# Causal Inference and Machine Learning
# Session 6: Merging causal discovery and causal effect estimation

Alessandro Mascaro

February 17th, 2025

Barcelona School of Economics,
Master's degree in Data Science Methodology

# Introduction to the course

# Welcome!

- So far, you have learnt
  - **(i)** How to identify and estimate causal effects from experimental and observational data when knowledge of the causal structure is known or assumed (**causal inference**)
  - **(ii)** How to learn as much as possible of a causal structure from observational data (**causal discovery**)
- In this second part of the course, we will focus on
  - **O1** How to merge **(i)** and **(ii)** to identify and estimate causal effects when the causal structure is unknown;
  - **O2** How to use experimental data to improve the identifiability of causal structures.
- We will do so both in the frequentist and in the Bayesian setting, trying to understand the relative advantages of both approaches!

# Applications

- The combination of causal discovery and causal effect estimation has applications across a wide range of fields. For instance:
  - **Genomics** Interest in estimating the causal effect of activating/deactivating one gene on the activation/deactivation of another gene which is associated with a particular phenotypical feature. Causal structure among genes is unknown and must be estimated from observational data or from combinations of observational and other experimental data;
  - **Economics** Interest in identifying direct causal effects. For instance, one may want to understand whether the class size influences students outcomes: if smaller classes boost outcomes primarily through higher teacher-student interaction (as opposed to improved peer dynamics), then policies that directly enhance teacher-student engagement could yield similar improvements at lower cost.

# Goal and disclaimers

- The goal of this short course is to provide you with the fundamental concepts and intuition that you need to understand this class of methods;
- Because of this, we will mainly focus on simple parametric settings, often assuming to work on jointly continuous Gaussian i.i.d. data. This is clearly limiting in applications, but necessary to focus on the fundamental ideas;
- Applications of the methods taught in this course to more challenging settings can be part of a thesis project!

# Syllabus

**L1** Causal inference with unknown structure: identification of causal effects via covariate adjustment, the IDA algorithm.

**L2** Bayesian causal inference with unknown structure: introduction to Bayesian model selection for causal discovery. Sampling schemes and Bayesian model averaging estimation of causal effects.

**L3** Combining observational and experimental data for causal discovery: the GIES algorithm. Causal effect identification and estimation in this setting.

**L4** Combining observational and experimental data in the Bayesian setting: The unknown-targets case. Experimental design for causal discovery.

**L5** Beyond faithfulness: permutation-based methods for causal discovery (GSP, IGSP, UT-IGSP) and Bayesian equivalents (minimal I-MAP MCMC).

# Evaluation

- This part of the course will be evaluated through one assignment and, for those who choose it, one project:
- Assignment due by February 28th, based on **L1** and **L2**. Rules and modalities are the same as for the first part of the course;
- Project due by March 18th, based on **L3**, **L4** and **L5**.

# Causal discovery and causal effect estimation brushup

# Structural Causal Models

- A **Structural Causal Model** (SCM) describes causal relationships among variables.
- For a variable $X_j$, with $j \in [q] := \{1, \ldots, q\}$, the SCM specifies:

$$X_j = f_j(X_{C_j}, \epsilon_j), \quad j \in [q]$$

  where:

  - $X_{C_j}$ is the set of direct causes of $X_j$,
  - $\epsilon_j$ is the exogenous (error) component.
  - $f_j(\cdot)$ is the *mechanism*, a function mapping the causes and exogenous component to the $j$-th variable;

- Each mechanism is assumed to be **stable and autonomous**: It remains invariant under interventions on other parts of the system $\rightarrow$ essential feature to define and identify causal effects!

# Graphical representation

- The **causal structures** of SCMs can be represented as **Directed Graphs**:
  - Each variable $X_j$ is a node.
  - Directed edges are drawn from each cause $X_{C_j}$ to its effect $X_j$
- If **acyclicity** is assumed, the causal structure is a **Directed Acyclic Graph** (DAG) and the causes of $X_j$ correspond to its parents in the DAG: $X_{C_j} = X_{pa_j(\mathcal{D})}$

# From Causal to Statistical Models - 1

- As defined, the SCM says very little about what a particular causal structure may imply on an observed sample. However, we can make some **additional assumptions**, for instance
  1. **Causal Sufficiency**: All common causes are included in the model (i.e., no hidden confounders).
  2. **Independent Errors**: The exogenous components are independent:

$$p(\epsilon_1, \ldots, \epsilon_p) = \prod_{j=1}^{p} p(\epsilon_j).$$

  3. **Additive Errors**: For all $j \in [q]$, the SCM has mechanisms of the form:

$$X_j = f_j(X_{C_j}) + \epsilon_j$$

# From Causal to Statistical Models - 2

instead of full joint distrtribution where we need all conditional dependencies, we only need the parents node now!

