21 Causal Structure Discovery I

In the last couple of lectures, we discussed learning the structure of undirected graphical models from data. We will now focus on learning causal relationships, i.e., structure in directed graphical models. In this chapter, we will consider the case where we observe all the nodes and we have observational data. In the next lecture, we will discuss cases where we also have interventional data (obtained from do-operations), as well as the case where only some part of the nodes are observed.

21.1 DAGs and Markov equivalence classes (MECs)

Let us briefly recall what we learned from past lectures on directed graphical models. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG. As usual, we have $\mathcal{V} = \{1, \dots, N\}$ and $\mathcal{E}$ denote the set of directed edges. Then, the fact that a probability distribution $p(x_1, \dots, x_N)$ factorizes according to the DAG $\mathcal{G}$, i.e.,

$$p(x_1, \dots, x_N) = \prod_{i=1}^{N} p(x_i \mid x_{\pi_i}),$$

is equivalent to the (directed) global Markov property with respect to $\mathcal{G}$, i.e., $x_{\mathcal{A}} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}}$ holds for any $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathcal{V}$ such that $\mathcal{A}$ and $\mathcal{B}$ are d-separated by $\mathcal{C}$ in $\mathcal{G}$.

In order to learn structure from data, we can use the data to test conditional independence relations, and from those, we can infer the structure of the DAG. However, we have seen in one of the problem sets that there can be *multiple* DAGs that imply exactly the same set of conditional independence relations. These graphs are said to be *Markov equivalent*, and they are in the same *Markov equivalence class* (MEC). Thus, when learning structure of a DAG from data, in general the best we can hope for is to find the right MEC, because in general we cannot identify a unique DAG for a given set of conditional independence relations. On the other hand, if there are additional assumptions on the distribution, we can sometimes identify the unique DAG in the MEC; we will see such examples shortly. Also, one can use *interventional* data to infer causal structures of the DAG that are not identified using observational data; we discuss this in the next lecture.

In one of the problem sets, we proved that two DAGs are in the same MEC if and only if (1) they have the same skeleton, and (2) they have the same set of immoralities. Natural questions then arise: given a skeleton and a set of immoralities, how many Markov equivalent DAGs are there in the MEC? How big is a given MEC? How many MECs are there for N variables? For an illustration, Fig. 1 enumerates all the possible MECs for DAGs with $N = 3$. One can note that there are large MECs (e.g., the complete graph, which has 6 Markov equivalent DAGs) and also small MECs of

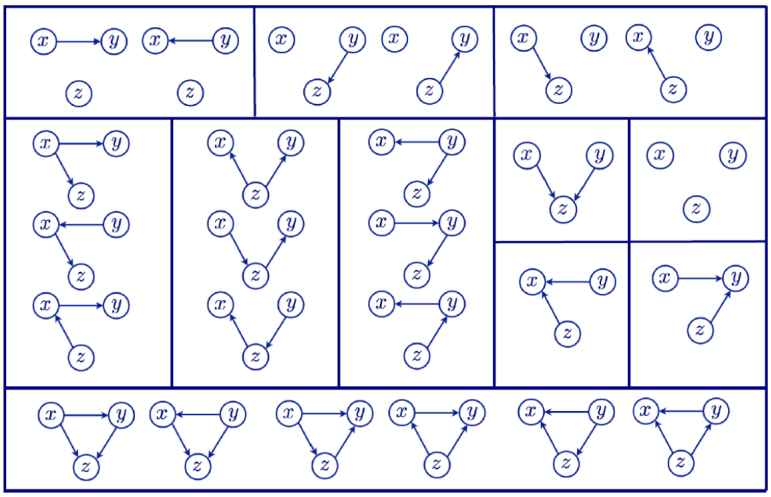

Figure 1: MECs in three-variable DAGs.

size 1 (only one DAG in the MEC). Even though recent papers (Gillispie & Perlman, 2001; Radhakrishnan, Solus, Uhler, 2017; Katz-Rogozhnikov, Shanmugam, Squires, Uhler, 2019) studied the number, size distribution, etc. of MECs for N up to 12 or 13, there are still a lot of open questions left to be answered, especially for larger N.

