Causal Inference and Machine Learning
Session 9: Bayesian causal discovery from observational and experimental data

Alessandro Mascaro

March 4th, 2025

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

Introduction

Bayesian causal discovery

- Suppose we observe a (n, p) data matrix of observations $\boldsymbol{X}$, containing i.i.d. samples of a random vector $X := (X_1, \dots, X_p)$, with X distributed according to some DAG model;
- In the Bayesian setting, causal discovery can be tackled as a Bayesian model selection problem, where our target is

$$\boxed{p(\mathcal{D} \mid \boldsymbol{X}) = \frac{m(\boldsymbol{X} \mid \mathcal{D})\, p(\mathcal{D})}{\sum_{\mathcal{D} \in \mathcal{S}_q} m(\boldsymbol{X} \mid \mathcal{D})\, p(\mathcal{D})} \propto m(\boldsymbol{X} \mid \mathcal{D})\, p(\mathcal{D})}$$

  i.e. the posterior distribution over DAG models;
- The posterior distribution must be approximated, typically via **sampling methods**;

Using experimental data

- Experimental data are data measured after an intervention modifying the original mechanisms of the Structural Causal Model;
- When doing a **hard intervention** on a variable, the relationship of that variable with its **causes** is destroyed, while the one with its effects is **preserved**;

$\Longrightarrow$ We can use this asymmetry to learn causal directions using experimental data!

- **Greedy Interventional Equivalence Search** (GIES) is a score-based method for causal discovery that uses both observational and experimental data;

In the Bayesian setting

- Today, we will see how the same problem of causal discovery from observational and experimental data can be tackled in the Bayesian setting;
- Actually, in the second part of the lecture, we will go beyond what GIES does and assume no knowledge of the targets of intervention of each experimental setting!

Bayesian Causal Discovery from experimental data with known intervention targets

Setting

- Suppose we observe a set of K data matrices $\{\boldsymbol{X}^{(k)}\}_{k=1}^{K}$, where
  - $\boldsymbol{X}^{(1)}$ contains i.i.d. samples from the observational distribution of X;
  - $\boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(K)}$ contain samples from the post-intervention distributions of X given a hard intervention on the target nodes $\boldsymbol{T}^{(k)}$.

  and let $\mathcal{T} = \{\boldsymbol{T}^{(k)}\}_{k=1}^{K}$ be the multi-set of intervention targets;
- **Goal:** Derive the posterior distribution

$$\boxed{p(\mathcal{D} \,|\, \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}, \mathcal{T}) \propto m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \mathcal{D}, \mathcal{T})\, p(\mathcal{D} \,|\, \mathcal{T})}$$

  i.e., the posterior distribution over DAGs given data from different experimental settings and knowledge of the intervention targets $\mathcal{T}$;
- **Again**, this will be a **Bayesian Model Selection** (BMS) problem!

Marginal likelihood

- As in all BMS problems, the **marginal likelihood** is of fundamental importance. Assuming that X is distributed according to a parametric distribution with parameter $\Theta_{\mathcal{D}}$, it is defined as:

$$m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \mathcal{D}, \mathcal{T}) = \int_{\Theta_{(\mathcal{D},\mathcal{T})}} p(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \Theta_{(\mathcal{D},\mathcal{T})}, \mathcal{D}, \mathcal{T}) \cdot$$
$$p(\Theta_{(\mathcal{D},\mathcal{T})} \,|\, \mathcal{D}, \mathcal{T}) \, d\Theta_{(\mathcal{D},\mathcal{T})}$$

  where

  - $p(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \Theta_{(\mathcal{D},\mathcal{T})}, \mathcal{D}, \mathcal{T})$ is the likelihood of all the independent measurements in the matrices $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$;
  - $\Theta_{(\mathcal{D},\mathcal{T})}$ is the set of parameters associated with each pre- and post-intervention distribution;

Bayesian Causal Discovery: How to

- To perform BMS, two steps:
  (i) Specify a Bayesian model that will define the posterior distribution. This consists of:
    - $p(\{\boldsymbol{X}\}_{k=1}^{K} \,|\, \Theta, \mathcal{D}, \mathcal{T})$: the statistical model;
    - $p(\Theta_{(\mathcal{D},\mathcal{T})} \,|\, \mathcal{D}, \mathcal{T})$: the parameter prior;
    - $p(\mathcal{D} \,|\, \mathcal{T})$: the model prior;

  (ii) Approximated the posterior distribution via **sampling methods**.
- In what follows, we will focus again on the **Gaussian** setting;

Bayesian causal discovery: Model specification in the Gaussian setting

Statistical model - 1

- Suppose X is generated by a linear Gaussian Structural Equation Model (SEM) with independent error components and causal structure represented by the DAG $\mathcal{D}$.
- As we saw in **L1**, this implies that X is distributed as a Gaussian DAG model, i.e.