- Under these assumptions, the joint pdf of $X := (X_1, \ldots, X_p)$ factorizes as:

$$p(x_1, \ldots, x_p) = \prod_{j=1}^{p} p\big(x_j \mid \boldsymbol{x}_{pa_j}\big),$$

  i.e., the causal structure $\mathcal{D}$ implies a set of conditional independencies $\mathcal{I}(\mathcal{D})$, namely that all the nodes are independent of their non-descendants given their parents.
  This is known as the **causal Markov assumption**.

# The Gaussian case - 1

- In this course, we will focus for semplicity on the Gaussian setting;
- In this case, the SCM is just a linear Structural Equation Model (SEM) with Normal error components:

$$X = \boldsymbol{B}^T X + \epsilon, \qquad \epsilon \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{D})$$

where

1. $X$ is the vector $(X_1, \ldots, X_p)$ of endogenous variables;
2. $\epsilon$ is the vector $(\epsilon_1, \ldots, \epsilon_p)$ of exogenous, independent Gaussian random variables with diagonal covariance matrix $\boldsymbol{D}$;
3. $\boldsymbol{B}$ is a triangular (up to a permutation) matrix of regression coefficients such that $\boldsymbol{B}_{ij} \neq 0 \iff i \to j \in \mathcal{D}$

# The Gaussian case - 2

- Equivalently, we can also write

$$\boldsymbol{L}^T X = \epsilon, \qquad \epsilon \sim \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{D}\right)$$

where $\boldsymbol{L} = \boldsymbol{I}_p - \boldsymbol{B}$ or

$$X \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{D}}), \qquad \boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{L}^{-T}\boldsymbol{D}\boldsymbol{L}^{-1}$$

# The Gaussian case - 2

- Equivalently, we can also write

$$\boldsymbol{L}^T X = \epsilon, \qquad \epsilon \sim \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{D}\right)$$

  where $\boldsymbol{L} = \boldsymbol{I}_p - \boldsymbol{B}$ or

$$X \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{D}}), \qquad \boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{L}^{-T}\boldsymbol{D}\boldsymbol{L}^{-1}$$

- In the Gaussian case, the Markov property is reflected in the sparsity pattern of the matrix $\boldsymbol{L}$, where the non-null off-diagonal elements of the $j$-th column correspond to the parents (i.e. causes) of $X_j$;

# Causal discovery

- **Causal Discovery** is the game of inferring the underlying causal DAG from data and weaker assumptions on the SCM;

# Causal discovery

- **Causal Discovery** is the game of inferring the underlying causal DAG from data and weaker assumptions on the SCM;
- Typically, when only observational data is available, it relies on the following assumptions:
  - **Sufficiency**
  - **Markovianity**
  - **Faithfulness**: no more conditional independencies in the data than the one implied by the Markov assumption;

# Causal discovery

- **Causal Discovery** is the game of inferring the underlying causal DAG from data and weaker assumptions on the SCM;
- Typically, when only observational data is available, it relies on the following assumptions:
  - **Sufficiency**
  - **Markovianity**
  - **Faithfulness**: no more conditional independencies in the data than the one implied by the Markov assumption;
- Under these assumption, there is a mapping between conditional independencies and causal structures: we can "test" for conditional independencies to learn about the causal structure!

# SCMs and causal inference

# Hard interventions

- **Structural Causal Models** (SCMs) are made of stable and autonomous mechanisms.
- A **hard intervention** fixes the value of the target variable $X_t$ to a constant $\tilde{x}_t$
- We denote this intervention by $\text{do}(X_t = \tilde{x}_t)$ or $\text{do}(\tilde{x}_t)$
- Under such an intervention, only the mechanism for $X_t$ is modified; all other mechanisms remain unchanged.

# The Post-Intervention Distribution

- A hard intervention induces a post-intervention distribution;
- By the stability and autonomy assumptions, the SCM implies that the post-intervention joint distribution **truncates** only the factor corresponding to the intervened variable.
- Specifically, for a realization $x$ of $X$, the post-intervention pdf becomes:

$$p(x \mid do(X_t = \tilde{x}_t)) = \prod_{j \neq t} p\big(x_j \mid x_{\mathrm{pa}_j(\mathcal{D})}\big) \mathbf{1}\{x_t = \tilde{x}_t\}$$

# Defining total causal effects

- The **total causal effect** $\gamma_{ty}$ of a continuous treatment $X_t$ on an outcome $Y$ (with $X_o \equiv Y$) is defined as the derivative of the post-intervention expected value of $Y$ with respect to the fixed value $\tilde{x}_t$:

$$\gamma_{ty} = \frac{\partial \mathbb{E}\big[Y \mid do(X_t = \tilde{x}_t)\big].}{\partial \tilde{x}_t}$$

- This derivative quantifies how the expected outcome $Y$ changes as we vary the intervention value.

