Potential Outcomes & Structure Identifiability

Konstantin Ibadullaev

19.06.2024

Agenda

1. Potential Outcomes
   - Description & Examples
   - Relation between Potential Outcomes and SCMs
2. Structure Identifiability
   - Additive Noise Models
   - Linear Gaussian Models with Equal Error Variances
   - Linear Non-Gaussian Acyclic Models
   - Nonlinear Gaussian Additive Noise Models

Potential Outcomes. Description & Examples

Potential Outcomes is an alternative approach to causal inference For the illustration purposes let us consider the **Example 3.4**

Example 3.4 Eye disease

There exists a rather effective treatment for an eye disease.
For 99% of all patients, the treatment works and the patient gets cured (B =0);
if untreated, these patients turn blind within a day (B = 1).
For the remaining 1%, the treatment has the opposite effect and they turn blind (B = 1) within a day. If untreated, they regain normal vision (B = 0).

Potential Outcomes. Description & Examples

1. Consider $\mathbf{u=1,...n}$ a group of $\mathbf{n}$ patients instead of random variables
2. Assign two potential outcomes to each patient $\mathbf{u}$:
   - $\mathbf{B_u(t = 1) = 1}$ - patient would go blind if receives treatment $(\mathbf{T = 1})$
   - $\mathbf{B_u(t = 1) = 0}$ - patient would get cured if receives treatment $(\mathbf{T = 1})$

   Analogously :
   - $\mathbf{B_u(t = 0) = 1}$ - patient would go blind if receives no treatment $(\mathbf{T = 0})$
   - $\mathbf{B_u(t = 0) = 0}$ - patient would get cured if receives no treatment $(\mathbf{T = 0})$

Remarks

- Both of these potential outcomes are assumed to be **deterministic**.
- If $\mathbf{B_u(t = 1) = 0}$ or $\mathbf{B_u(t = 0) = 1}$ the treatment has a **positive effect** on **u**
- In practice one can not check these assumptions.
  According to the *"fundamental problem of causal inference"* [Holland, 1986], for each unit **u** we can observe either $\mathbf{B_u(t = 1)}$ or $\mathbf{B_u(t = 0)}$ and never both of them at the same time.
- One can observe **only one** of the potential outcomes; the unobserved quantity becomes a counterfactual.

Potential Outcomes. Description & Examples

Unit u	Treatment T	Pot. Outcome $B_u(t=0)$	Pot. Outcome $B_u(t=1)$	Unit-Level Causal Effect $B_u(t=1)-B_u(t=0)$
1	1	1	0	-1
2	0	1	0	-1
3	1	1	0	-1
⋮				
43	1	1	0	-1
44	0	0	1	1
45	0	1	0	-1
⋮				
119	1	1	0	-1
120	1	0	1	1
121	0	1	0	-1
⋮				
200	0	1	0	-1

Figure 1: Table for the Example 3.4 using potential outcomes. For each patient **u**, we observe only one of the two potential outcomes. The observed information has a gray background.

Potential Outcomes. Description & Examples

To justify latter results one needs to fulfill **the stable unit treatment value assumption (SUTVA)**

1. The units do not interfere (e.g., the potential outcome of a unit does not depend on which treatment any other unit received)
2. The potential outcomes do not depend on how or why the treatment has been received.

SUTVA is satisfied when the data are generated from an SCM
For this example,we have sampled 200 i.i.d. units using Bernoulli distributions $N_T \sim Ber(0.6)$ and $N_B \sim Ber(0.01)$. The i.i.d. assumption implies that the units do not interfere with each other and modularity (intervening on T changes only the structural assignment for T) yields that the way the treatment is taken does not influence the result

Potential Outcomes. Description & Examples

The potential outcomes tell us the effect of a treatment on an individual basis;

Unit-Level Causal Effect

$$\mathbf{ULCE} = \mathbf{B_u(t = 1) - B_u(t = 0)} \tag{1}$$

Average Causal Effect

$$\mathbf{CE} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{u=1}}^{\mathbf{n}} \mathbf{B_u(t = 1) - B_u(t = 0)} \tag{2}$$