$$\begin{aligned} X_1, \dots, X_q \,|\, \boldsymbol{\Sigma}_{\mathcal{D}} &\sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{D}}) \\ \boldsymbol{\Sigma}_{\mathcal{D}} &= \boldsymbol{L}^{-T}\boldsymbol{D}\boldsymbol{L}^{-1} \end{aligned}$$

  where $\boldsymbol{L} = (\boldsymbol{I}_p - \boldsymbol{B})$ and $\boldsymbol{B}$ is the matrix of coefficients of the SEM and $\boldsymbol{B}_{ij} \neq 0 \iff i \to j \in \mathcal{D}$;

Statistical model - 2

- The joint pdf of X *before any intervention* factorizes as

$$p(x \,|\, (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}) = \prod_{j=1}^{p} d\,\mathcal{N}(x_j\,; -\boldsymbol{L}^{\top}_{\mathrm{pa}_{\mathcal{D}}(j),j} x_{\mathrm{pa}_j(\mathcal{D})}, \boldsymbol{D}_{jj})$$

- Consequently, we can write the likelihood of the (n_1, p) data matrix $\boldsymbol{X}^{(1)}$, containing i.i.d. samples from X as

$$p(\boldsymbol{X}^{(1)} \,|\, (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}) = \prod_{j=1}^{p} d\mathcal{N}_{n_1}(\boldsymbol{X}^{(1)}_{\cdot j}\,; -\boldsymbol{X}^{(1)}_{\cdot \mathrm{pa}_j(\mathcal{D})} \boldsymbol{L}_{\mathrm{pa}_j(\mathcal{D}),j}, \boldsymbol{D}_{jj} \boldsymbol{I}_{n_1})$$

Statistical model - 3

- In what follows, we will consider **stochastic** *hard interventions* that set the value of the target nodes to a constant with some noise;
- The post-intervention pdf after a stochastic hard intervention on $\boldsymbol{T}^{(k)}$ factorizes as

$$
\begin{aligned}
p_k(x \mid (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}, \mathcal{T}) = p(x \mid \mathsf{do}(X_{\boldsymbol{T}^{(k)}} = \tilde{x}_{\boldsymbol{T}^{(k)}}, (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}, \mathcal{T}) \\
\prod_{j \notin \boldsymbol{T}^{(k)}} d\mathcal{N}(x_j \,; -\boldsymbol{L}^{\top}_{\mathrm{pa}_{\mathcal{D}}(j),j} x_{\mathrm{pa}_j(\mathcal{D})}, \boldsymbol{D}_{jj}) \cdot \\
\prod_{l \in \boldsymbol{T}^{(k)}} d\mathcal{N}(x_l \,; \tilde{x}_l, \tilde{\boldsymbol{D}}^{(k)}_{ll})
\end{aligned}
$$

  where $\tilde{\boldsymbol{D}}_{ll}$ is the variance associated with the intervention on node l;

Statistical model - 4

- Accordingly, the likelihood of the (n_k, p) data matrix $\boldsymbol{X}^{(k)}$ containing i.i.d. samples from the post-intervention distribution of X is

$$p(\boldsymbol{X}^{(k)} \,|\, (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}, \mathcal{T}) = \prod_{j \notin \boldsymbol{T}^{(k)}} d\mathcal{N}_{n_k}(\boldsymbol{X}^{(k)}_{\cdot j}\,; -\boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_j(\mathcal{D})} \boldsymbol{L}_{\mathrm{pa}_j(\mathcal{D}), j}, \boldsymbol{D}_{jj} \boldsymbol{I}_{n_k})$$

$$\prod_{l \notin \boldsymbol{T}^{(k)}} d\mathcal{N}_{n_k}(\boldsymbol{X}^{(k)}_{\cdot l}\,; \boldsymbol{0}, \tilde{\boldsymbol{D}}^{(k)}_{ll} \boldsymbol{I}_{n_k})$$

  where, for simplicity and wlog, we are assuming that the hard intervention fixes the value of the target variables to 0;

Statistical model - 5

- Putting it all together, we have:

$$
\begin{aligned}
p(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}, \mathcal{T}) &= \prod_{k=1}^{K} p(\boldsymbol{X}^{(k)} \,|\, (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}, \mathcal{T}) \\
&= \prod_{k=1}^{K} \Bigg(\prod_{j \notin \boldsymbol{T}^{(k)}} d\mathcal{N}_{n_k}(\boldsymbol{X}_{.j}^{(k)}; -\boldsymbol{X}_{.\mathrm{pa}_j(\mathcal{D})}^{(k)} \boldsymbol{L}_{\mathrm{pa}_j(\mathcal{D}),j}, \boldsymbol{D}_{jj} \boldsymbol{I}_{n_k}) \cdot \\
&\qquad \prod_{l \notin \boldsymbol{T}^{(k)}} d\mathcal{N}_{n_k}(\boldsymbol{X}_{.l}^{(k)}; \boldsymbol{0}, \tilde{\boldsymbol{D}}_{ll}^{(k)} \boldsymbol{I}_{n_k}) \Bigg)
\end{aligned}
$$

- Can we write it more compactly?

