Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data

Siyuan Guo*

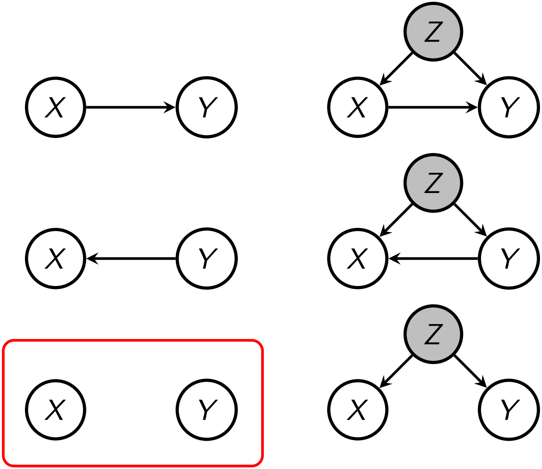
Joint work with

Viktor Tóth*, Ferenc Huszár and Bernhard Schölkopf

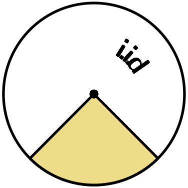
University of Cambridge
Max Planck Institute for Intelligent Systems

syg26@cantab.ac.uk *syguoML*

Causal Structure Identification



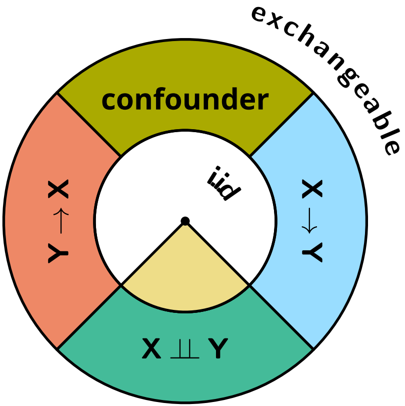
Current Limitation



$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$$

- I. I. D: We can only differentiate whether X and Y are independent.

Key Takeaways



- I. I. D: We cannot differentiate between $X \rightarrow Y$ and $Y \rightarrow X$.
- Exchangeable: We can.

What is this talk about?

- What does $X \to Y$ mean in exchangeable data generating process?
  - Formalization of Independent Causal Mechanism Principle
- Establishes connection between invariant causal structure and conditional independence in exchangeable process
  - **Causal de Finetti**
    - Bivariate
    - Multivariate
- How Causal de Finetti can be used in practice?
  - Exchangeable generative models vs 'Grouped data'
  - Algorithm for recovering invariant causal structure from grouped data via conditional independence

Background

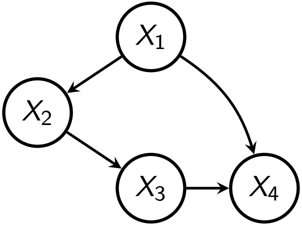
Structural Causal Model

Definition

A structural causal model (SCM) M is given by a set of variables $X_1, \ldots, X_N$ and corresponding structural assignments of the form

$$X_i := f_i(PA_i, U_i), i = 1, \ldots, N \tag{1}$$

where, PA_i are **parents** or **direct causes** of X_i and U_i are **noise** variables, which we require to be jointly independent.



Conditional Independence

Conditional Independence in Probability:

$$X \perp\!\!\!\perp Y \mid Z$$

$$\Leftrightarrow$$

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

Conditional Independence Assumption Encoded in DAG:

$$X \perp\!\!\!\perp_{\mathscr{G}} Y \mid Z$$

represents a conditional independence relationship assumption encoded by a DAG $\mathscr{G}$.

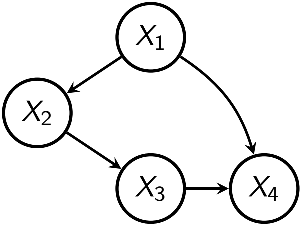
D-separation

Definition

A path p is d-separated by a block of node Z if and only if one of the two conditions holds:

1. p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ s.t. $m \in Z$
2. p contains $i \rightarrow m \leftarrow j$ s.t. the middle node $m \notin Z$ and $de(m) \notin Z$

We say Z d-separates X and Y if it blocks every path from a node in X to a node in Y.



