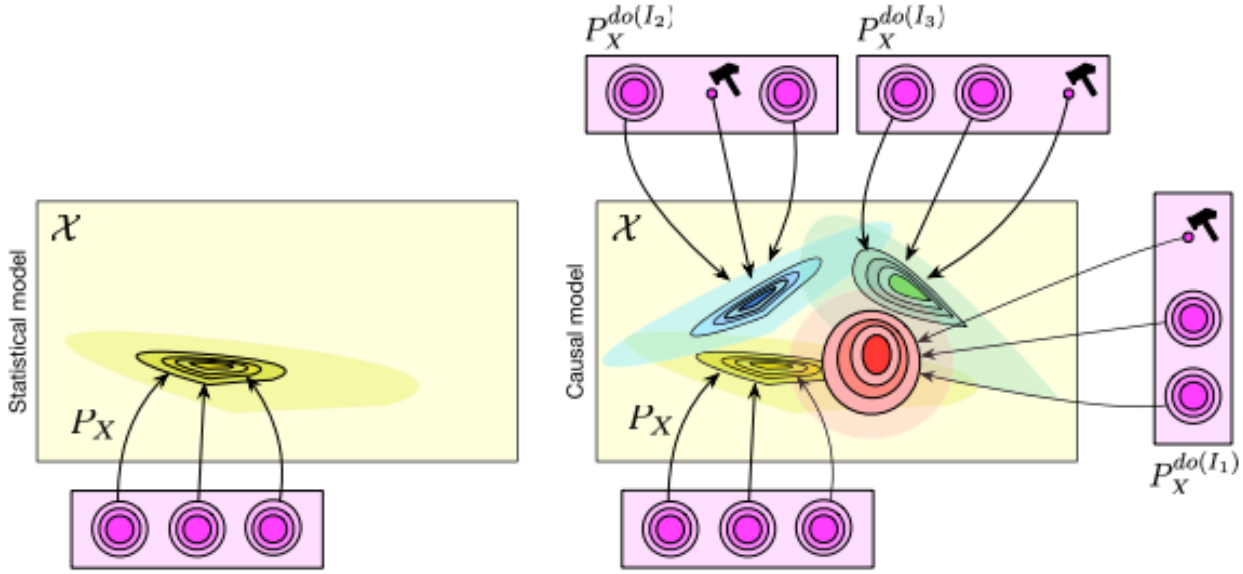
Toward Causal Representation Learning



Content from Schoelkopf et al, Toward Causal Representation Learning

ECE324, Winter 2023

Michael Guerzhoy

The usual supervised learning setting

- Data: $\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\}$
- Assume data are i.i.d., i.e. $(x, y) \sim P((x, y))$
- Want to learn (or compute from a generative model) $P(y|x)$

Issues with the usual setting

- Robustness
  - Want to be robust to changes in the test distribution
    - In Vision: camera blur, noise, shifts, rotations…
    - In Vision: adversarial examples
  - Change in test distribution: $P(x, y)$ is different from what's in the training set

Issues with the usual setting

- Learning reusable mechanisms
  - Infants learn that physical objects can be tracked over time and behave consistently
  - Want to camture such mechanisms, with few examples

Issues with the usual setting

- Want to predict the outcomes of counterfactual scenarios
  - P(rain|umbrellas) is high in the training set, but want to know P(rain|do(umbrellas))

Taxonomy of models

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

- Physical model: $\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d$
  - Earlier x's cause later x's
- Statistical model: an approximation of P(X, Y)
  - "The frequency of storks is a reasonable predictor of human birth rates in Europe…. A change to the stork population would not affect the birth rates"
  - What went wrong there? Changing the stork population makes the distribution of P(babies, storks) different
- Structural causal: a DAG that encodes causal relationships
- Causal graphical: a DAG that does not always encode causal relationships

Structural Causal Models

- Each edge in the graph is

$$X_i := f_i(PA_i, U_i)$$

- X_i is caused by its parents, U_i is unexplained noise, the U's are jointly independent