Statistical model - 6

- Denoting with $\mathcal{A}(j) := \{k \in [K] : j \notin \boldsymbol{T}^{(k)}\}$, we can write

$$p(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}, \mathcal{T}) = \prod_{j=1}^{p} \Bigg(d\mathcal{N}_{|\mathcal{A}(j)|}(\boldsymbol{X}_{.j}^{(\mathcal{A}(j))}; -\boldsymbol{X}_{.\mathrm{pa}_j(\mathcal{D})}^{(\mathcal{A}(j))} \boldsymbol{L}_{\mathrm{pa}_j(\mathcal{D}),j}, \boldsymbol{D}_{jj} \boldsymbol{I}_{n_k}) \cdot \prod_{k \notin \mathcal{A}(j)} d\mathcal{N}_{n_k}(\boldsymbol{X}_{.j}^{(k)}; \boldsymbol{0}, \tilde{\boldsymbol{D}}_{jj}^{(k)} \boldsymbol{I}_{n_k}) \Bigg)$$

  where $|\mathcal{A}(j)| = \sum_{k \in A(j)} n_k$
- $\mathcal{A}(j) \subseteq [K]$ is the index-set that, for each node j, specifies those experimental settings in which its conditional distribution is unaffected by the intervention performed;

$\Longrightarrow$ It captures all the **invariances** holding across experimental settings!

Parameter prior - 1

- When only observational data is available, Gaussian DAG models are parameterized in terms of the matrices $(\boldsymbol{L}, \boldsymbol{D})$, where $\boldsymbol{L}$ follows a specific sparsity pattern defined by the DAG $\mathcal{D}$;
- Ben-David et al. (2011) developed the **DAG-Wishart** distribution, which is defined exactly on the space of matrices $(\boldsymbol{L}, \boldsymbol{D})$ of a DAG.
- The DAG-Wishart distribution has a lot of desirable properties: its **marginal likelihood** is available in closed-form, it is decomposable, score equivalent and **consistent** when used as a score;

Parameter prior - 2

- In our setting, the parameter space is a little bit more complicated: there is no "nice" distribution defined on our space that we can use off-the-shelf;
- However, we can still try to specify a prior on the non-null elements of $(\boldsymbol{L}, \boldsymbol{D})$ that are invariant across experimental settings and on the parameters induced by the interventions ($\tilde{\boldsymbol{D}}_{ll}^{(k)}$, for all $k : l \in \boldsymbol{T}^{(k)}$)
- We would still like to have all the nice properties of the DAG-Wishart distribution!

Parameter prior - 3

- Castelletti & Peluso (2024) propose adapting the procedure initially proposed by Geiger & Heckerman (2002) for the case of observational data to our setting;
- The procedure is quite involved, but for each DAG $\mathcal{D}$ and set of node-specific parameters $(\boldsymbol{L}_{\mathrm{pa}_{\mathcal{D}}(j),j}, \boldsymbol{D}_{jj})$ it can be broken down in four main steps:
  1. Identify a complete DAG $\mathcal{C}$ where $\mathrm{pa}_{\mathcal{C}}(j) = \mathrm{pa}_{\mathcal{D}}(j)$ and denote its parameters with $(\bar{\boldsymbol{L}}, \bar{\boldsymbol{D}})$;
  2. Specify an Inverse-Wishart distribution on its unconstrianed covariance matrix $\boldsymbol{\Sigma}_{\mathcal{C}}$

$$\boldsymbol{\Sigma}_{\mathcal{C}} \sim \text{I-W}(a, \boldsymbol{U})$$

  3. Derive the induced distribution on $(\bar{\boldsymbol{L}}_{\mathrm{pa}_{\mathcal{C}}(j),j}, \bar{\boldsymbol{D}}_{jj})$;
  4. Specify as a prior on $(\boldsymbol{L}_{\mathrm{pa}_{\mathcal{D}}(j),j}, \boldsymbol{D}_{jj})$ the distribution derived in 3.

