\omega(\iota)}} \\ P_{\mathcal{M}_\iota} \arrow[r, "\tau"'] & \tau(P_{\mathcal{M}_\iota}) \end{tikzcd}$$

where $\tau(P_{\mathcal{M}_\iota}) = P_{\mathcal{M}_{\omega(\iota)}}$, $\forall \iota \in \mathcal{I}$. We have *interventional consistency*.

The *transformation* approach: example

Lung cancer scenario example:



$$\tau : \mathcal{M}[S] \times \mathcal{M}[T] \times \mathcal{M}[C] \to \mathcal{M}'[S'] \times \mathcal{M}'[C']$$

$$\tau : (s, t, c) \mapsto (s, c)$$

Set of interventions: $\mathcal{I} = \{\emptyset, do(S = 0)\}$

$$\omega : \begin{cases} \emptyset \mapsto \emptyset \\ do(S = 0) \mapsto do(S' = 0) \end{cases}$$

Consistency condition:

$$\begin{array}{ccc} P_{\mathcal{M}}(S, T, C) & \xrightarrow{\tau} & P_{\mathcal{M}'}(S', C') \\ \Big\downarrow \iota & & \Big\downarrow \omega(\iota) \\ P_{\mathcal{M}}(T, C | do(S = 0)) & \underset{\tau}{\rightarrow} & P_{\mathcal{M}'}(C' | do(S' = 0)) \end{array}$$

The Φ-*abstraction* approach: mapping [4]

An SCM $\mathcal{M}$ can be formalized as a *functor* from a syntactic category:

$$F_{\mathcal{M}} : \mathtt{Syn}_{\mathcal{M}} \to \mathtt{FinStoch}$$

In this formalization, an intervention is an *endofunctor* on the syntactic category:

$$\mathrm{cut}_X : \mathtt{Syn}_{\mathcal{M}} \to \mathtt{Syn}_{\mathcal{M}}$$

The Φ-*abstraction* approach: consistency [4]

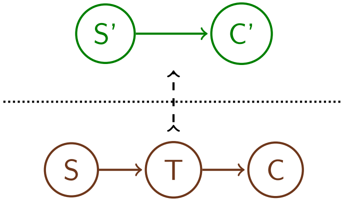
Given two SCMs $\mathcal{M}$ and $\mathcal{M}'$ with a homomorphism ϕ between their DAGs, an abstraction exists if we have a *natural transformation* between the respective functors:

$$
\begin{tikzcd}
\mathtt{Syn}_{\mathcal{M}} \arrow[r, "F_{\mathcal{M}}"] \arrow[d, "\Phi"'] \arrow[dr, phantom, "\Downarrow"] & \mathtt{FinStoch} \arrow[d, "\mathrm{id}"] \\
\mathtt{Syn}_{\mathcal{M}'} \arrow[r, "F_{\mathcal{M}'}"'] & \mathtt{FinStoch}
\end{tikzcd}
$$

Given a Φ-abstraction, the homomorphism ϕ guarantees *interventional consistency*.

The Φ-*abstraction* approach: example

Lung cancer scenario example:



$$\mathtt{Syn}_{\mathcal{M}} : \bullet_S \longrightarrow \bullet_T \longrightarrow \bullet_C$$

$$\mathtt{Syn}_{\mathcal{M}'} : \bullet_{S'} \longrightarrow \bullet_{C'}$$

$$F_{\mathcal{M}} : \begin{cases} \bullet \mapsto \{0,1\} \\ \longrightarrow \mapsto \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \end{cases}$$

$$F_{\mathcal{M}'} : \begin{cases} \bullet \mapsto \{0,1\} \\ \longrightarrow \mapsto \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \end{cases}$$

$$\Phi : \bullet_S \mapsto \bullet_{S'}, \bullet_T \mapsto \bullet_{S'}, \bullet_C \mapsto \bullet_{C'}$$

A natural transformation is a *collection of maps* in `FinStoch`.

The α-*abstraction* approach: mapping [9]

Let $\mathcal{M}$ and $\mathcal{M}'$ be two finite SCMs with finite domains. An **abstraction** is a tuple

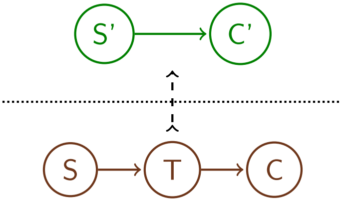
$$(R, a, \alpha)$$

where:

- $R \subseteq \mathcal{X}_{\mathcal{M}}$ is a subset of *relevant nodes* among the endogenous nodes of $\mathcal{M}$.
- $a : R \to \mathcal{X}_{\mathcal{M}'}$ is a *surjective function* mapping a low-level node in $\mathcal{M}$ to a high-level node in $\mathcal{M}'$.
- α is a *collection of surjective functions*, one for each high-level node X', defined as $\alpha_{X'} : \mathcal{M}[a^{-1}(X')] \to \mathcal{M}'[X']$.
  α'_X maps an output of the low-level nodes sent onto X' by a onto an output of X'.