# Marginal post-intervention distribution

- The definition of $\gamma_{ty}$ includes a post-intervention expected value. To compute it, we first need to compute the marginal post-intervention distribution of $Y$, which can be easily obtained by marginalization

$$p\left(y \mid do(X_t = \tilde{x}_t)\right) = \int p\left(x \mid do(X_t = \tilde{x}_t)\right) \, d\{x \backslash y\},$$

# Marginal post-intervention distribution

- The definition of $\gamma_{ty}$ includes a post-intervention expected value. To compute it, we first need to compute the marginal post-intervention distribution of $Y$, which can be easily obtained by marginalization

$$p\left(y \mid do(X_t = \tilde{x}_t)\right) = \int p\left(x \mid do(X_t = \tilde{x}_t)\right) \, d\{x \backslash y\},$$

- Given the post-intervention distribution of $Y$, its expected value is

$$\mathbb{E}\big[Y \mid do(X_t = \tilde{x}_t)\big] = \int y \, p\left(y \mid do(X_t = \tilde{x}_t)\right) \, dy,$$

# Adjustment sets

- Let $X_t, Y$ be a treatment and an outcome node, and $\boldsymbol{Z}$ be a set of nodes in a causal DAG $\mathcal{D}$ such that $\boldsymbol{Z} \cap \{X_t, Y\} = \emptyset$. $\boldsymbol{Z}$ is an **adjustment set** relative to $X_t$ and $Y$ in $\mathcal{D}$ if for any pdf consistent with $\mathcal{D}$ we have:

$$p\left(y \mid do(X_t = \tilde{x}_t)\right) = \int p(y \mid \tilde{x}_t, \boldsymbol{z})\, p(\boldsymbol{z})\, d\boldsymbol{z},$$

post intervention(with do)

observed in data

# Adjustment sets

- Let $X_t, Y$ be a treatment and an outcome node, and $\boldsymbol{Z}$ be a set of nodes in a causal DAG $\mathcal{D}$ such that $\boldsymbol{Z} \cap \{X_t, Y\} = \emptyset$. $\boldsymbol{Z}$ is an **adjustment set** relative to $X_t$ and $Y$ in $\mathcal{D}$ if for any pdf consistent with $\mathcal{D}$ we have:

$$p\left(y \mid do(X_t = \tilde{x}_t)\right) = \int p(y \mid \tilde{x}_t, \boldsymbol{z})\, p(\boldsymbol{z})\, d\boldsymbol{z},$$

- Note that the right-hand side depends only on the **observational** (pre-intervention) distribution: the post-intervention densities can be identified and estimated from observational data!

# Backdoor Criterion and Covariate Adjustment

- It can be easily shown that the parents of the treatment node are a valid adjustment set

$$p\left(y \mid do(X_t = \tilde{x}_t)\right) = \int p(y \mid \tilde{x}_t, x_{pa_t(\mathcal{D})})\, p(x_{pa_t(\mathcal{D})})\, dx_{pa_t(\mathcal{D})},$$

# Backdoor Criterion and Covariate Adjustment

- It can be easily shown that the parents of the treatment node are a valid adjustment set

$$p\left(y \mid do(X_t = \tilde{x}_t)\right) = \int p(y \mid \tilde{x}_t, x_{pa_t(\mathcal{D})})\, p(x_{pa_t(\mathcal{D})})\, dx_{pa_t(\mathcal{D})},$$

- However, it is not the only one! It is an adjustment set if it satisfies the **backdoor criterion**, i.e.
  - No element in $\boldsymbol{Z}$ is a descendant of $X_t$;
  - $\boldsymbol{Z}$ blocks all back-door paths from $X_t$ to $Y$;