Parameter prior - 4

- To adapt the prior elicitation procedure of Geiger & Heckerman (2002) to our setting, it is sufficient to use it also to specify a prior on $\tilde{\boldsymbol{D}}$, using as a complete DAG $\mathcal{D}$ where $\mathrm{pa}_{\mathcal{C}}(l) = \mathrm{pa}_{\tilde{\mathcal{D}}}(l) = \emptyset$;
- It can be shown that, for each $j \in [p]$, this procedure leads to the following prior specification:

$$\tilde{\boldsymbol{D}}_{jj}^{(k)} \mid \mathcal{D} \sim \text{Inv-Ga}\left(\frac{a-p+1}{2}, \frac{\boldsymbol{U}_{jj}}{2}\right), \qquad k \in [K] \backslash \mathcal{A}(j)$$

$$\boldsymbol{D}_{jj} \mid \mathcal{D} \sim \text{Inv-Ga}\left(\frac{a-p+|pa_j(\mathcal{D})|+1}{2}, \frac{\boldsymbol{U}_{j|\mathrm{pa}_j(\mathcal{D})}}{2}\right),$$

$$\boldsymbol{L}_{\mathrm{pa}_j(\mathcal{D}),j} \mid \boldsymbol{D}_{jj}, \mathcal{D} \sim \mathcal{N}_{|\mathrm{pa}_j(\mathcal{D})|}\left(-\boldsymbol{U}^{-1}_{\mathrm{pa}_j(\mathcal{D})}\boldsymbol{U}_{\mathrm{pa}_j(\mathcal{D}),j}, \boldsymbol{D}_{jj}\boldsymbol{U}^{-1}_{\mathrm{pa}_j(\mathcal{D})}\right),$$

where $\boldsymbol{U}_{j|\mathrm{pa}_j(\mathcal{D})} := \boldsymbol{U}_{jj} - \boldsymbol{U}_{j,\mathrm{pa}_j(\mathcal{D}}(\boldsymbol{U}_{\mathrm{pa}_j(\mathcal{D}),\mathrm{pa}_j(\mathcal{D})})^{-1}\boldsymbol{U}_{\mathrm{pa}_j(\mathcal{D}),j}$

- It is **extremely similar** to the DAG-Wishart prior!

Parameter prior - 5

- Using the prior specification procedure, it can be shown that the **marginal likelihood** factorises as

$$
\begin{aligned}
m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \mid \mathcal{D}, \mathcal{T}) &= \prod_{j=1}^{p} \left(m(\boldsymbol{X}_{.j}^{(\mathcal{A}(j))} \mid \boldsymbol{X}_{.\mathrm{pa}_j(\mathcal{D})}^{(\mathcal{A}(j))}) \cdot \prod_{k \notin \mathcal{A}(j)} m(\boldsymbol{X}_{.j}^{(k)}) \right) \\
&= \prod_{j=1}^{p} \left(\frac{m\left(\boldsymbol{X}_{.\mathrm{fa}_j(\mathcal{D})}^{(\mathcal{A}(j))}\right)}{m\left(\boldsymbol{X}_{.\mathrm{pa}_j(\mathcal{D})}^{(\mathcal{A}(j))}\right)} \cdot \prod_{k \notin \mathcal{A}(j)} m\left(\boldsymbol{X}_{.j}^{(k)}\right) \right)
\end{aligned}
$$

- The marginal likelihood induced by the proposed prior elicitation procedure is **decomposable**;

Parameter prior - 6

- In the Gaussian setting, for any $k \in [K]$ and $B \subseteq [p]$, we have

$$m(\boldsymbol{X}^{(k)}_{.B}) = \pi^{-n|B|/2} \frac{\det(\boldsymbol{U}_{BB})^{(a-p+|B|)/2}}{\det(\boldsymbol{U}_{BB} + \boldsymbol{S}^{(k)}_{BB})^{(a-p+|B|+n)/2}} \frac{\Gamma\left(\frac{a-p+|B|+n}{2}\right)}{\Gamma\left(\frac{a-p+|B|}{2}\right)}$$

  where $\boldsymbol{S}^{(k)} = (\boldsymbol{X}^{(k)})^T(\boldsymbol{X}^{(k)})$
- Plugging this formula into the one for the marginal likelihood returns the **marginal likelihood** for the **Gaussian setting** in **closed-form**;
- Castelletti & Peluso (2024) show that it also satisfies
  - **Score equivalence**: two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ have the same marginal likelihood **if and only if** they are I-Markov equivalent;
  - **Consistency**: As $n \to \infty$, the true I-Markov equivalence class is assigned highest marginal likelihood;

DAG prior

- $p(\mathcal{D} \,|\, \mathcal{T})$ can be specified exactly as in the case with no experimental data, so that $p(\mathcal{D} \,|\, \mathcal{T}) = p(\mathcal{D})$;
- Consider a collection of Bernoulli distributions on 0-1 elements indicating absence/presence of edges in DAG $\mathcal{D}$
- Let $\boldsymbol{S}^{\mathcal{D}}$ be the 0-1 *adjacency matrix* of the skeleton of $\mathcal{D}$:

$$\boldsymbol{S}^{\mathcal{D}}_{u,v} = \begin{cases} 1 & \text{if } u \to v \in \mathcal{D} \text{ or } u \leftarrow v \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

