Causal Inference with Graphical Models

Jose M. Peña
STIMA, IDA, LiU

Lecture 1: Causal Models and Learning Algorithms

Contents

- Causal Models
- IC Algorithm
- Projections
- IC* Algorithm
- Restricted Causal Models
- RESIT Algorithm
- Score Based Algorithms

Literature

- Main sources
  - Pearl, J. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press, 2009. Chapters 1 and 2.
  - Peters, J., Janzing, D. and Schölkopf, B. *Elements of Causal Inference.* MIT Press, 2017. Chapters 3, 4, 6 and 7.
- Additional sources
  - Verma, T. S. Graphical Aspects of Causal Models. Technical Report R-191, UCLA, 1993.
  - Pearl, J. *Causality: Models, Reasoning, and Inference* (1st ed.). Cambridge University Press, 2000. Chapters 1 and 2.

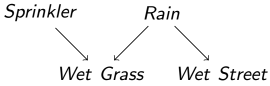
Causal Models

- A causal structure over a set of variables V is a DAG over V.
- A causal model consists of a causal structure, a set of functions $x_i = f_i(pa_i, u_i)$ for each $X_i \in V$, and a distribution $p(u_i)$ for each U_i.
- The functions are also called structural equations, which is **different** from algebraic equations since the equality sign should be read as an assignment or determination, i.e. it is asymmetric.
- For now, the error, noise or disturbance terms U_i are assumed to be **independent** one of another. They may be seen as representing unmodeled or unobserved causes.
- Note that $f_i(pa_i, u_i)$ and $p(u_i)$ together define a conditional distribution $p(x_i|pa_i)$. Then, a causal model defines a distribution over V:

$$p(v) = \prod_i p(x_i|pa_i).$$

- A causal model can be obtained from knowledge of the physics behind the phenomenon being modeled, from interventional experiments such as randomized control trials, or from passive observations.

Causal Models

DAG	Parameter values for the conditional probability distributions
	$p(s) = (0.3, 0.7) = (\theta_{s_0}, \theta_{s_1})$ $p(r) = (0.5, 0.5) = (\theta_{r_0}, \theta_{r_1})$ $p(wg\|r_0, s_0) = (0.1, 0.9) = (\theta_{wg_0\|r_0,s_0}, \theta_{wg_1\|r_0,s_0})$ $p(wg\|r_0, s_1) = (0.7, 0.3) = (\theta_{wg_0\|r_0,s_1}, \theta_{wg_1\|r_0,s_1})$ $p(wg\|r_1, s_0) = (0.8, 0.2) = (\theta_{wg_0\|r_1,s_0}, \theta_{wg_1\|r_1,s_0})$ $p(wg\|r_1, s_1) = (0.9, 0.1) = (\theta_{wg_0\|r_1,s_1}, \theta_{wg_1\|r_1,s_1})$ $p(ws\|r_0) = (0.1, 0.9) = (\theta_{ws_0\|r_0}, \theta_{ws_1\|r_0})$ $p(ws\|r_1) = (0.7, 0.3) = (\theta_{ws_0\|r_1}, \theta_{ws_1\|r_1})$ $p(s, r, wg, ws) = p(s)p(r)p(wg\|s, r)p(ws\|r)$

IC Algorithm

- A distribution p is stable or **faithful** or isomorphic wrt a DAG D when $X \perp_p Y | Z$ iff $X \perp_D Y | Z$.
- In other words, the independences in p are **structural** and not formed by incidental parameter equalities. The unstable distributions have measure zero when the parameters are chosen at random.
- A pattern is a mixed graph that represents an equivalence class of DAGs:
  - It contains the edge $A \to B$ if the edge is in every member of the class, and
  - the edge $A - B$ if $A \to B$ is in some members and $A \leftarrow B$ in some others.

Inductive causation (IC) algorithm

Input: A distribution p over V that is stable wrt some DAG D

Output: The pattern G corresponding to the equivalence class of D

```
Let $G$ be a complete undirected graph
For each pair of nodes $A, B \in V$
    If $A \perp_p B | S_{AB}$ for some $S_{AB} \subseteq V$, then delete the edge $A - B$ from $G$
For each pair nodes $A, B \in V$ st $A - B$ is not in $G$
    If $A \multimap C \multimapinv B$ is in $G$ and $C \notin S_{AB}$, then add the orientations $A \to C \leftarrow B$ to $G$
Orient as many edges in $G$ as possible without creating inmoralities or directed cycles
```

- The algorithm's steps can be systematized and optimized.