Theorem (Markov Property)

Given a DAG $\mathscr{G}$ and a joint distribution P, this distribution is said to satisfy:

- ***markov factorization property*** *with respect to a DAG $\mathscr{G}$ if*

$$P(X_1,\ldots,X_N) = \prod_i \underbrace{P(X_i|PA_i)}_{\text{causal conditional}} \tag{2}$$

- ***global markov property*** *with respect to a DAG $\mathscr{G}$ if*

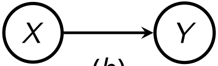
$$A \perp\!\!\!\perp_{\mathscr{G}} B|C \implies A \perp\!\!\!\perp B|C \tag{3}$$

If P has a density p, then above markov properties are equivalent.

Example in I.I.D

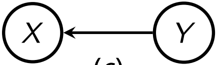
(a)

$\{X \perp\!\!\!\perp Y\}$



(b)

$\emptyset$



(c)

$\emptyset$

I.I.D $\to$ Exchangeable

Exchangeable

Definition

A finite sequence of random variables $X_1, X_2, \ldots, X_N$ is called **exchangeable**, if for any permutation π of $\{1, \ldots, N\}$, we have that

$$P(X_{\pi(1)}, \ldots, X_{\pi(N)}) = P(X_1, \ldots, X_N) \tag{4}$$

We say we have an **infinite exchangeable** sequence if for any $N \in \mathbb{N}$, the finite sequence with length N is exchangeable.

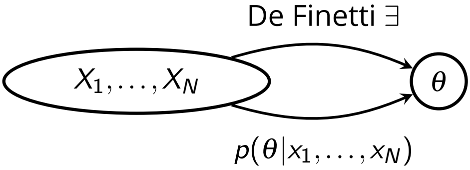
1. the order of observations does not matter
2. i.i.d data is exchangeable but not all exchangeable data is i.i.d

Theorem (De Finetti)

*Let $(X_n)_{n\in\mathbb{N}}$ be an infinite sequence of binary random variables. The sequence is **exchangeable** if and only if there exists θ such that $X_1, X_2, \ldots$ are conditionally i.i.d given θ, with a prob. measure μ on θ. i.e. Given any sequence $(\mathbf{x}_1, .., \mathbf{x}_N) \in \{0,1\}^N$, we have*

$$P(X_1 = \mathbf{x}_1, \ldots, X_N = \mathbf{x}_N) = \int \prod_{i=1}^{N} p(\mathbf{x}_i|\theta)\, d\mu(\theta) \tag{5}$$

1. Justifies Bayesian statistics



Independent Causal Mechanisms (ICM) Principle

It states that the causal generative process of a system's variables is composed of autonomous modules that

1. do **not inform** each other
2. do **not influence** each other

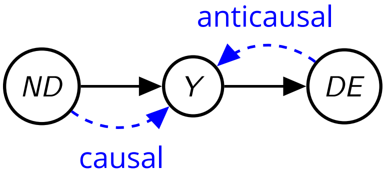
$$P(X_1, \ldots, X_N) = \prod_i \underbrace{P(X_i|PA_i)}_{\text{causal conditional}} \tag{6}$$

The principle says that the causal conditionals should be independent in the sense:

1. **not inform**: knowing some other mechanisms $P(X_j|PA_j)(i \neq j)$ does not give us information about a mechanism $P(X_i|PA_i)$
2. **not influence**: changing one mechanism $P(X_i|PA_i)$ does not change any of the other mechanisms $P(X_j|PA_j)(i \neq j)$

ICM Related Work

- On causal and anticausal learning [Schölkopf et al. 2012]



Connection to ML:

	Causal $X \to Y$	Anticausal $Y \to X$
Semi-supervised Learning (SSL)	✗ "$P_X \perp\!\!\!\perp P_{Y\|X}$"	✔ $P_X = P_Y \circ P_{X\|Y}$
Covariate Shift $P_X \to Q_X$	✔	✗

- Invariant Risk Minimization [Arjovsky et al. 2019]
- Invariant Causal Prediction [Peters et al. 2016]
- Learning Independent Causal Mechanisms [Parascandolo et al. 2018]
- Fast and Slow Learning of Recurrent Independent Mechanisms [Madan et al. 2021]

ICM Formalization

ICM Principle, though understand intuitively, lacks a formal testable definition. For example, consider a bivariate example $C \rightarrow E$, ICM states that two mechanisms are independent, i.e. "$P_{E|C} \perp\!\!\!\perp P_C$".

What does "$P_{E|C} \perp\!\!\!\perp P_C$" mean?