# The Gaussian case - 3

- If $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, conditional expectations are linear, hence, if $\boldsymbol{Z}$ is an adjustment set:

$$\begin{aligned}
\mathbb{E}\big[Y \mid do(X_t = \tilde{x}_t)\big] &= \int_y y \int_z p\left(y \mid \tilde{x}_t, \boldsymbol{z}\right)\, p(\boldsymbol{z})\, d\boldsymbol{z}\, dy \\
&= \int_z \int_y y\, p\left(y \mid \tilde{x}_t, \boldsymbol{z}\right)\, dy\, p(\boldsymbol{z})\, d\boldsymbol{z} \\
&= \int \mathbb{E}\big[Y \mid \tilde{x}_t, \boldsymbol{z}\big]\, p(\boldsymbol{z})\, d\boldsymbol{z} \\
&= \int (\beta_0 + \beta_{ty}\tilde{x}_t + \boldsymbol{\beta}_{zy}\boldsymbol{z}) p(\boldsymbol{z}) d(z) \\
&= \beta_0 + \beta_{ty}\tilde{x}_t + \boldsymbol{\beta}_{zy}\mathbb{E}(\boldsymbol{Z})
\end{aligned}$$

# The Gaussian case - 3

- If $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, conditional expectations are linear, hence, if $\boldsymbol{Z}$ is an adjustment set:

$$\begin{aligned}
\mathbb{E}\big[Y \mid do(X_t = \tilde{x}_t)\big] &= \int_y y \int_z p\left(y \mid \tilde{x}_t, \boldsymbol{z}\right)\, p(\boldsymbol{z})\, d\boldsymbol{z}\, dy \\
&= \int_z \int_y y\, p\left(y \mid \tilde{x}_t, \boldsymbol{z}\right)\, dy\, p(\boldsymbol{z})\, d\boldsymbol{z} \\
&= \int \mathbb{E}\big[Y \mid \tilde{x}_t, \boldsymbol{z}\big]\, p(\boldsymbol{z})\, d\boldsymbol{z} \\
&= \int (\beta_0 + \beta_{ty}\tilde{x}_t + \boldsymbol{\beta}_{zy}\boldsymbol{z}) p(\boldsymbol{z}) d(z) \\
&= \beta_0 + \beta_{ty}\tilde{x}_t + \boldsymbol{\beta}_{zy}\mathbb{E}(\boldsymbol{Z})
\end{aligned}$$

- By the definition of causal effect given before

$$\gamma_{ty} = \frac{\partial \mathbb{E}\big[Y \mid do(X_t = \tilde{x}_t)\big]}{\partial \tilde{x}_t} = \beta_{ty}$$

# Estimating the Causal Effect

- The do-calculus provides an **estimand**, a population-quantity of interested to be estimated;
- Once defined the estimand, one has to choose the **estimator**

# Estimating the Causal Effect

- The do-calculus provides an **estimand**, a population-quantity of interested to be estimated;
- Once defined the estimand, one has to choose the **estimator**
- In the case of identification via covariate adjustment, a natural choice would be to use **Ordinary Least Squares** (OLS), so that for a given $(n, p)$ data matrix $\boldsymbol{X}$, we have

$$\hat{\gamma}_{ty}^{OLS} := \left[ (\boldsymbol{X}_{\tilde{\boldsymbol{Z}}}^T \boldsymbol{X}_{\tilde{\boldsymbol{Z}}})^{-1} (\boldsymbol{X}_{\tilde{\boldsymbol{Z}}} \boldsymbol{X}_Y) \right]_1$$

where $\tilde{\boldsymbol{Z}} := (X_t, \boldsymbol{Z})$, $\boldsymbol{Z}$ is an adjustment set and $\boldsymbol{X}_{\tilde{\boldsymbol{Z}}}$ denotes the corresponding columns of $\boldsymbol{X}$;

# Estimating the Causal Effect

- The do-calculus provides an **estimand**, a population-quantity of interested to be estimated;
- Once defined the estimand, one has to choose the **estimator**
- In the case of identification via covariate adjustment, a natural choice would be to use **Ordinary Least Squares** (OLS), so that for a given $(n, p)$ data matrix $\boldsymbol{X}$, we have

$$\hat{\gamma}_{ty}^{OLS} := \left[ (\boldsymbol{X}_{\tilde{\boldsymbol{Z}}}^T \boldsymbol{X}_{\tilde{\boldsymbol{Z}}})^{-1} (\boldsymbol{X}_{\tilde{\boldsymbol{Z}}} \boldsymbol{X}_Y) \right]_1$$

  where $\tilde{\boldsymbol{Z}} := (X_t, \boldsymbol{Z})$, $\boldsymbol{Z}$ is an adjustment set and $\boldsymbol{X}_{\tilde{\boldsymbol{Z}}}$ denotes the corresponding columns of $\boldsymbol{X}$;
- As there are many adjustment sets, there are also many unbiased estimators of $\hat{\gamma}_{ty}$;