Potential Outcomes. Description & Examples

The "fundamental problem of causal inference" prevents us from computing **(2)** directly.
Assume that in a completely randomized experiment, units $\mathbf{u \in U_o \subset \{1, \ldots n\}}$ received no treatment $\mathbf{T{=}0}$ and $\mathbf{u \in U_1 = U_o^C}$ received treatment $\mathbf{T{=}1}$

Unbiased Estimator for CE

$$\hat{\mathbf{CE}} := \frac{1}{\#\mathbf{U_1}} \sum_{\mathbf{u \in U_1}} \mathbf{B_u(t = 1)} - \frac{1}{\#\mathbf{U_0}} \sum_{\mathbf{u \in U_0}} \mathbf{B_u(t = 0)} \qquad (3)$$

Remarks

- The randomness in $\hat{\mathbf{CE}}$ comes from the random assignments that determine, which of the unit's two potential outcomes we observe;
- The outcomes themselves are considered hidden, not random.
- Note that **(3)** contains only observed quantities and can therefore be computed after the study has been conducted

Relation between Potential Outcomes and SCMs

In SCMs, we can represent potential outcomes using the language of counterfactuals. Recall the definition of the SCM:

SCM $\mathfrak{C}$ for the eye disease

$$\mathbf{T} = \mathbf{N_T}$$

$$\mathbf{B} = \mathbf{T} \cdot \mathbf{N_B} + (\mathbf{1} - \mathbf{T}) \cdot (\mathbf{1} - \mathbf{N_B})$$

Relation between Potential Outcomes and SCMs

For example, patient 43 has $N_T = 1$ and $N_B = 0$, while patient 44 has $N_T = 0$ and $N_B = 1$. That is two terms $t = 0$ and $t = 1$ correspond to **interventions** on **T**. Summarizing, we have the following

SCM $\mathfrak{C}$ for the eye disease

$$\underbrace{B_u(t = \tilde{t})}_{\text{potential outcome}} = \underbrace{B \text{ in } SCM\ \mathfrak{C}|N = n_u \ \ do(T := \tilde{t})}_{\text{counterfactual SCM}}$$

where $\mathbf{n_u}$ characterizes unit **u** [Pearl, 2009, Equation (3.51)]. Since in the counterfactual SCM all noise terms are deterministic, the entailed distribution of B is degenerate, too, and B is deterministic (as required)

Relation between Potential Outcomes and SCMs

According to Pearl [2009, 7.3.1] and Halpern [2000] if certain properties(axioms) hold for both SCMs and potential outcomes frameworks, it can be shown that these properties are complete for counterfactual SCMs.

We can conclude that any theorem that holds for counterfactual SCMs holds in the world of potential outcomes and vice versa.

Relation between Potential Outcomes and SCMs

The two worlds differ in their language. Even if every theorem holds true in both frameworks, some theorems might be "easier" to prove in one world than in the other.

- Working with settings, in which the average causal effect is zero but the individual causal effects are non-zero, seems to be easier for potential outcomes.
- The graphical representation of SCMs, on the other hand, might be beneficial to exploit assumptions on the causal relations between random variables
- Richardson and Robins [2013] propose to use **single world intervention graphs**. These graphs allow us to set variables to certain values and therefore construct graphical correspondences to counterfactual variables.

Structure Identifiability

Problem: The class of SCMs is too flexible. Given a distribution $\mathbf{P_X}$ over random variables $\mathbf{X} = (X_1, ..., X_d)$, can different SCMs entail this distribution?

Answer: indeed, usually for many different graph structures, there is an SCM that induces the distribution $\mathbf{P_X}$.