- Algorithmic independence [Janzing and Schölkopf 2010]: encode each mechanism as a bit string, and require that joint compression of these strings does not save space relative to independent compressions.
- Can we have a **statistical** formalization for "$P_{E|C} \perp\!\!\!\perp P_C$"?

Causal de Finetti

Theorem (Causal de Finetti - bivariate)

Let $\{X_i, Y_i\}_{i\in\mathbb{N}}$ *be an infinite sequence of binary r.vs.*
Suppose:

1. *the sequence is infinitely exchangeable*
2. $\forall n \in \mathbb{N} : Y_{[n]} \perp\!\!\!\perp X_{n+1} | X_{[n]}$ *Note* $[n] := \{1, \ldots, n\}$

Then $\exists$ *suitable* μ, ν *such that:*

$$P(X_1 = x_1, Y_1 = y_1, \ldots, X_N = x_N, Y_N = y_N) = \int \prod_{n=1}^{N} p(y_n|x_n, \psi) p(x_n|\theta) d\mu(\theta) d\nu(\psi) \tag{7}$$

1. Encode ICM
2. Identify bivariate causal structure

Theorem

Let $\{X_i, Y_i\}_{i\in\mathbb{N}}$ be an infinite sequence of binary r.vs. Suppose:

1. *the sequence is infinitely exchangeable*
2. $\forall n \in \mathbb{N} : Y_{[n]} \perp\!\!\!\perp X_{n+1} | X_{[n]}$

Then $\exists$ suitable μ, ν such that:

$$P(X_1 = x_1, Y_1 = y_1, \ldots, X_N = x_N, Y_N = y_N) = \int \prod_{n=1}^{N} p(y_n|x_n, \psi)p(x_n|\theta) \underbrace{d\mu(\theta)d\nu(\psi)}_{\theta \perp\!\!\!\perp \psi} \tag{8}$$

1. Encode ICM
2. Identify bivariate causal structure

Causal de Finetti - bivariate

Theorem

Let $\{X_i, Y_i\}_{i\in\mathbb{N}}$ *be an infinite sequence of binary r.vs. Suppose:*

1. *the sequence is infinitely exchangeable*
2. $\forall n \in \mathbb{N} : Y_{[n]} \perp\!\!\!\perp X_{n+1} | X_{[n]} \implies \text{"}P_{Y|X} \perp\!\!\!\perp P_X\text{"}$

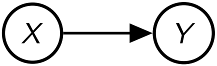
Then $\exists$ *suitable* μ, ν *such that:*

$$P(X_1 = x_1, Y_1 = y_1, \ldots, X_N = x_N, Y_N = y_N) \\ = \int \prod_{n=1}^{N} p(y_n|x_n, \psi)p(x_n|\theta)d\mu(\theta)d\nu(\psi) \tag{9}$$

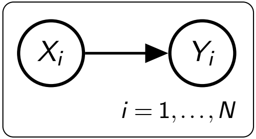
1. Encode ICM
2. Identify bivariate causal structure

Causal Graph with Data Generating Process

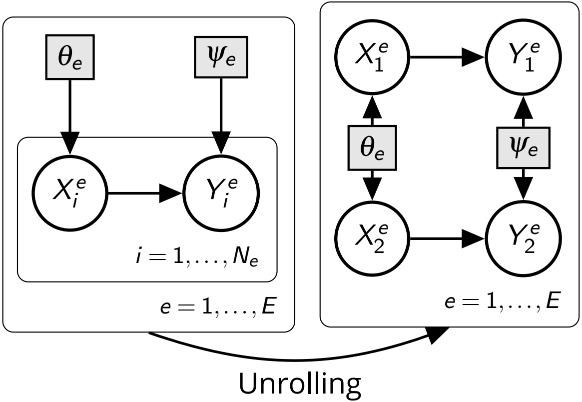
Causal Graph



i. i. d. process

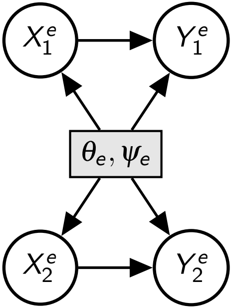


exchangeable process

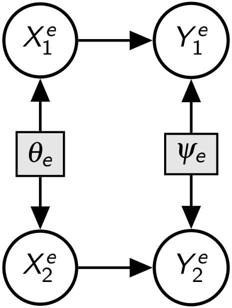


Disentangle the Latents

De Finetti:



Causal De Finetti:

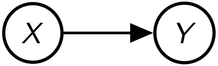


Remark

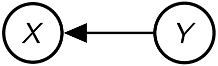
Mechanisms are independent in the sense that the latents governing different mechanisms are statistically independent.