- We assign a prior on the DAGs based on their skeleton. In particular: $\boldsymbol{S}^{\mathcal{D}}_{u,v} \,|\, \pi \overset{\text{iid}}{\sim} \text{Ber}(\pi)$, $u < v$, $\pi \in (0,1)$
- The prior probability assigned to each DAG is thus:

$$\boxed{p(\mathcal{D} \,|\, \pi) = \pi^{|\boldsymbol{S}^{\mathcal{D}}|} (1-\pi)^{\frac{q(q-1)}{2} - |\boldsymbol{S}^{\mathcal{D}}|}}$$

Model specified

- We have now specified the whole Bayesian model for our Bayesian causal discovery problem!
- We just need to compute the posterior distribution;

Sampling from the posterior

- As in the case with no experimental data, we can only approximate the posterior distribution;
- We will use again Markov Chain Monte Carlo algorithms to approximate it via sampling;
- In particular, as the marginal likelihood is available in closed-form, we will use again a **Metropolis-Hastings** (MH) algorithm, based on the following steps
    - Start from an (arbitrary) initial DAG;
    - Given a current DAG $\mathcal{D}$ propose a new candidate DAG $\widetilde{\mathcal{D}}$
    - Accept/reject $\widetilde{\mathcal{D}}$ with probability given by the MH acceptance ratio:
    - Iterate the previous steps for a number of times S

Proposing a new DAG

- The proposal scheme is **exactly the same** as in the case with no experimental data;
- Suppose $\mathcal{D}$ is the current DAG. We **propose** a new candidate DAG $\tilde{\mathcal{D}}$ by **inserting**, **deleting** or **reversing** at random an edge in $\mathcal{D}$ and **checking that the resulting graph is a DAG**!
- In practice, we build the set $\mathcal{O}_{\mathcal{D}}$ of all possible DAGs that can be reached from $\mathcal{D}$ and sample uniformly at random from it;

Acceptance/Rejection step

- Given a current DAG $\mathcal{D}$, a new DAG $\widetilde{\mathcal{D}}$ drawn from the proposal $q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})$ is accepted with probability

$$\alpha_{\widetilde{\mathcal{D}},\mathcal{D}} = \min\left\{1; \frac{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \widetilde{\mathcal{D}}, \mathcal{T})}{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \mathcal{D}, \mathcal{T})} \cdot \frac{p(\widetilde{\mathcal{D}})}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \,|\, \widetilde{\mathcal{D}})}{q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})}\right\}$$

  which depends on:

  - The marginal likelihood ratio;
  - The prior ratio;
  - The proposal ratio;

$\Longrightarrow$ The only difference from the case with no experimental data is in the **marginal likelihood**!

Acceptance/Rejection step

- The **proposal ratio**

$$\frac{q(\mathcal{D} \,|\, \widetilde{\mathcal{D}})}{q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})} = \frac{|\mathcal{O}_{\mathcal{D}}|}{|\mathcal{O}_{\widetilde{\mathcal{D}}}|}$$

  requires the enumeration of all operators that can be applied to $\mathcal{D}$ and lead to a valid graph (i.e. a DAG).
- It is usually computationally expensive, but for p large it can be approximated to 1;

Acceptance/Rejection step

- As we are using local moves and thanks to the decomposability of the marginal likelihood, the **marginal likelihood** ratio simplifies to the components which are affected by the local move.
- If, for instance, $\tilde{\mathcal{D}}$ differs from $\mathcal{D}$ for the addition of an edge pointing towards node t, we have:

$$\begin{aligned}
\frac{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \mid \widetilde{\mathcal{D}}, \mathcal{T})}{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \mid \mathcal{D}, \mathcal{T})} &= \frac{m(\boldsymbol{X}_{.t} \mid \boldsymbol{X}^{(\mathcal{A}(t))}_{.\mathrm{pa}_t(\widetilde{\mathcal{D}})}, \widetilde{\mathcal{D}}, \mathcal{T})}{m(\boldsymbol{X}_{.t} \mid \boldsymbol{X}^{(\mathcal{A}(t))}_{.\mathrm{pa}_t(\mathcal{D})}, \mathcal{D}, \mathcal{T})} \cdot \prod_{k \notin \mathcal{A}(t)} \frac{m(\boldsymbol{X}^{(k)}_{.t})}{m(\boldsymbol{X}^{(k)}_{.t})}, \\
&= \frac{m(\boldsymbol{X}_{.t} \mid \boldsymbol{X}^{(\mathcal{A}(t))}_{.\mathrm{pa}_t(\widetilde{\mathcal{D}})}, \widetilde{\mathcal{D}}, \mathcal{T})}{m(\boldsymbol{X}_{.t} \mid \boldsymbol{X}^{(\mathcal{A}(t))}_{.\mathrm{pa}_t(\mathcal{D})}, \mathcal{D}, \mathcal{T})}
\end{aligned}$$