The α-*abstraction* approach: example (I)

Lung cancer scenario example:



$$R = \{S, C\} \subseteq \mathcal{X}_{\mathcal{M}}$$

$$a : R \to \mathcal{X}_{\mathcal{M}'}$$

$$a : \begin{cases} S \mapsto S' \\ C \mapsto C' \end{cases}$$

$$\alpha : \begin{cases} \alpha_{S'} : \{0,1\} \to \{0,1\} \\ \alpha_{S'} : s \mapsto s \\ \alpha_{C'} : \{0,1\} \to \{0,1\} \\ \alpha_{C'} : c \mapsto c \end{cases}$$

The α-*abstraction* approach: abstraction error

We evaluate the *quality* of an abstraction in terms of *interventional consistency*.

The **abstraction error** wrt $P(\mathbf{Y}'|do(\mathbf{X}'))$ is the maximum *distance between interventional distributions* in the base and abstracted model.

$$\begin{tikzcd}
\mathcal{M}[\mathbf{X}] \arrow[r, "\mu", "P_{\mathcal{M}_\iota}(\mathbf{Y}|do(\mathbf{X}))"'] \arrow[d, "\alpha_{\mathbf{X}'}"'] & \mathcal{M}[\mathbf{Y}] \arrow[d, "\alpha_{\mathbf{Y}'}"] \\
\mathcal{M}'[\mathbf{X}'] \arrow[r, "\nu", "P_{\mathcal{M}'_{\iota'}}(\mathbf{Y}'|do(\mathbf{X}'))"'] & \mathcal{M}'[\mathbf{Y}']
\end{tikzcd}$$

$$E(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') = \max_{\mathbf{x} \in \mathcal{M}[\mathbf{X}]} D_{JSD}(\alpha_{\mathbf{Y}'} \cdot \mu, \nu \cdot \alpha_{\mathbf{X}'})$$

The α-*abstraction* approach: abstraction error [9]

An abstraction implies multiple *abstraction errors*.

(Global) abstraction error $e(\alpha)$ is the maximum abstraction error over all disjoint sets of variables.

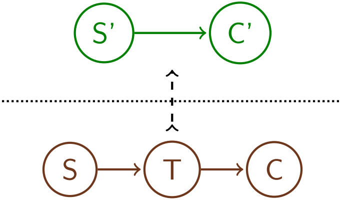
$$\begin{tikzcd} \mathcal{M}[X] \arrow[r, "\mu"] \arrow[d, "\alpha_{X'}"'] & \mathcal{M}[Y] \arrow[d, "\alpha_{Y'}"] \\ \mathcal{M}'[X'] \arrow[r, "\nu"] & \mathcal{M}'[Y'] \end{tikzcd} \qquad \begin{tikzcd} \mathcal{M}[Y] \arrow[r, "\mu'"] \arrow[d, "\alpha_{Y'}"'] & \mathcal{M}[Z] \arrow[d, "\alpha_{Z'}"] \\ \mathcal{M}'[Y'] \arrow[r, "\nu'"] & \mathcal{M}'[Z'] \end{tikzcd}$$

$$\begin{tikzcd} \mathcal{M}[X] \arrow[r, "\mu' \circ \mu"] \arrow[d, "\alpha_{X'}"'] & \mathcal{M}[Z] \arrow[d, "\alpha_{Z'}"] \\ \mathcal{M}'[X'] \arrow[r, "\nu' \circ \nu"] & \mathcal{M}'[Z'] \end{tikzcd}$$

$$e(\boldsymbol{\alpha}) = \sup_{\mathbf{X}', \mathbf{Y}' \subseteq \mathcal{X}'} E(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$$

The α-*abstraction* approach: example (II)

Lung cancer scenario example:



Assuming no commutativity

$$\begin{tikzcd} \mathcal{M}[S] \arrow[r, "\mu_C"] \arrow[d, "\alpha_{S'}"'] & \mathcal{M}[C] \arrow[d, "\alpha_{C'}"] \\ \mathcal{M}'[S'] \arrow[r, "\nu_{C'}"'] & \mathcal{M}'[C'] \end{tikzcd}$$

I can compute *abstraction error*:
$E(\boldsymbol{\alpha}, S', C') = D_{JSD}(\alpha_{C'} \circ \mu_C, \nu_{C'} \circ \alpha_{S'})$

Since there are not other subsets this is also the *overall abstraction error*:
$e(\boldsymbol{\alpha}) = E(\boldsymbol{\alpha}, S', C')$

Summary of approaches

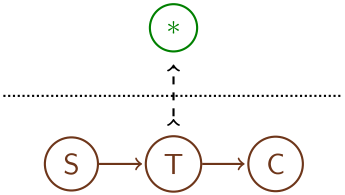
- **Transformation approach:** works at the *distributional* level.
- **Φ-abstraction approach:** works at the *structural* level.
- **α-abstraction approach:** works at the *distributional*/ *structural* level.