Proposition 7.1 (Non-uniqueness of graph structures)

Consider a random vector $\mathbf{X} = (X_1, ..., X_d)$ with distribution $\mathbf{P_X}$ that has as density with respect to Lebesgue measure and assume it is Markovian with respect to $\mathcal{G}$. Then there exists an SCM $\mathfrak{C} = (\mathbf{S}, \mathbf{P_N})$ with graph $\mathcal{G}$ that entails the distribution $\mathbf{P_X}$

Structure Identifiability

Proposition 4.1 (Non-uniqueness of graph structures)

For every joint distribution $\mathbf{P}_{\mathbf{X},\mathbf{Y}}$ of two real-valued variables, there is an SCM

$$Y = f_Y(X, N_Y), \quad X \perp\!\!\!\perp Y$$

where f_Y is a measurable function and N_Y is a real-valued noise variable

Structure Identifiability

If the distribution $\mathbf{P_X}$ is **Markovian** and **faithful** with respect to the underlying DAG $\mathcal{G}$, we have a one-to-one correspondence between d-separation statements in the graph $\mathcal{G}$ and the corresponding conditional independence statements in the distribution. All graphs outside the correct Markov equivalence class of $\mathcal{G}$ can therefore be rejected because they impose a set of d-separations that does not equal the set of conditional independences in $\mathbf{P_X}$.

Structure Identifiability

Since both the Markov condition and faithfulness put restrictions only on the conditional independences in the joint distribution, we are not able to distinguish between two Markov equivalent graphs, that is, between two graphs that entail exactly the same set of conditional independences.

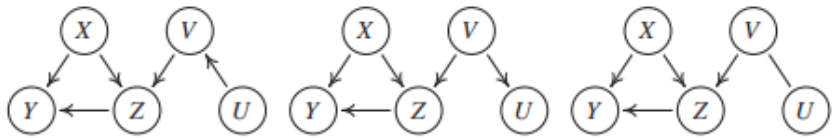

Figure 2: Two Markov equivalent DAGs (left and center) and CPDAG on the right-hand side

Structure Identifiability

Lemma 7.2 (Identifiability of Markov equivalence class)

Assume that $\mathbf{P_X}$ is Markovian and faithful with respect to $\mathcal{G}$. Then, for each graph $\mathcal{G} \in \text{CPDAG}(\mathcal{G})$, we find an SCM that entails the distribution $\mathbf{P_X}$. Furthermore, there is no graph $\mathcal{G}$ with $\mathcal{G} \notin \text{CPDAG}(\mathcal{G})$, such that $\mathbf{P_X}$ is Markovian and faithful with respect to $\mathcal{G}$

Additive Noise Models

Proposition 7.1 shows that a given distribution could have been entailed from several SCMs with different graphs.

Definition 7.3 (ANMs)

We call an SCM $\mathfrak{C}$ an ANM if the structural assignments are of the form

$$X_j = f_j(\mathbf{PA_j}) + N_j, \quad j = 1, ..., d \tag{4}$$

if the noise is additive. For simplicity, the functions f_j are differentiable and the noise variables N_j have a strictly positive density

We obtain non-trivial identifiability results, if we restrict the function class.

Additive Noise Models

Some of the following identifiability results assume causal minimality ($A \perp\!\!\!\perp B|C \Rightarrow A \perp\!\!\!\perp_{\mathcal{G}} B|C$). For ANMs, this means that each function f_j is not constant in any of its arguments.

Proposition 7.4 (Causal minimality and ANMs)

Consider a distribution induced by a model (7.1) and assume that the functions $\mathbf{f_j}$ are not constant in any of its arguments, that is, for all $\mathbf{j}$ and $\mathbf{i} \in \mathbf{PA_j}$ there is some value $\mathbf{pa_{j,-i}}$ of the variables $\mathbf{PA_j \backslash \{i\}}$ and some $\mathbf{x_i \neq x_i'}$ such that

$$\mathbf{f_j(pa_{j,-i}, x_i) \neq f_j(pa_{j,-i}, x_i')}$$

Then the joint distribution satisfies causal minimality with respect to the corresponding graph. Conversely, if there are nodes $\mathbf{j}$ and $\mathbf{i}$ such that for all $\mathbf{pa_{j,-i}}$ the function $\mathbf{f_j(pa_{j,-i}, \cdot)}$ is constant, causal minimality is violated.

Linear Gaussian Models with Equal Error Variances

There is another deviation from linear Gaussian SEMs that makes the graph identifiable via restricting the noise variables to have the same variance is sufficient to recover the graph structure.