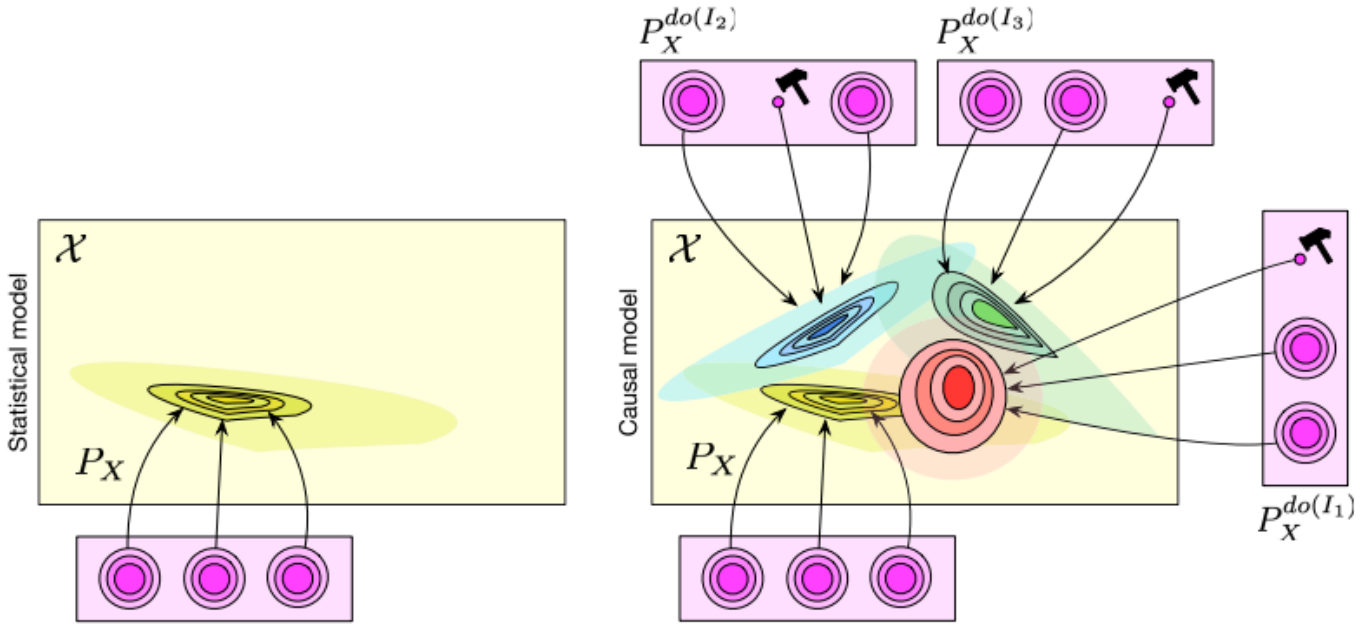


Fig. 1. ***Difference between statistical (left) and causal models (right) on a given set of three variables. While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention (indicated with a 🔨).***

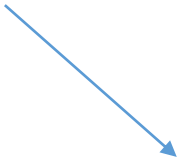
Difference between SCM and a causal graphical model

- A: altitude
- T: temperature
- $P(A, T) = P(A|T)P(T) = P(T|A)P(A)$
- $P(A, T)$ might be different for Austria and Switzerland, but $P(T|A)$ might be the same
- The SCM $P(T|A)P(A)$ encodes the mechanism of generating the temperature from the altitude that can generalize across countries

ICM principle: The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

1) Changing (or performing an intervention upon) one mechanism $P(X_i|\mathbf{PA}_i)$ does not change any of the other mechanisms $P(X_j|\mathbf{PA}_j)$ $(i \neq j)$ [220].
2) Knowing some other mechanisms $P(X_i|\mathbf{PA}_i)$ $(i \neq j)$ does not give us information about a mechanism $P(X_j|\mathbf{PA}_j)$ [124].