Following we will focus on *α-abstraction approach*.

3. Measuring Abstraction Error

Measuring Abstraction Error [13]

In the *α-abstraction* framework, does **abstraction error** tell us the whole story about abstraction?



Let $\mathcal{M}'$ be the trivial singleton model.

Then, $e_\alpha = 0$.

We want other *quantitative measures* for an abstraction.

Generalizing Abstraction Error [13]

The abstraction error can be expressed more generally as:

$$E_{\boldsymbol{\alpha}}(\mathbf{X}', \mathbf{Y}') = \underset{x' \in \mathbf{X}'}{\text{agg}} \; D(p, q)$$

$$e(\boldsymbol{\alpha}) = \underset{(\mathbf{X}', \mathbf{Y}') \in \mathcal{J}}{\text{agg}} \; E_{\boldsymbol{\alpha}}(\mathbf{X}', \mathbf{Y}')$$

$$\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}$$

parametrized by aggregation functions, distances, intervention sets, pseudo-inverse, and paths.

Parameters for a Generalized Abstraction Error

- Aggregation functions:
  - Which *guarantees* do we want?
  - How do we *weight* errors?
- Distances:
  - What *metric* do we use on the statistical manifold?
  - Which *properties* does each measure entail?
- Intervention sets:
  - Which interventions are *non-redundant*?
  - Which interventions are *relevant*?
- Pseudo-inverse:
  - How should be an *inverse* defined at all?

Paths: new error measures

If we consider different *paths*, we derive *new error measures*:

Interventional consistency (IC)

$$\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}$$

Consistency projected on the abstracted model.

Interventional information loss (IIL)

$$\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}$$

Loss in abstracting and reconstructing.

Paths: new error measures

Interventional superresolution information loss (ISIL)

$$
\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}
$$

Loss in reconstructing and abstracting.

Interventional superresolution consistency (ISC)

$$
\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}
$$

Consistency projected on the base model.

Some properties of these new error measures

For all the measures above (IC,IIL,ISIL,ISC) with supremum aggregation:

- *Non-monotonicty*: not given that $e(\boldsymbol{\beta\alpha}) \geq e(\boldsymbol{\alpha})$
- *Triangle inequality*: $e(\boldsymbol{\beta\alpha}) \leq e(\boldsymbol{\alpha}) + e(\boldsymbol{\beta})$
- *Ordering:* IIL $\geq$ IC, IIL $\geq$ ISC, IC $\geq$ ISIL, ISC $\geq$ ISIL
- *Finiteness condition*: error is finite if a is order-preserving
- *Different minima:* IC, IIL, ISC, ISIL may disagree on minima

4. Conclusion

Conclusions

Large space for conceptual and practical development of **causal abstraction frameworks**:

- *Foundations* of the framemorks
- *Characterization* of these frameworks
- *Algorithmic and empirical* development

More about abstraction:
https://github.com/FMZennaro/CausalAbstraction/

Thanks!

Thank you for your attention!

References I

[1] Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

[2] Erik P Hoel. When the map is better than the territory. *Entropy*, 19(5):188, 2017.

[3] Erik P Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

[4] Jun Otsuka and Hayato Saigo. On the equivalence of causal models: A category-theoretic approach. *arXiv preprint arXiv:2201.06981*, 2022.

[5] Jun Otsuka and Hayato Saigo. The process theory of causality: an overview. 2022.

[6] Judea Pearl. *Causality*. Cambridge University Press, 2009.

References II

[7] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

[8] Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. *arXiv preprint arXiv:2103.15758*, 2021.

[9] Eigil Fjeldgren Rischel. The category theory of causal models. 2020.

[10] Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

[11] Fabio Massimo Zennaro. Abstraction between structural causal models: A review of definitions and properties. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

References III

[12] Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *2nd Conference on Causal Learning and Reasoning*, 2023.

[13] Fabio Massimo Zennaro, Paolo Turrini, and Theo Damoulas. Quantifying consistency and information loss for causal abstraction learning. In *Proceedings of the Thrity-Second International Conference on International Joint Conferences on Artificial Intelligence*, 2023.