IC Algorithm

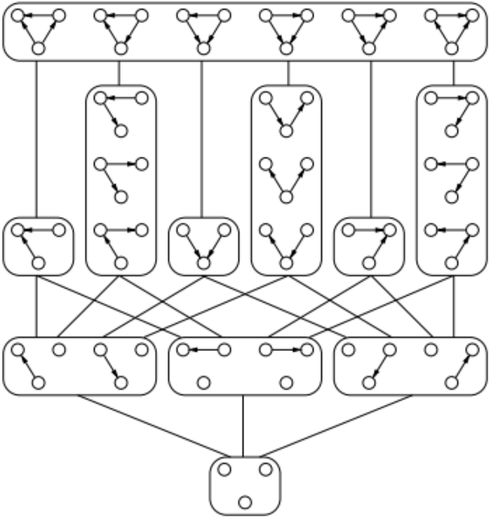


Figure 2: Hasse diagram of the space of Markov equivalence classes of Bayesian network structures over three variables.

Projections

- A latent structure L is a causal structure over $V \cup U$ st V are observable and U are latent. The variables in U are **not necessarily independent** anymore.
- Note that L induces a distribution over V:

$$p(v) = \sum_U \prod_i p(x_i | pa_i).$$

- It may be convenient to work with the projection of L onto V.

Projection algorithm

Input: A latent structure L over $V \cup U$

Output: The projection G of L over V

```
Let $G$ be the empty graph over $V$
For each pair of nodes $A, B \in V$
    If $L$ has a directed path from $A$ to $B$ st every internal node is in $U$,
    then add the edge $A \rightarrow B$ to $G$
    If $L$ has a divergent path between $A$ and $B$ st every internal node is in $U$,
    then add the edge $A \leftrightarrow B$ to $G$
```

- The separation criterion for DAGs can be extended to projections: Simply, redefine the term collider as follows.
  - A node B in a path ρ is a collider when $A \rightarrow B \leftarrow C$ or $A \rightarrow B \leftrightarrow C$ or $A \leftrightarrow B \leftrightarrow C$ is a subpath of ρ.
- Interestingly, L and G represent the **same** independence model over V.

IC* Algorithm

```
IC* algorithm
Input: A distribution $p$ over $V$ that is stable wrt the projection of some latent structure $L$
Output: A marked pattern $G$ of the projection of $L$ over $V$

0. Let $G$ be a complete undirected graph
1. For each pair of nodes $A, B \in V$
     If $A \perp_p B | S_{AB}$ for some $S_{AB} \subseteq V$, then delete the edge $A - B$ from $G$
2. For each pair nodes $A, B \in V$ st $A - B$ is not in $G$
     If $A \circ\!\!-\!\!\circ C \circ\!\!-\!\!\circ B$ is in $G$ and $C \notin S_{AB}$, then add the arrowheads $A \circ\!\!\rightarrow C \leftarrow\!\!\circ B$ to $G$
3. Add as many arrowheads and marks to $G$ as possible according to the following rules:
     3.1. If $G$ has a marked directed path from $A$ to $B$ and $A \circ\!\!- B$,
            then add the arrowhead $A \circ\!\!\rightarrow B$ to $G$
     3.2. If $A \circ\!\!\rightarrow C \multimap B$ is in $G$ and $A \circ\!\!-\!\!\circ B$ is not in $G$,
            then add the arrowhead $C \rightarrow B$ to $G$ and mark the edge with $*$
```

- Every $*$-marked directed edge in G corresponds to a directed path in L, i.e. genuine causation.
- Every unmarked directed edge in G corresponds to a directed or divergent path in L, i.e. potential causation.
- Every bidirected edge in G corresponds to a divergent path in L, i.e. spurious association.
- Undirected edges in G correspond to undetermined relationships.
- The algorithm's steps can be systematized and optimized.
- There exist more sophisticated algorithms that allow even selection bias.

IC* Algorithm

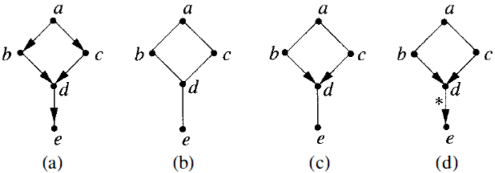


Figure 2.3 Graphs constructed by the IC* algorithm. (a) Underlying structure. (b) After step 1. (c) After step 2. (d) Output of IC*.

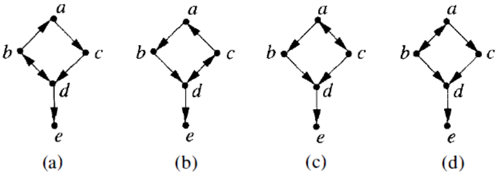


Figure 2.4 Latent structures equivalent to those of Figure 2.3(a).