Proposition 7.5 (Identifiability with equal error variances)

Consider an SCM with graph $\mathcal{G}_0$ and assignments

$$X_j := \sum_{k \in \mathbf{PA_j}^{\mathcal{G}}} \beta_{jk} X_k + N_j, \quad j = 1, ..d, \tag{5}$$

where all N_j are i.i.d. and follow a Gaussian distribution. In particular, the noise variance σ^2 does not depend on j. Additionally, for each $j \in \{1...p\}$ we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA_j}$ Then, the graph $\mathcal{G}_0$ is identifiable from the joint distribution.

Linear Non-Gaussian Acyclic Models

Shimizu et al. [2006] prove the following statement using independent component analysis (ICA) [Comon, 1994, Theorem 11]

Theorem 7.6 (Identifiability of LiNGAMs)

Consider an SCM with graph $\mathcal{G}_0$ and assignments

$$X_j := \sum_{k \in \mathbf{PA_j}^{\mathcal{G}}} \beta_{jk} X_k + N_j, \quad j = 1, ..d, \tag{6}$$

where all N_j are are jointly independent and non-Gaussian distributed with strictly positive density.Additionally, for each $j \in \{1...p\}$ we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA_j}$ Then, the graph $\mathcal{G}_0$ is identifiable from the joint distribution.

Nonlinear Gaussian Additive Noise Models

The graph structure of an ANM becomes identifiable if the assignments are linear and the noise variables are non-Gaussian. Alternatively, we can also exploit nonlinearity. The result is easiest to state with Gaussian noise:

Nonlinear Gaussian Additive Noise Models

Theorem 7.7 (Identifiability of nonlinear Gaussian ANMs)

- Let $\mathbf{P_X} = \mathbf{P_{X_1,\dots,X_d}}$ be induced by an SCM with

$$X_j = f_j(\mathbf{PA_j}) + N_j,$$

with normally distributed noise variables $N_j \sim N(0, \sigma_j^2)$ and three times differentiable functions f_j that are not linear in any component in the following sense. Denote the parents $\mathbf{PA_j}$ of X_j by $X_{k_1}, \dots, X_{k_l}$ then the function $f_j(x_{k_1}, \dots x_{k_{a-1}}, \cdot, x_{k_{a+1}} \dots, x_{k_l})$ is assumed to be nonlinear for all a and some $x_{k_1}, \dots x_{k_{a-1}}, \cdot, x_{k_{a+1}} \dots, x_{k_l} \in \mathbb{R}^{l-1}$

Nonlinear Gaussian Additive Noise Models

Theorem 7.7 (Identifiability of nonlinear Gaussian ANMs)

- As a special case, let $\mathbf{P_X} = \mathbf{P_{X_1,\ldots,X_d}}$ be induced by an SCM with

$$X_j := \sum_{k \in \mathbf{PA_j}} f_{jk}(X_k) + N_j, \quad j = 1, ..d, \tag{7}$$

  with normally distributed noise variables $N_j \sim N(0, \sigma_j^2)$ and three times differentiable, nonlinear functions f_{jk}. This model is known as a causal additive model (CAM).

In both cases one can identify the corresponding graph $\mathcal{G}_0$ from the distribution $\mathbf{P_X}$. The statements remain true if the noise distributions for source nodes, that is, nodes without parents, are allowed to have a non-Gaussian density with full support on the real line $\mathbb{R}$.

Summary of Some Known Identifiability Results for GN

Type of structural assignment		Condition on funct.	DAG identif.
(General) SCM:	$X_j := f_j(X_{\mathbf{PA}_j}, N_j)$	—	✗
ANM:	$X_j := f_j(X_{\mathbf{PA}_j}) + N_j$	nonlinear	✓
CAM:	$X_j := \sum_{k \in \mathbf{PA}_j} f_{jk}(X_k) + N_j$	nonlinear	✓
Linear Gaussian:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✗
Lin. G., eq. error var.:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✓

Figure 3: Summary of some known identifiability results for Gaussian noise. Results for non-Gaussian noise identifiability results are available, too, but they are more technical.

Thank you for your attention!

Reference

1. Elements of Causal Learning; Petersonas; Janzing, Dominik; Schölkopf, Bernhard.