Identify causal structure

I.I.D:

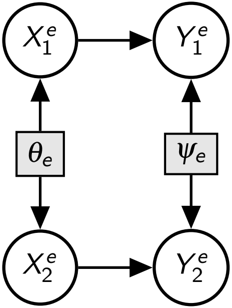


$\emptyset$

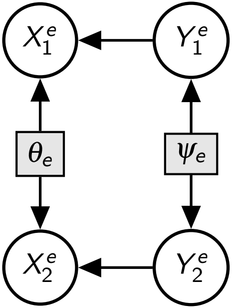


$\emptyset$

Exch-angeable

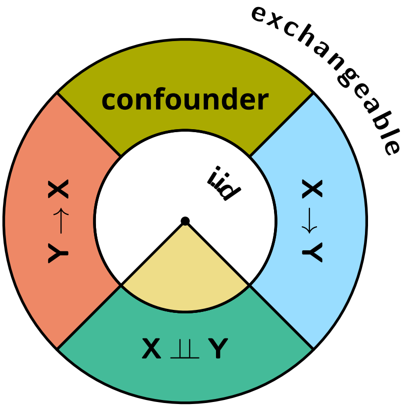


$$\{X_1^e \perp\!\!\!\perp Y_2^e | X_2^e, X_2^e \perp\!\!\!\perp Y_1^e | X_1^e\}$$



$$\{X_1^e \perp\!\!\!\perp Y_2^e | Y_1^e, X_2^e \perp\!\!\!\perp Y_1^e | Y_2^e\}$$

Key Takeaways



- I.I.D: We cannot differentiate between $X \to Y$ and $Y \to X$.
- Exchangeable: We can.

Theorem (Causal de Finetti - multivariate)

Let $\{X_{1;n}, X_{2;n}, \ldots, X_{d;n}\}_{n\in\mathbb{N}}$ *be an infinite sequence of d-tuple binary r.vs.* *variable index* *sample index*

Suppose:

1. *the sequence is infinitely exchangeable*
2. *If there exists a DAG* $\mathscr{G}$ *such that* $\forall i \in [d], \forall n \in \mathbb{N}$:

$$X_{i;[n]} \perp\!\!\!\perp \overline{ND}_{i;[n]}, ND_{i;n+1} | PA_{i;[n]}$$

where PA_i*: parents of node i,*
ND_i*: non-descendants of node i,*
$\overline{ND}_i$*: non-descendants of node i excluding its own parents.*

Then $\exists$ *suitable* ν_i *s.t. the joint probability can be written as*

$$(\ldots) = \int \cdots \int \prod_{n=1}^{N} \prod_{i=1}^{d} p(x_{i;n} | pa_{i;n}, \theta_i) d\nu_1(\theta_1) \ldots d\nu_d(\theta_d) \tag{10}$$

Understanding the conditions

$$X_{i;[n]} \perp\!\!\!\perp \overline{ND}_{i;[n]}, ND_{i;n+1} | PA_{i;[n]}$$

1. $X_{i;[n]} \perp\!\!\!\perp \overline{ND}_{i;[n]} | PA_{i;[n]}$
   - PA_i is a markov blanket for $\overline{ND}_i$
2. $X_{i;[n]} \perp\!\!\!\perp ND_{i;n+1} | PA_{i;[n]}$
   - Encodes "$P_{X_i|PA_i} \perp\!\!\!\perp P_{ND_i}$"
   - Note $PA_i \subseteq ND_i$, so above implies $P_{X_i|PA_i} \perp\!\!\!\perp P_{PA_i}$

Grouped Data

$$
\begin{array}{lcccc}
\text{Env}_1: & (X_1^1, Y_1^1) & (X_2^1, Y_2^1) & \cdots & (X_N^1, Y_N^1) \\
\text{Env}_2: & (X_1^2, Y_1^2) & (X_2^2, Y_2^2) & \cdots & (X_N^2, Y_N^2) \\
\vdots & & \vdots & & \\
\text{Env}_E: & (X_1^E, Y_1^E) & (X_2^E, Y_2^E) & \cdots & (X_N^E, Y_N^E)
\end{array}
$$

- Multi-environments
- Related through some *invariant causal structure*

Clinical Data:

- Env = Hospital
- (X, Y) = Medical Variables

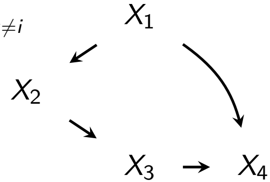
Algorithm

Suppose we have multiple environments. Each environment are independent with each other. Suppose further within each environment, our observed data is an exchangeable process and each sample shares the same causal structure.