# Maximum Likelihood Estimator

- Another, more involved, estimator is the Maximum Likelihood Estimator. Recall that, in the Gaussian setting, for $j \in [q]$ we have

$$X_j = \boldsymbol{B}^T_{pa_{\mathcal{D}}(j),j} X_{pa_{\mathcal{D}}(j)} + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, \boldsymbol{D}_{jj})$$

# Maximum Likelihood Estimator

- Another, more involved, estimator is the Maximum Likelihood Estimator. Recall that, in the Gaussian setting, for $j \in [q]$ we have

$$X_j = \boldsymbol{B}^T_{pa_{\mathcal{D}}(j),j} X_{pa_{\mathcal{D}}(j)} + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, \boldsymbol{D}_{jj})$$

- In this case, we can derive the MLE as follows:
  1. For each $j \in [q]$, compute the MLE $(\hat{\boldsymbol{B}}_{pa_{\mathcal{D}}(j),j}, \hat{\boldsymbol{D}}_{jj})$;
  2. Construct $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{D}})$ and derive the MLE of $\boldsymbol{\Sigma}_{\mathcal{D}}$ as

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{D}} = (\boldsymbol{I}_p - \hat{\boldsymbol{B}})^{-T} \hat{\boldsymbol{D}} (\boldsymbol{I}_p - \hat{\boldsymbol{B}})^{-1}$$

  3. From $\hat{\boldsymbol{\Sigma}}_{\mathcal{D}}$, derive

$$\gamma_{ty}^{MLE} = (\hat{\boldsymbol{\Sigma}}_{\mathcal{D}})^{-1}_{\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}} (\hat{\boldsymbol{\Sigma}}_{\mathcal{D}})_{\tilde{\boldsymbol{Z}}Y}$$

# Which estimator? - 1

- Different estimators have different statistical properties, which would you choose? Why?
- Let's make some basic numerical experiments...

# Which estimator? - 2

There are many OLS estimators as valid adjustments sets. Each one of them is **unbiased**, but has **different variance**. Their variance is also strictly higher than the MLE, as they not explicitly include the conditional independence information implied by the DAG;

On the other hand, the MLE is unique, but it can become cumbersome to compute when the dimensionality of the data increases!

# Summary and final thoughts

- The starting point is the **Structural Causal Model**
- Each Structural Causal Model comes with a modularity assumption that allows us to define external interventions and causal effects;
- The associated do-calculus allows us to identify causal effects from observational data when possible. Once the estimand is defined, we can estimate it in different ways!
- If the SCM is not known, we can make some weaker assumptions that imply some testable probabilistic relationships among variables, and use them to learn as much as possible of the causal structure! (**causal discovery**);
- If we want to identify and estimate causal effects when the SCM structure is not known, the strategy seems obvious...

# Merging causal discovery and causal effect estimation

# A naive strategy

- If we don't know the causal structure, a natural strategy is to first use a causal discovery algorithm, and then use the structure learnt to identify and estimate the causal effect of interest
- **Problem:**

# A naive strategy

- If we don't know the causal structure, a natural strategy is to first use a causal discovery algorithm, and then use the structure learnt to identify and estimate the causal effect of interest
- **Problem:** even in the simpler settings, usually the assumptions made only allow us to recover an equivalence class of DAGs. Each DAG **may** imply a different causal effect estimand;
- **Simple solution:**

# A naive strategy

- If we don't know the causal structure, a natural strategy is to first use a causal discovery algorithm, and then use the structure learnt to identify and estimate the causal effect of interest
- **Problem:** even in the simpler settings, usually the assumptions made only allow us to recover an equivalence class of DAGs. Each DAG **may** imply a different causal effect estimand;
- **Simple solution:** Enumerate all possible DAGs within a Markov equivalence class, and for each of them derive an estimand and the corresponding estimate, resulting in a set of estimates $\mathbf{\Gamma}_{ty}$;

# A naive strategy

- If we don't know the causal structure, a natural strategy is to first use a causal discovery algorithm, and then use the structure learnt to identify and estimate the causal effect of interest
- **Problem:** even in the simpler settings, usually the assumptions made only allow us to recover an equivalence class of DAGs. Each DAG **may** imply a different causal effect estimand;
- **Simple solution:** Enumerate all possible DAGs within a Markov equivalence class, and for each of them derive an estimand and the corresponding estimate, resulting in a set of estimates $\mathbf{\Gamma}_{ty}$;
- Is this actually feasible?

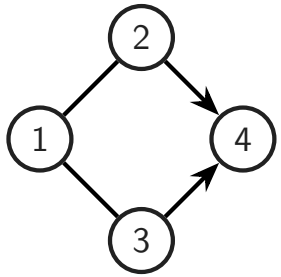
# How big are Markov Equivalence classes

- Determining the number of DAGs in a Markov Equivalence class is a difficult task:
- **Extreme example**: any complete DAG - any DAG without missing edges - implies no conditional independencies and any other complete DAG is Markov equivalent to it.