Posterior inference

- Output of the algorithm is a collection of DAGs $\left\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(S)}\right\}$
- We can provide an estimate of the posterior probability of $\mathcal{D} \in \mathcal{S}_q$ as

$$\boxed{\widehat{p}(\mathcal{D} \mid \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\left\{\mathcal{D}^{(s)} = \mathcal{D}\right\}}$$

  i.e. the proportion of DAGs, visited by the MCMC, equal to $\mathcal{D}$
- Other summaries:
  - Estimate of the (marginal) posterior probability of edge inclusion for each $u \to v$

$$\boxed{\widehat{p}(u \to v \mid \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\left\{u \to v \in \mathcal{D}^{(s)}\right\}}$$

    computed as the proportion of DAGs, visited by the MCMC, containing $u \to v$;

Posterior inference

- DAG point estimates can be obtained by:
  - including those edges whose posterior probability is higher than some threshold, e.g. 0.5 (Median Probability DAG Model, MPM)

$$\widehat{\boldsymbol{S}}_{u,v} = \begin{cases} 1 & \text{if } \widehat{p}(u \to v \,|\, \boldsymbol{X}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

  - selecting the DAG having the highest posterior probability (Maximum A Posteriori DAG, MAP)

$$\widehat{\mathcal{D}}_{MAP} = \underset{\mathcal{D}}{\operatorname{argmax}}\, \widehat{p}(\mathcal{D} \,|\, \boldsymbol{X})$$

Bayesian methods and consistency

- We said that GIES can be inconsistent despite using a consistent score: the local moves it uses may lead to get stuck in local modes;
- In the Bayesian setting, we are not optimizing the score, but using an MCMC scheme that samples from the posterior distribution,
- The only requirements that this chain has to satisfy is that it is **reversible**, **aperiodic** and **irreducible**:
  - Reversibility and aperiodicity follow from the properties of MH;
  - To prove irreducibility, it is sufficient to show that there is a positive probability of moving from each state (DAG) to any other state in a finite number of steps $\rightarrow$ satisfied by our scheme!
- **However**, this does not mean that a **finite** MCMC will not get stuck in regions of the posterior distribution.

$\Longrightarrow$ One can define MCMC schemes on the space of I-Essential graphs! See, for instance, Castelletti & Consonni (2019)

Software

- Unfortunately, no ready-made implementation of the method, we have to write everything from scratch!
- Let's move to `R`!

Bayesian causal discovery from experimental data with unknown intervention targets

Setting

- Suppose we observe a set of K data matrices $\{\boldsymbol{X}^{(k)}\}_{k=1}^{K}$, where
  - $\boldsymbol{X}^{(1)}$ contains i.i.d. samples from the observational distribution of X;
  - $\boldsymbol{X}^{(2)}, \dots, \boldsymbol{X}^{(K)}$ contain samples from the post-intervention distributions of X given a hard intervention on the target nodes $\boldsymbol{T}^{(k)}$.

  and let $\mathcal{T} = \{\boldsymbol{T}^{(k)}\}_{k=1}^{K}$ be the multi-set of intervention targets;
- **Goal:** Derive the posterior distribution

$$\boxed{p(\mathcal{D}, \mathcal{T} \,|\, \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}) \propto m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \mathcal{D}, \mathcal{T})\, p(\mathcal{D}, \mathcal{T})}$$

  i.e., the posterior distribution over DAGs and possible multi-set of targets $\mathcal{T}$ given data from different experimental settings;
- **Again**, this will be a **Bayesian Model Selection** (BMS) problem!

Bayesian causal discovery from unknown targets: How

- **Again**, to do causal discovery via BMS, we need two steps:
  (i) Specify a Bayesian model that will define the posterior distribution. This consists of:
    - $p(\{\boldsymbol{X}\}_{k=1}^{K} \mid \Theta_{(\mathcal{D},\mathcal{T})}, \mathcal{D}, \mathcal{T})$: the statistical model;
    - $p(\Theta_{(\mathcal{D},\mathcal{T})} \mid \mathcal{D}, \mathcal{T})$: the parameter prior;
    - $p(\mathcal{D}, \mathcal{T})$: the model prior;

  (ii) Approximated the posterior distribution via **sampling methods**.
- The only difference is that our model is now defined by both the DAG $\mathcal{D}$ and the multiset of intervention targets $\mathcal{T}$;
- We focus again on the **Gaussian** setting, following Castelletti & Peluso (2023);

Bayesian Model Specification

- Both the **statistical model** and the **parameter prior** are specified conditionally on the multi-set $\mathcal{T}$ of intervention targets;

$\Longrightarrow$ They are exactly the same objects as in the previous case, where the targets were known;

$\Longrightarrow$ We can specify them exactly as before, and inherit all their nice properties;

$\Longrightarrow$ The only thing that remains to be specified is the prior over $\mathcal{T}$!