21.2 Identifying DAGs: linear model and additive noise

Now, back to the question of structure learning. We discussed above that for general distributions and observational data, one can only identify the MEC. Then, can we identify more than the MEC with additional assumptions on the distribution?

Let us first consider the case where $x_1, \ldots, x_N$ are jointly Gaussian:

$$p(x_1, \ldots, x_N) \sim \mathtt{N}(0, \mathbf{\Sigma}).$$

We assume that the mean of the Gaussian is zero, because it doesn't affect conditional independences of the variables. We know that any Gaussian DAG model is linear, i.e., we can write each variable x_j as $x_j = \sum_{i \in \pi_j} a_{ij} x_i + \epsilon_j$, where ϵ_j is an additive Gaussian "noise" independent of x_{π_j}. The following simple example highlights that in the jointly Gaussian setting, identifying the DAG out of the MEC is *impossible*. This is in contrast to the usual "Gaussians are nice" stories that we have seen in class!

Example 1. Consider a DAG $X \to Y$, where

$$Y = \phi X + N, \quad X \perp\!\!\!\perp N,$$

where X and N are mean-zero Gaussian random variables. It can be checked that

$$X = \tilde{\phi} Y + \tilde{N}, \quad Y \perp\!\!\!\perp \tilde{N},$$

with $\tilde{\phi} = \frac{\phi \operatorname{var}(X)}{\phi^2 \operatorname{var}(X) + \operatorname{var}(N)}$ and $\tilde{N} = X - \tilde{\phi} Y$, implying that the distribution can be represented with a DAG $Y \to X$.

This simple example already shows that even if we know that the variables are jointly Gaussian, we cannot tell which of the two DAGs is the right one; both DAGs are equally valid. In fact, rather surprisingly, the following proposition tells us that the Gaussian case is the *only* "bad" case with respect to identifiability. If we consider X and N that are non-Gaussian, then the structure (i.e., edge direction) of the DAG becomes completely identifiable.

Proposition 1. *Let X and Y be two random variables, for which*

$$Y = \phi X + N, \quad X \perp\!\!\!\perp N, \quad \phi \neq 0$$

holds. Then, we can reverse the process, i.e., there exists $\psi \in \mathbb{R}$ and a noise $\tilde{N}$ such that

$$X = \psi Y + \tilde{N}, \quad Y \perp\!\!\!\perp \tilde{N}$$

if and only if X and N are Gaussian.

We will not present the proof here, but this is a consequence of the following theorem that characterizes the Gaussian distribution, by Skitovič and Darmois (Skitovič, 1954; 1962; Darmois, 1953).

Theorem 1 (Darmois-Skitovič)**.** *Let $x_1, \ldots, x_d$ be independent, non-degenerate random variables. If there are non-vanishing coefficients $a_1, \ldots, a_d$ and $b_1, \ldots, b_d$ (i.e., $a_i \neq 0$ and $b_i \neq 0$ for all i) such that the two linear combinations*

$$y_1 = a_1 x_1 + \cdots + a_d x_d, \quad y_2 = b_1 x_1 + \cdots + b_d x_d,$$

are independent, then each x_i is Gaussian distributed.

Proposition 1 extends to more than two variables. Shimizu et al. (2006) proved the following theorem, which shows that in any linear model with additive noise, the underlying DAG is identifiable if and only if the noise is non-Gaussian.

Theorem 2 (Shimizu et al., 2006)**.** *For some DAG $\mathcal{G}$, let $x_j = \sum_{i \in \pi_j} a_{ij} x_i + \epsilon_j$, $i = 1, \ldots, N$ where all ϵ_i's are jointly independent, and $a_{ij} \neq 0$ for all $i \in \pi_j$. Then, the DAG $\mathcal{G}$ is identifiable from the joint distribution if and only if the ϵ_i's are non-Gaussian distributed.*

The authors call this model a linear non-Gaussian acyclic model (LiNGAM) and provide a practical method based on independent component analysis (ICA) that can be applied to a finite amount of data. Later, an improved version of this method has been proposed in (Shimizu et al., 2011).