Restricted Causal Models

- Let $X \sim F_X$, where F_X is a continuous CDF. Then, $Y = F_X(x) \sim U(0,1)$.
- Proof:

$$F_Y(y) = p(Y \le y) = p(F_X(x) \le y) = p(X \le F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

- Useful for sampling random variables (a.k.a. inverse CDF method): Let $Y \sim U(0,1)$ and let F_X be a continuous CDF. Then, $X = F_X^{-1}(y) \sim F_X$.
- Any joint probability distribution $p(x,y)$ admits causal models **in both directions**, i.e.

$$X \to Y : x = f_X(u_X) \text{ and } y = f_Y(x, u_Y) \text{ with } X \perp U_Y.$$

$$Y \to X : y = f_Y(u_Y) \text{ and } x = f_X(y, u_X) \text{ with } Y \perp U_X.$$

- Proof: Let $F_{Y|x}(y) = p(Y \le y | X = x)$ and let $f_Y(x, u_Y) = F_{Y|x}^{-1}(u_Y)$ where $U_Y \sim U(0,1)$ and $X \perp U_Y$.
- So, there is no chance of identifying the true causal model from observations alone unless **further assumptions are made**.

Restricted Causal Models

- Assume that $p(x, y)$ admits the causal model

$$y = \alpha x + u_Y \text{ with } X \perp U_Y$$

where the random variables are continuous. Then,

$$x = \beta y + u_X \text{ with } Y \perp U_X$$

iff X and U_Y are Gaussian.

- In other words, identifiability for **linear non-Gaussian models** is possible.

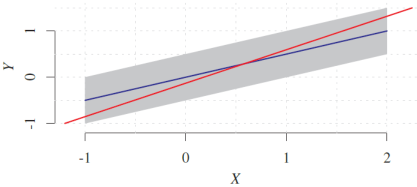


Figure 4.1: Joint density over X and Y for an identifiable example. The blue line is the function corresponding to the forward model $Y := 0.5 \cdot X + N_Y$, with uniformly distributed X and N_Y; the gray area indicates the support of the density of (X, Y). Theorem 4.2 states that there cannot be any valid backward model since the distribution of (X, N_Y) is non-Gaussian. The red line characterized by (b, c) is the least square fit minimizing $\mathbb{E}[X - bY - c]^2$. This is not a valid backward model $X = bY + c + N_X$ since the resulting noise N_X would not be independent of Y (the size of the support of N_X would differ for different values of Y).

Restricted Causal Models

- Assume that $p(x, y)$ admits the causal model

$$y = f_Y(x) + u_Y \text{ with } X \perp U_Y$$

  where the random variables are continuous. Then, $p(x, y)$ does not admit **in general** a model of the same form in the backward direction.
- In other words, identifiability for **non-linear additive models** is possible in general, i.e. for all but some "rare" or "non-generic" or "fine-tuned" cases.
- The precise characterization is rather technical. Exception: When X and U_Y are Gaussian, $p(x, y)$ admits the backward model iff f_Y is linear.
- Assume that $p(x, y)$ admits the causal model

$$y = g_Y(f_Y(x) + u_Y) \text{ with } X \perp U_Y$$

  where the random variables are continuous. Then, $p(x, y)$ does not admit **in general** a model of the same form in the backward direction.
- In other words, identifiability for **post-nonlinear models** is possible in general.

RESIT Algorithm

Regression with subsequent independence test (RESIT) algorithm

Input: A sample from $p(x, y)$

Output: The non-linear additive model $X \to Y$ or $Y \to X$ or nothing

```
Perform a non-linear regression from $Y$ on $X$ to write $y = \hat{f}_Y(x) + \hat{u}_Y$
Perform the hypothesis test $H_0 : X \perp \hat{U}_Y$
Repeat the two steps above exchanging the roles of $X$ and $Y$
If $H_0$ is accepted in one direction and rejected in the other,
    then infer the former as the causal direction
```

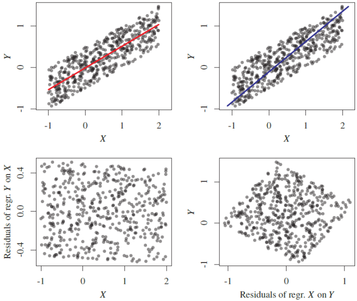


Figure 4.5: We are given a sample from the underlying distribution and perform a linear regression in the directions $X \to Y$ (left) and $Y \to X$ (right). The fitted functions are shown in the top row, the corresponding residuals are shown in the bottom row. Only the direction $X \to Y$ yields independent residuals; see also Figure 4.1.