- **Input**: $(X^e_{1;n}, \ldots, X^e_{d;n})_{n=1}^{N_e}, \forall e \in \mathscr{E}$. Assume $N_e \geq 2, \forall e$.
- **Output**: A DAG $\mathscr{G}$

1. Identify observed variable's topological ordering
    - Identify first-order sinks S_1.
      We say $i \in S_1$ if $\forall j \neq i, \forall e \in \mathscr{E}$

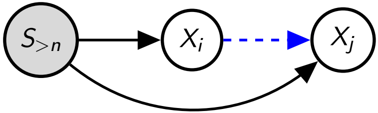
$$X^e_{i;1} \perp\!\!\!\perp X^e_{j;2} | \{X^e_{k;1}\}_{k \neq i}$$



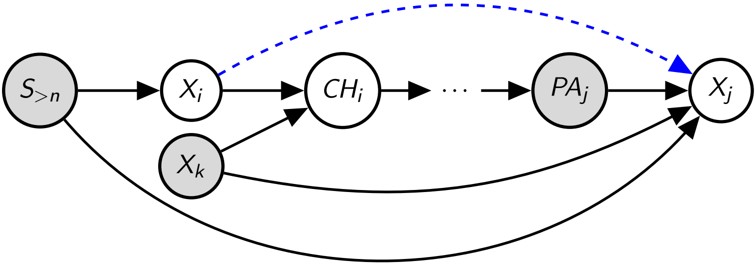
    - Remove identified S_1 nodes
    - Find new S_1 nodes as S_2
    - Iteratively repeat above

Algorithm

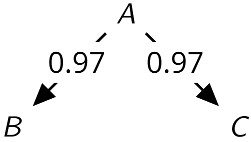
2. Identify edges between different topological orders
   Suppose $X_i \in S_n$ and $X_j \in S_m$, where $n > m$.
   - Suppose $t = n - m = 1$



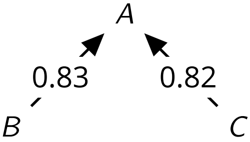
   - Suppose $t > 1$



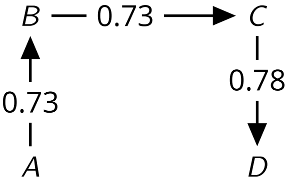
Experiments



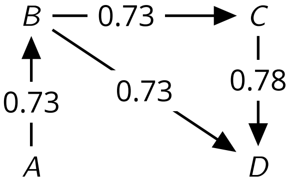
Acc: 91%



Acc: 71%

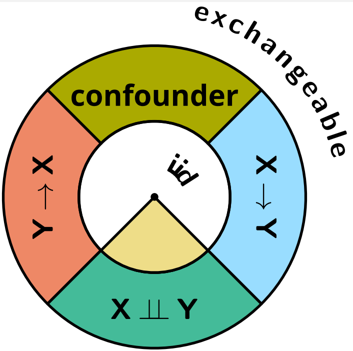


Acc: 60%



Acc: 65%

What is this talk about?



- What does $X \to Y$ mean in exchangeable data generating process?
- Connection between invariant causal structure and conditional independence in exchangeable process
- How Causal de Finetti can be used in practice?

References I

Schölkopf et al. (2012)
On Causal and Anticausal Learning
Proceedings of the 29th International Conference on Machine Learning

Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D. (2019)
Invariant risk minimization.
arXiv preprint arXiv:1907.02893.

Peters, J., Bühlmann, P., Meinshausen, N. (2016)
Causal inference by using invariant prediction: identification and confidence intervals.
*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*78(5), 947-1012.

Janzing, D., Schölkopf, B. (2010)
Causal inference using the algorithmic Markov condition
IEEE Transactions on Inf. Theory 56.

References II

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., Schölkopf, B.(2018)
Learning Independent Causal Mechanisms
Proceedings of the 35th International Conference on Machine Learning

Madan, K., Ke, R., Goyal, A., Schölkopf, B., Bengio, Y. (2021)
Fast and Slow Learning of Recurrent Independent Mechanisms
ICLR 2021

Thank you!