Throughout this class, we have seen many benign properties of Gaussian distributions. Most of the time, these good properties come from the symmetries (e.g.,

rotational invariance) that the Gaussian distribution has. However, when it comes to identifying DAG structures, these symmetries actually hurt. In order to identify the true DAG from data, we need something to "break" the symmetries so that we can tell different DAGs in the same MEC apart. One example is when we have non-Gaussian noise in the linear model, as seen above. We will now see another way of breaking symmetries, which is when the model is *nonlinear*.

21.3 Identifying DAGs: nonlinear model and additive noise

So far, we have looked at linear models with additive noise, i.e., $\mathsf{x}_j = \sum_{i \in \pi_j} a_{ij} \mathsf{x}_i + \epsilon_j$. We will now consider the case where the model is nonlinear: $\mathsf{x}_j = f_j(\mathsf{x}_{\pi_j}) + \epsilon_j$ for some nonlinear function $f_j : \mathbb{R}^{|\pi_j|} \to \mathbb{R}$. The following result shows that nonlinearity helps in breaking the symmetry, so that even with Gaussian noise ϵ_j we can identify the DAG.

Theorem 3 (Peters et al., 2014)**.** *For some DAG $\mathcal{G}$, let $\mathsf{x}_j = f(\mathsf{x}_{\pi_j}) + \epsilon_j$, $i = 1, \ldots, N$ where all Gaussian noise terms $\epsilon_i \sim \mathtt{N}(0, \sigma_j^2)$ are jointly independent, and three-times differentiable functions f_j are not linear in any component: denoting the parents x_{π_j} of x_j by $\mathsf{x}_{k_1}, \ldots, \mathsf{x}_{k_\ell}$, the function $f_j(x_{k_1}, \ldots, x_{k_{a-1}}, \cdot, x_{k_{a+1}}, \ldots, x_{k_\ell})$ is assumed to be nonlinear for all $a \in \{1, \ldots, \ell\}$ and some $(x_{k_1}, \ldots, x_{k_{a-1}}, x_{k_{a+1}}, \ldots, x_{k_\ell}) \in \mathbb{R}^{\ell-1}$. Then, we can identify the DAG $\mathcal{G}$ from the joint distribution.*

It is also shown in the same paper that a similar result holds for non-Gaussian noise. This theorem tells us that for nonlinear models, additive Gaussian noise still allows identifiability; thus, *linear models with Gaussian noise* are basically the only "bad" cases when it comes to identifiability of DAGs.

Example 2. Consider a simple nonlinear example where

$$\mathsf{Y} = \mathsf{X}^3 + \mathsf{N}_\mathsf{Y}, \quad \mathsf{X} \perp\!\!\!\perp \mathsf{N}_\mathsf{Y}, \quad \mathsf{N}_\mathsf{Y} \sim \mathtt{N}(0, 1).$$

The top-left graph of Fig. 2 shows how different samples of (X, Y) look like. If you plot the residual $\mathsf{Y} - \mathsf{X}^3$ (which is equal to N_Y) versus X, we get the bottom-left plot and we can see that X and N_Y are independent. The top-right and bottom-right plots show what happens if you "flip" X and Y. Namely, we run a nonlinear regression on the data and fit X as a function of Y, say $\mathsf{X} = g(\mathsf{Y}) + \mathsf{N}_\mathsf{X}$. The bottom-right panel plots the residual $\mathsf{X} - g(\mathsf{Y})$ (i.e., the "noise" N_X) versus Y, and we can clearly see that N_X and Y are not independent. From this, we can see that the correct DAG should be $\mathsf{X} \to \mathsf{Y}$; nonlinearity of the model breaks the symmetry.

Even though we can identify the DAGs from observational data when we have nonlinear models or non-Gaussian noise, these methods typically have a higher computational price which grows rapidly with the size of the graph. Often this extra computational cost for DAG identification is prohibitively large, so that we have to resort to methods that only allow identifying the MEC. We will now look at methods for learning MECs.