  $\Longrightarrow$ For a complete DAG with $p$ vertices, there are $p!$ DAGs in its Markov equivalence class! For $p = 10$, it means $3628800$ DAGs.

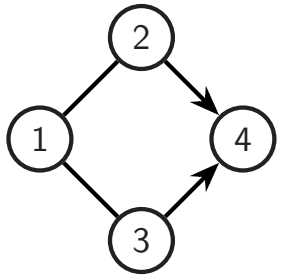
# Enumerating DAGs within the same Markov Equivalence class

- Enumerating all DAGs within a Markov equivalence class is difficult. Let's consider, for instance, the following example

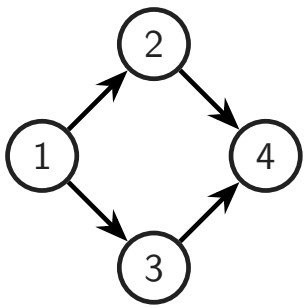


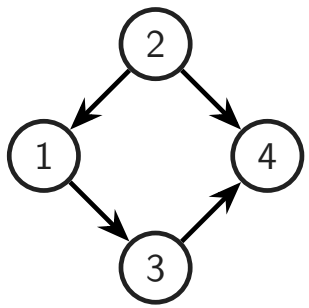
# Enumerating DAGs within the same Markov Equivalence class

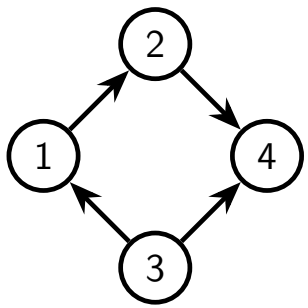
- Enumerating all DAGs within a Markov equivalence class is difficult. Let's consider, for instance, the following example

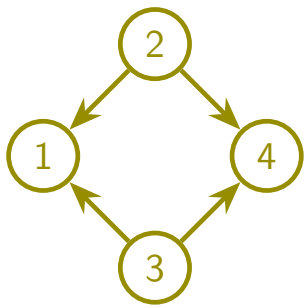



- One may think that, as there are two undirected edges, there must be $2^2$ DAGs within the Markov equivalence class. But:



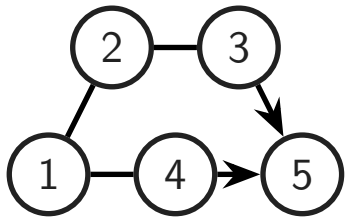



# Overcoming difficulties

- So, enumerating each DAG in the Markov equivalence class and for each identifying the causal effects can easily become a nightmare!
- **Idea:** We actually don't need to enumerate all the DAGs. As the parent set of the treatment node is a valid adjustment set, we can just enumerate all the possible parent sets of the treatment node and compute the set of estimates of the causal effect $\mathbf{\Gamma}_{ty}^{L}$
- For a given CPDAG $\mathcal{G}$, we denote with $\text{sib}_{\mathcal{G}}(j)$ the **siblings** of $j$ in $\mathcal{G}$, i.e. those nodes which are connected to $j$ via an **undirected edge**:
- For any node $j$, its **possible parent set** in the Markov equivalence class represented by $\mathcal{G}$ is $\text{pa}_j(\mathcal{G}) \cup \boldsymbol{S}$, where $\boldsymbol{S} \subseteq \text{sib}_j(\mathcal{G})$

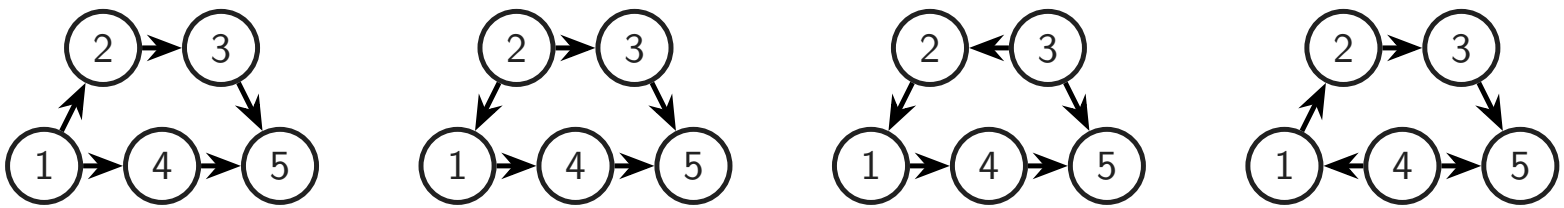
# Possible parents - Example

- Suppose $X := (X_1, \ldots, X_5)$ and that we interested in $\gamma_{15}$.
  A causal discovery algorithm returns the following CPDAG $\mathcal{G}$:




  where $\text{sib}_1(\mathcal{G}) = \{2, 4\}$;