RESIT Algorithm

Regression with subsequent independence test (RESIT) algorithm

Input: A sample from a distribution $p(x_1, \dots, x_n)$ that is generated by a non-linear additive model that is faithful to a DAG G

Output: The DAG G (in the large sample limit and using a consistent regression method and independence test)

```
$S = \{1, \dots, n\}$
$\pi = [\,]$
// Phase 1: Determine a topological order by identifying a sink node, i.e.
// a node whose residual is independent of the rest of the nodes
Repeat
    For $k \in S$ do
        Regress $X_k$ on $X_{S \setminus \{k\}}$
        Measure the dependence between the residuals and $X_{S \setminus \{k\}}$
    Let $k^*$ be the $k$ with the weakest dependence
    $S = S \setminus \{k^*\}$
    $Pa_{k^*} = S$
    $\pi = [k^*, \pi]$
Until $S = \varnothing$
// Phase 2: Remove superfluous edges without violating the sink condition, i.e.
// the residual of a node is independent of its predecessors or non-descendants
    For $k \in \{2, \dots, n\}$ do
        For $\ell \in Pa_{\pi(k)}$ do
            Regress $X_{\pi(k)}$ on $X_{Pa_{\pi(k)} \setminus \{\ell\}}$
            If the residuals are independent of $X_{\pi(1:k-1)}$, then remove $\ell$ from $Pa_{\pi(k)}$
```

- Note that the noise variables are jointly independent (cf. IC* algorithm).

Score Based Algorithms

- Choose a DAG G with maximum posterior probability given some data $d_{1:N}$ (a.k.a. Bayesian score), i.e.

$$p(G|d_{1:N}) = p(d_{1:N}|G)p(G)/P(d_{1:N}) \propto p(d_{1:N}|G)p(G)$$

  where $p(d_{1:N}|G)$ is the marginal likelihood of $d_{1:N}$ given G, $p(G)$ is a prior probability distribution over DAGs, and $p(d_{1:N})$ is a normalization constant.

- Moreover,

$$p(d_{1:N}|G) = \int p(d_{1:N}|\theta_G, G)p(\theta_G|G)d\theta_G$$

  where $p(d_{1:N}|\theta_G, G)$ is the likelihood function of $d_{1:N}$ given G and θ_G, and $p(\theta_G|G)$ is a prior probability distribution over the parameter values of G.

- For discrete variables X_i of cardinality k_i, and **assuming** that $p(\theta_G|G) = \prod_i \prod_j p(\theta_{x_i|pa_i=j}|G)$ and $p(\theta_{x_i|pa_i=j}|G) \sim Dirichlet(\alpha_{ij1}, \ldots, \alpha_{ijk_i})$, we have that

$$p(d_{1:N}|G) = \prod_i \prod_j \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_k \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

  where $\alpha_{ij} = \sum_k \alpha_{ijk}$, N_{ijk} is the number of instances in $d_{1:N}$ where $X_i = k$ and $Pa_i = j$, and $N_{ij} = \sum_k N_{ijk}$.

Score Based Algorithms

- Two DAGs are called Markov equivalent if they represent the same set of separations.
- The marginal likelihood is the same for Markov equivalent DAGs iff

$$\alpha_{ijk} = \alpha p'(ijk)$$

  where α is the user-defined imaginary sample size and $p'(ijk)$ is a prior probability distribution, e.g. $p'(ijk) = 1/(k_i \prod_{X_\ell \in Pa_i} k_\ell)$.
- Under the Dirichlet parameter prior assumption and when $N \to \infty$, we get the Bayesian information criterion (BIC):

$$\log p(d_{1:N}|G) \approx \log p(d_{1:N}|\theta_G^{ML}, G) - \frac{\log N}{2} dim(G)$$

  where θ_G^{ML} are the maximum likelihood estimates of the parameters (i.e., proportions in $d_{1:N}$), and $dim(G)$ is the dimension or number of free parameters of G, i.e. $\sum_i (k_i - 1) \prod_{X_\ell \in Pa_i} k_\ell$.
- Similar results exist for Gaussian random variables.

Score Based Algorithms

- Number of DAGs with 1-12 nodes: 1, 3, 25, 543, 29281, 3781503, 1138779265, 783702329343, 1213442454842881, 4175098976430598143, 31603459396418917607425, 521939651343829405020504063
- Then, an exhaustive search is prohibitive and, thus, a heuristic search must be performed instead.

```
Hill-climbing
Input: A sample $d_{1:N}$ from a distribution $p(v)$
Output: A DAG $G$ over $V$

Let $G$ be the empty DAG
Repeat until no change occurs
    Add, remove or reverse any edge in $G$ that improves the Bayesian score the most
```

- The log Bayesian score is **decomposable** if $\log p(G)$ is so, i.e.

$$\log p(G|d_{1:N}) = \sum_i f(X_i, Pa_i, d_{1:N})$$

  and, thus, adding, removing or reversing a edge in G implies recomputing only one or two factors.
- Unfortunately, hill-climbing is not asymptotically correct.
- Note that the noise variables are jointly independent (cf. IC* algorithm).

Summary

- Causal Models
- IC Algorithm
- Projections
- IC* Algorithm
- Restricted Causal Models
- RESIT Algorithm
- Score Based Algorithms

Thank you