hypothesis

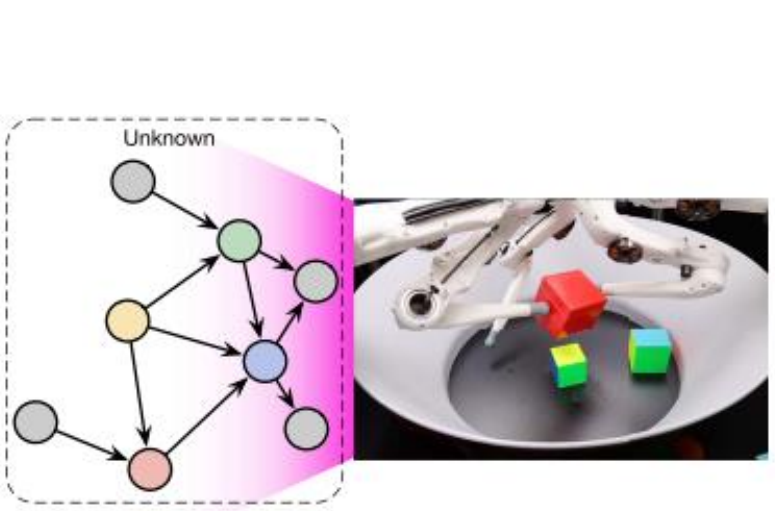


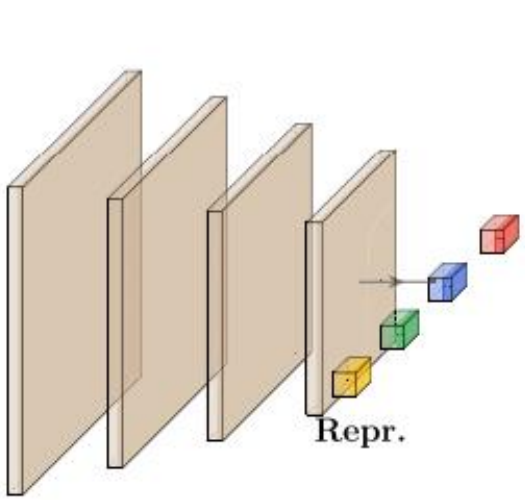
SMS: Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization [see (4)], that is, they should usually not affect all factors simultaneously.

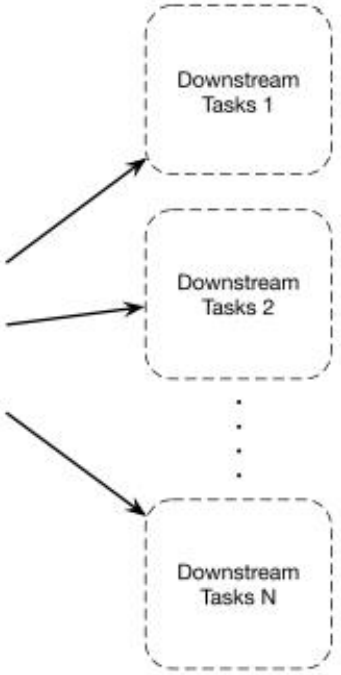
Causal discovery and machine learning

- In a very large dataset, can obtain conditional independence results, gaining insights about the causal graphical model
  - But in general this is very difficult
- Model assumptions are necessary
  - E.g. $Y = f(X) + V$ (note that in this case we can tell that the model is not $X = g(Y) + V$)

Causal representation learning problem setting







- Goal: make neural networks learn representations where neurons are governed by SCM

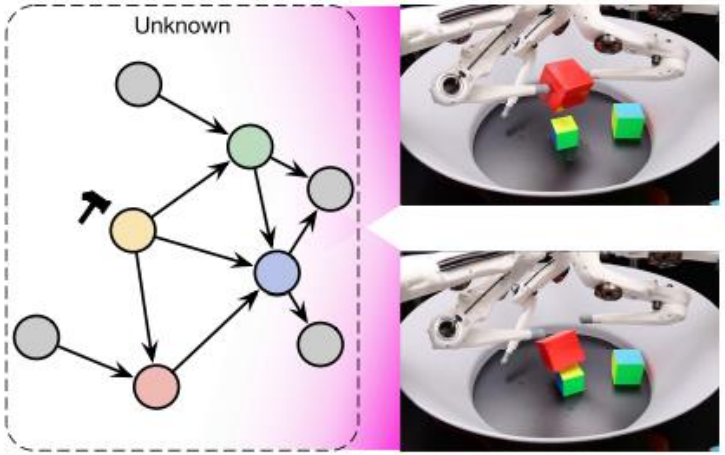


Fig. 3. ***Example of the SMS hypothesis where an intervention (which may or may not be intentional/observed) changes the position of one finger (🔨), and as a consequence, the object falls. The change in pixel space is entangled (or distributed), in contrast to the change in the causal model.***

- Approach 1: use autoencoders to learn a disentangled representation
- Approach 2: use object-centric representation
  - Object detection as a submodule of the system
- Approach 3: Explicitly incorporate view invariance

- Goal: learn transferable representations
  - Make the system modular. E.g. deal with lighting separately

- Goal: Learning Interventional World Models and Reasoning
  - Really difficult; requires reasoning