- The DAGs in the corresponding equivalence class are:




- $4$ DAGs in the Markov equivalence class, **but** 3 different parent sets for $X_1$, namely $\{\emptyset, \{2\}, \{4\}\}$. Hence $\mathbf{\Gamma}_{ty}^L = \{\hat{\beta}_{ty}, \hat{\beta}_{ty|2}, \hat{\beta}_{ty|4}\}$

# Local validity

- In the previous example, there is no possible parent set with $\boldsymbol{S} = \{2, 4\}$, as it creates a new v-structure on $X_1$, thus going outside of the Markov equivalence class!
- Let $\mathcal{G}_{S \to t}$ the DAG obtained by directed all the nodes in $\boldsymbol{S}$ towards $t$;
- We say that $\mathcal{G}_{S \to t}$ is **locally valid** if $\mathcal{G}_{S \to t}$ does not contain a new v-structure with $t$ as a collider;
- In our example $\mathcal{G}_{\{2,4\} \to 1}$ is not locally valid!:

# Important Lemma

- Maathuis et al. (2009) prove that $\mathcal{G}_{\boldsymbol{S}\to t}$ is locally valid if and **only if** there is a DAG $\mathcal{D}$ in the equivalence class of $\mathcal{G}$ such that $\mathrm{pa}_{\mathcal{D}}(t) = \mathrm{pa}_{\mathcal{G}}(t) \cup \boldsymbol{S}$.
- The **only if** side of this proposition is particularly important: it means that if $\mathcal{G}_{\boldsymbol{S}\to t}$ is locally valid, then there exists at least one DAG in the Markov equivalence such that the causal effect $\gamma_{ty}$ can be estimated using $\mathrm{pa}_{\mathcal{G}}(t) \cup \boldsymbol{S}$ as adjustment set!

# The IDA Algorithm

- Maathuis et al. (2009) then propose the following algorithm, that they called IDA (Identification when DAG is Absent):

**Algorithm 1:** IDA Algorithm

**Input:** CPDAG $\mathcal{G}$, treatment $t$, outcome $y$

**Output:** Set $\Gamma_{ty}^{L}$

$\mathbf{\Gamma}_{ty}^{L} \leftarrow \emptyset$ **for each** *subset* $S$ *of* $\mathrm{sib}_{\mathcal{G}}(t)$ **do**

  Check if $G_{S \to t}$ is locally valid;

  **if** $G_{S-i}$ *is locally valid* **then** Add $\hat{\beta}_{ty|\mathrm{pa}_{\mathcal{G}}(t) \cup S)}$ to $\mathbf{\Gamma}_{ty}^{L}$;

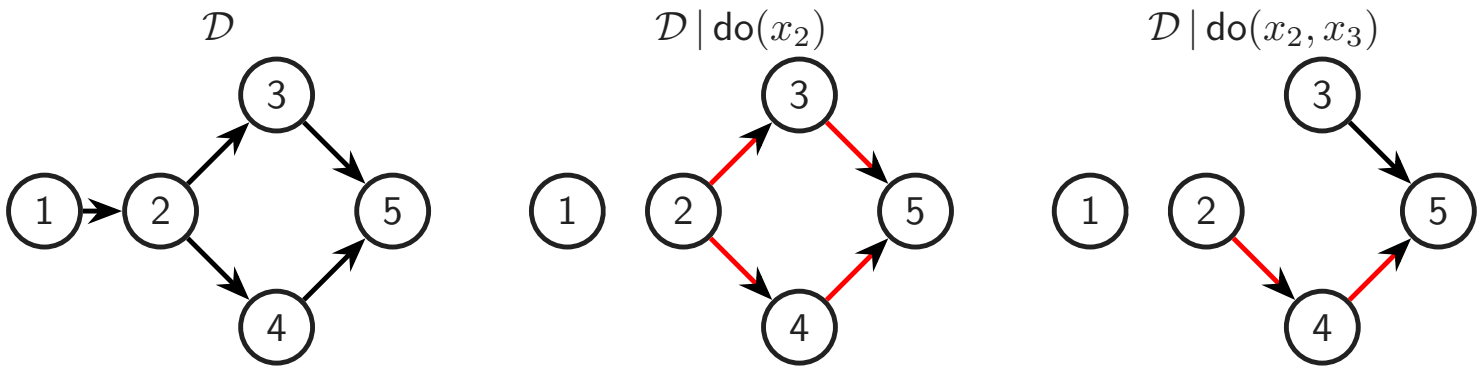
- Algorithm 1 just requires checking whether $\mathcal{G}_{S \to t}$ is locally valid for all the possible subsets of $\boldsymbol{S}$, which is much faster then enumerating all the DAGs and deriving $\mathbf{\Gamma}_{ty}$!
- They also prove that $\mathbf{\Gamma}_{ty} = \mathbf{\Gamma}_{ty}^{L}$!