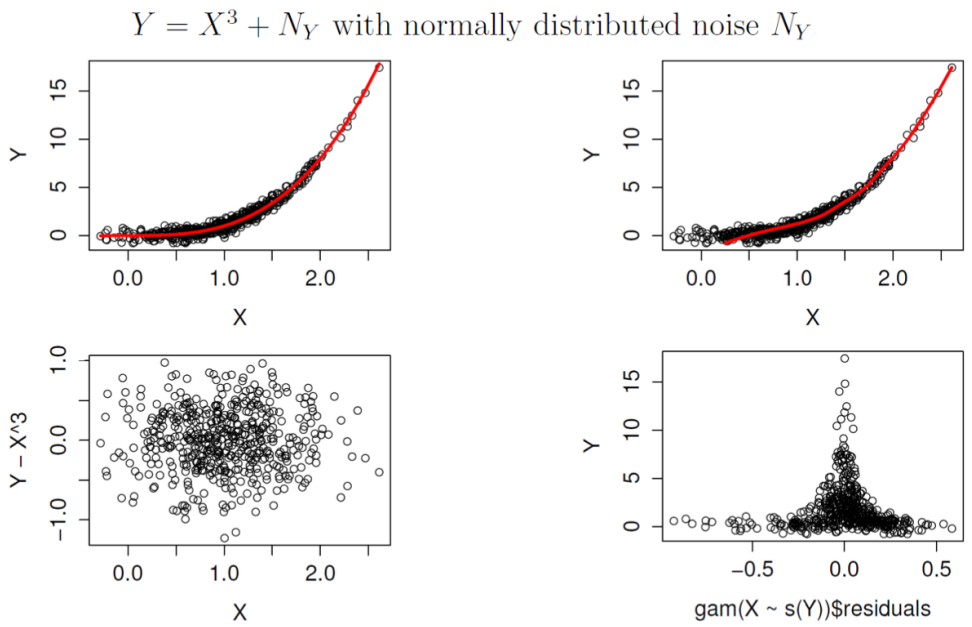

Figure 2: Identifiability of DAG in nonlinear models (Example 2).

21.4 Identifying MECs: constraint-based approaches

We now switch gears and discuss two different approaches for learning MECs from data. The first set of approaches is called *constraint-based* approaches. Here, constraints refer to conditional independence (CI) relations; we infer CI relations from data and determine the MEC that satisfies as many of the CI relations as possible.

As you will see in the problem set, one such approach for identifying MECs proceeds by removing an edge (i, j) from the graph whenever there is a set $\mathcal{S} \in \mathcal{V} \setminus \{i, j\}$ satisfying $x_i \perp\!\!\!\perp x_j \mid x_{\mathcal{S}}$. This is how you learn the skeleton of the DAG, and the algorithm further learns immoralities to find out the MEC. For such approaches to work, we actually need a strong assumption on the distribution. Recall from the global Markov property that whenever there is no edge (i, j) between nodes i and j, there must exist a set $\mathcal{S} \subset \mathcal{V}$ such that $x_i \perp\!\!\!\perp x_j \mid x_{\mathcal{S}}$. For structure learning, the algorithm assumes that the converse also holds: if you have $x_{\mathcal{A}} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{S}}$, then $\mathcal{A}$ and $\mathcal{B}$ are d-separated by $\mathcal{S}$. This assumption is called the *faithfulness assumption*.

If there is an infinite amount of data, then the fathfulness assumption is not too restrictive, because we can correctly identify CI relations (unless the parameters exactly cancel out) and there is a unique MEC that satisfies the exact set of conditional independence relations coming from the DAG. In practice, there is only finite amount of data available, so as we discussed a couple of lectures ago, we have to perform hypothesis testing based on the empirical mutual information between random variables. Basically, we calculate the empirical mutual information, and if it is below

some threshold we conclude that the CI relation is likely to hold and remove the corresponding edge. Of course, mistakes are expected to happen in these tests leading to the removal of edges that should not be removed. These makes this type of approach sensitive to mistakes in individual CI tests which can accumulate over the course of the algorithm.