Prior over targets

- For convenience, we represent each set of intervention targets $\boldsymbol{T}^{(k)}$ as a binary vector $\boldsymbol{I}^{(k)} := \left(\boldsymbol{I}^{(k)}(1), \ldots, \boldsymbol{I}^{(k)}(p)\right)^T$ such that $\boldsymbol{I}^{(k)}(j) = 1$ if and only if $j \in \boldsymbol{T}^{(k)}$ and 0 otherwise;
- As $\boldsymbol{I}^{(k)}$ are binary quantities, we can specify independent Bernoulli priors over each of them

$$\boldsymbol{I}^{(k)}(j) \sim \text{Ber}(\eta) \qquad j \in [p]$$

where $\eta \in (0, 1)$ is the prior inclusion probability;
- The induced prior on $\mathcal{T}$ is:

$$\boxed{p(\mathcal{T}) = \prod_{k=2}^{K} \prod_{j=1}^{q} \eta^{\boldsymbol{I}^{(k)}(j)} (1-\eta)^{1-\boldsymbol{I}^{(k)}(j)}}$$

where $k \in \{2, \ldots, K\}$ as $k = 1$ is the observational setting;

Sampling from the posterior

- As the **marginal likelihood** is available in closed-form, it is still possible to use a Metropolis-Hastings scheme;
- However, that would require defining local moves that navigate the space of possible DAGs and targets simultaneously: not practical and possibly not efficient;
- **Solution:** Gibbs sampling scheme where, after initialising at an arbitrary pair $(\mathcal{D}^{(0)}, \mathcal{T}^{(0)})$, for each $s \in [S]$:
  - Sample $\mathcal{D}^{(s)}$ from $p(\mathcal{D} \mid \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}, \mathcal{T}^{(s-1)})$;
  - Sample $\mathcal{T}^{(s)}$ from $p(\mathcal{T} \mid \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}, \mathcal{D}^{(s)})$;

Gibbs sampling

- By sampling from its **full conditional distributions** $p(\mathcal{D} \,|\, \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}, \mathcal{T})$, $p(\mathcal{T} \,|\, \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}, \mathcal{D})$ the Gibbs sampler approximates the posterior distribution $p(\mathcal{D}, \mathcal{T} \,|\, \{\boldsymbol{X}^{(k)}\}_{k=1}^{K})$ at convergence;
- **Problem:** it is not possible to sample from our full conditional distributions directly
- **Solution:** Just use another layer of Metropolis-Hastings!

Full conditional of $\mathcal{D}$

- To sample from $p(\mathcal{D}\,|\,\{\boldsymbol{X}^{(k)}\}_{k=1}^{K}, \mathcal{T})$, we use
- Given a current DAG $\mathcal{D}$, a new DAG $\widetilde{\mathcal{D}}$ drawn from the proposal $q(\widetilde{\mathcal{D}}\,|\,\mathcal{D})$ is accepted with probability

$$\alpha_{\widetilde{\mathcal{T}},\mathcal{T}} = \min\left\{1; \frac{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K}\,|\,\widetilde{\mathcal{D}}, \mathcal{T}^{(s-1)})}{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K}\,|\,\mathcal{D}, \mathcal{T}^{(s-1)})} \cdot \frac{p(\widetilde{\mathcal{D}})}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D}\,|\,\widetilde{\mathcal{D}})}{q(\widetilde{\mathcal{D}}\,|\,\mathcal{D})}\right\}$$

Full conditional of $\mathcal{T}$

- To sample from $p(\mathcal{T} \,|\, \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}, \mathcal{D})$, again MH;
- Given a current multi-set $\mathcal{T}$, a new multi-set $\widetilde{\mathcal{T}}$ drawn from the proposal $q(\widetilde{\mathcal{T}} \,|\, \mathcal{T})$ is accepted with probability

$$\alpha_{\widetilde{\mathcal{T}},\mathcal{T}} = \min\left\{1; \frac{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \widetilde{\mathcal{T}}, \mathcal{D}^{(s)})}{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \mathcal{T}, \mathcal{D}^{(s)})} \cdot \frac{p(\widetilde{\mathcal{T}})}{p(\mathcal{T})} \cdot \frac{q(\mathcal{T} \,|\, \widetilde{\mathcal{T}})}{q(\widetilde{\mathcal{T}} \,|\, \mathcal{T})}\right\}$$