# Theoretical results

- Maathuis et al. (2009) prove consistency of their method in estimating the set of possible causal effects in high-dimensions (i.e. when $p >> n$), using the PC-algorithm for the causal discovery part and assuming Gaussianity;
- Their method is available in the R package `pcalg`;
- Many extension of this seminal work!

# E1 - Single and joint interventions

- For instance, it has been extended to the case of **joint interventions** and outside the Gaussian setting by Nandy et al. (2016);
- The causal effect of a node in joint interventions can differ from the ones on single nodes. For instance:



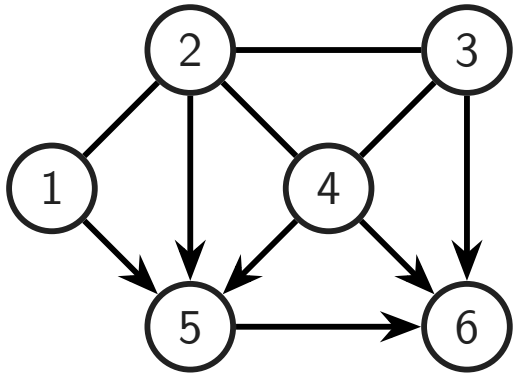

- $\gamma_{25}^{\{2,3\}}$, the causal effect of $X_2$ on $X_5$ in a joint intervention on $\{X_2, X_3\}$ differs because the intervention on $X_3$ "blocks" one of the causal path from $X_2$ to $X_5$ in $\mathcal{D}$!
- We'll see more of this tomorrow, in the Bayesian setting!

# E2 - Adjustment sets for CPDAGs

- IDA just uses the parent set of the treatment node, but it is not the only valid adjustment set!
- Perkovic et al. (2018) provide an adjustment criterion that is sound and complete for DAGs and CPDAGs.
  - **Sound**: If $\boldsymbol{Z}$ satisfies the adjustment criterion, it is a valid adjustment set;
  - **Complete**: If $\boldsymbol{Z}$ is a valid adjustment set, then it satisfies their criterion;
- In other words, $\boldsymbol{Z}$ is a valid adjustment set if and only if it satisfies their criterion;
- Their criterion is a little bit involved, but it allows to identify a valid adjustment set that is valid for an entire ME class (if it exists);

# Generalised adjustment criterion - Example

- Their criterion is a little bit involved, but to understand what it allows to do, consider the following CPDAG:




- By their criterion, in estimating the causal effect of $X_5$ on $X_6$ any superset of $\{2, 4\}$, $\{3, 4\}$ is a valid adjustment set for the entire Markov Equivalence class represented by $\mathcal{G}$.
- We can then choose which one to use based on other considerations, e.g. asymptotic variance of OLS estimates (Henckel et al. 2022);

# Moving to R

- Let's move to R!

# A potential issue - 1

- Isn't there something wrong about this strategy?

# A potential issue - 1

- Isn't there something wrong about this strategy?
- All these methods require a **preliminary causal discovery step** to derive the CPDAG!
- If the two steps are performed on the same data, this is a typical double-dipping problem! Using the data twice may lead, for example, to invalid confidence intervals;

# A potential issue - 2

- Tomorrow we will see how Bayesian approaches do not tackle the two parts as separated steps, but as a unique estimation problem;
- In your first assignment, you will be guided in exploring the consequences of using the data twice in this setting, comparing the performance of frequentist and Bayesian methods!

Thank you!

## References:


- Henckel, L., Perković, E., & Maathuis, M. H. (2022). *Graphical criteria for efficient total effect estimation via adjustment in causal linear models*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(2), 579-599.
- Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). *Estimating High-dimensional Intervention Effects from Observational Data*. The Annals of Statistics, 37(6A), 3133-3164.
- Perkovic, E., Textor, J., Kalisch, M., & Maathuis, M. H. (2018). *Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs.* Journal of Machine Learning Research, 18(220), 1-62.