21.5 Identifying MECs: score-based approaches

Another class of approaches are score-based methods, which were already introduced in the previous lectures. In particular, we discussed using the log-likelihood of the data over the edges of the DAG as a score for each DAG. Recall that the likelihood score of a DAG given data $\mathcal{D}$ can be written as

$$\hat{\ell}(\mathcal{G}) = n \sum_{i=1}^{N} \left[\hat{I}(\mathsf{x}_i; \mathsf{x}_{\pi_i}) - \hat{H}(\mathsf{x}_i) \right] = n \sum_{i=1}^{N} \hat{I}(\mathsf{x}_i; \mathsf{x}_{\pi_i}) - n \sum_{i=1}^{N} \hat{H}(\mathsf{x}_i),$$

where n is the number of data points in $\mathcal{D}$. Note that maximizing the likelihood score over all possible DAGs in the finite sample setting results in the complete DAG. This is the case since adding an edge can only increase the mutual information term. Always outputting the complete DAG is certainly not desirable, because our goal is to learn the MEC from data. The complete DAG carries no information since it is equivalent to all other complete DAGs. This motivates penalizing the size of a DAG.

There are many ways one can penalize the score to promote sparsity. We here discuss a particular one called the *Bayesian Information Criterion* (*BIC*) score. As seen in a previous installment of the notes, BIC has a very nice motivation in terms of Bayesian statistics. The BIC score is defined as

$$\ell_{\text{BIC}}(\mathcal{G}) = \hat{\ell}(\mathcal{G}) - \frac{\ln n}{2} \dim \mathcal{G}, \tag{1}$$

where $\dim \mathcal{G}$ is the number of independent parameters in the model corresponding to $\mathcal{G}$, which is roughly the number of edges $|\mathcal{E}|$. Note that $\hat{\ell}(\mathcal{G}; \mathcal{D})$ scales linearly in n, which is faster than $\ln n$. This means that when $n \to \infty$, then we are okay to use the likelihood score, but if n is small, we want to penalize graphs that have too many parameters.

There are some nice and useful properties of BIC scores:

- **Score equivalence**: For any graphs $\mathcal{G}$ and $\mathcal{G}'$ in the same MEC, we have

$$\ell_{\text{BIC}}(\mathcal{G}) = \ell_{\text{BIC}}(\mathcal{G}').$$

- **Consistency**: Let $\mathcal{G}^*$ be a perfect map of the true distribution p^*. Then, as $n \to \infty$, with probability 1, the perfect map $\mathcal{G}^*$ maximizes the score ℓ_{BIC}, and every DAG $\mathcal{G}$ that is not in the same MEC as $\mathcal{G}^*$ satisfies

$$\ell_{\text{BIC}}(\mathcal{G}) < \ell_{\text{BIC}}(\mathcal{G}^*).$$

- **Decomposability**: The BIC score can be decomposed into a sum of score terms that only depend on a random variable and its parents:

$$\ell_{\mathrm{BIC}}(\mathcal{G}) = \sum\nolimits_{i=1}^{N} \ell(\mathsf{x}_i, \mathsf{x}_{\pi_i}).$$

  This last property is important from a computational perspective, because when searching over the DAG space by adding/removing edges, we can update the score by only recomputing the small portion of a DAG that has been modified.

Now, how do we maximize the BIC score over all DAGs? Since the search space of all DAGs grows super-exponentially in the number of variables (e.g., Chickering, 2002), greedy search algorithms are applied to solve the problem. At each step there is a candidate graph and a set of neighboring graphs. For all these neighbors one computes the score and considers the best-scoring graph as the new candidate. If none of the neighbors obtains a better score, the search procedure terminates (not knowing whether one obtained only a local optimum). Clearly, one therefore has to define a neighborhood relation. Starting from a graph $\mathcal{G}$, we may define all graphs as neighbors from $\mathcal{G}$ that can be obtained by removing, adding or reversing one edge.

In the linear Gaussian case, for example, one cannot distinguish between Markov equivalent graphs. It turns out that in those cases it is beneficial to change the search space to Markov equivalence classes instead of DAGs. The greedy equivalence search (GES), proposed by Meek and analyzed by Chickering, starts with the empty graph and consists of two phases. In the first phase, edges are added until a local maximum is reached; in the second phase, edges are removed until a local maximum is reached, which is then given as an output of the algorithm. Quite remarkably, as sample size $n \to \infty$, this GES algorithm is known to find the correct MEC under the faithfulness condition.

In the next lecture we discuss how to make use of interventional data in the causal structure discovery process.