- The Bayes factor and the prior ratio are defined by the model. We need to devise a **proposal scheme**, from which the **proposal ratio** can be computed;

Full conditional of $\mathcal{T}$ - Proposal

- **Proposal scheme:** For each $k \in [K]$, sample $t \in [p]$ uniformly at random and set:
  - $\tilde{\boldsymbol{T}}^{(k)} = \{t \cup (\boldsymbol{T}^{(k)})^{(s-1)}\}$ if $t \notin (\boldsymbol{T}^{(k)})^{(s-1)}$;
  - $\tilde{\boldsymbol{T}}^{(k)} = \{(\boldsymbol{T}^{(k)})^{(s-1)} \backslash t\}$ otherwise;
- The proposal ratio induced by this scheme is always 1!

Full conditional of $\mathcal{T}$ - Marginal likelihood

- Suppose now that, for setting $\tilde{k}$ the node t was chosen uniformly at random and that $t \in \tilde{\boldsymbol{T}}^{(k)}$ and $t \notin \boldsymbol{T}^{(k)}$;
- Then, $\tilde{\mathcal{A}}(t) \neq \mathcal{A}(t)$ and the marginal likelihood ratio becomes

$$\frac{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \widetilde{\mathcal{T}}, \mathcal{D}^{(s)})}{m(\{\boldsymbol{X}^{(k)}\}_{k=1}^{K} \,|\, \mathcal{T}, \mathcal{D}^{(s)})} = \frac{m(\boldsymbol{X}_{.t} \,|\, \boldsymbol{X}_{.\mathrm{pa}_t(\mathcal{D})}^{(\tilde{\mathcal{A}}(t))}, \widetilde{\mathcal{T}}, \mathcal{D}^{(s)}) \prod_{k \notin \tilde{\mathcal{A}}(t)} m(\boldsymbol{X}_{.t}^{(k)})}{m(\boldsymbol{X}_{.t} \,|\, \boldsymbol{X}_{.\mathrm{pa}_t(\mathcal{D})}^{(\mathcal{A}(t))}, \mathcal{T}, \mathcal{D}^{(s)}) \prod_{k \notin \mathcal{A}(t)} m(\boldsymbol{X}_{.t}^{(k)})},$$

$$= \frac{m(\boldsymbol{X}_{.t} \,|\, \boldsymbol{X}_{.\mathrm{pa}_t(\mathcal{D})}^{(\tilde{\mathcal{A}}(t))}, \widetilde{\mathcal{T}}, \mathcal{D}^{(s)}) \cdot m(\boldsymbol{X}_{.t}^{(\tilde{k})})}{m(\boldsymbol{X}_{.t} \,|\, \boldsymbol{X}_{.\mathrm{pa}_t(\mathcal{D})}^{(\mathcal{A}(t))}, \mathcal{T}, \mathcal{D}^{(s)})}$$

- Again, the **decomposability** of the marginal likelihood allows huge computational savings!

Posterior inference

- Output of our scheme is a collection $(\mathcal{D}^{(s)}, \mathcal{T}^{(s)})$, for $s \in [S]$;
- We can provide an estimate of the posterior probability of $\mathcal{T}$ as

$$\boxed{\widehat{p}(\mathcal{T} \mid \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\left\{\mathcal{T}^{(s)} = \mathcal{T}\right\}}$$

  i.e. the prop. of sampled targets equal to $\mathcal{T}$ in the MCMC output;
- Other summaries:
  - Estimate of the (marginal) posterior probability of target inclusion for each $k \in [K]$ and $j \in [p]$

$$\boxed{\widehat{p}(j \in \boldsymbol{T}^{(k)} \mid \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\left\{j \in (\boldsymbol{T}^{(k)})^{(s)}\right\}}$$

    computed as the prop. of targets containing j in the MCMC output;

Software

- Unfortunately, again no ready-made implementation of the method, we have to write everything from scratch!
- Let's move to `R`!

Thank you!

References:

- Ben-David, E., Li, T., Massam, H., & Rajaratnam, B. (2011). *High dimensional Bayesian inference for Gaussian directed acyclic graph models*. arXiv preprint arXiv:1109.4371.
- Castelletti, F., & Consonni, G. (2019). *Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways* Journal of the Royal Statistical Society Series A, 183(4), 1727-1745..
- Castelletti, F., & Peluso, S. (2023). *Network structure learning under uncertain interventions*. Journal of the American Statistical Association, 118(543), 2117-2128.
- Castelletti, F., & Peluso, S. (2024). Bayesian learning of network structures from interventional experimental data. Biometrika, 111(1), 195-214.
- Geiger, D., & Heckerman, D. (2002). *Parameter priors for directed acyclic graphical models and the characterization of several probability distributions*. The Annals of Statistics, 30(5), 1412-1440.