Causal models

Marcel Maretić
Faculty of Mechanical Engineering and Naval Architecture
University of Zagreb
Ivana Lučića 5, 10002 Zagreb, Croatia
marcel.maretic@fsb.hr

Abstract. *We present a framework for causal modeling. Main objective is to define models augmented with graphs and present relevant results and important concepts in the theory of structural causal models. Causal models assume autonomy of mechanisms involved. This feature allows us to predict their behaviour even when some of these mechanisms break or change. We focus our attention on Markovian and semi-Markovian models. Causal effect is defined. We show how and when it can be calculated.*

Keywords. causality, Bayesian networks, Markovian models, semi-Markovian models, causal effect, identifiability

1 Introduction

Are random variables functionally (causally) related, dependent? Can we learn more about relations in the system? Can we make predictions about the system even when it malfunctions? Can we infer what would be systems new probability distribution after such change?

These are high-level questions that transcend ordinary probabilistic modeling. Answers require deep knowledge and understanding of the system. Such knowledge can not be acquired by observation alone.

Here, we would like to probabilistically[1] model a system with a finite set V of relevant random variables $X_1, \ldots, X_n$. We assume that joint probability distribution P for the set V is given in all our models and examples. Also, we also that we have all conditional probabilities $P(X \mid Y)$ where X and Y are sets of variables in V.

We start with a brief description of probabilistic graphical models known as Bayesian Networks (or belief networks) which can efficiently represent joint probability distributions. Later we build on them and describe a type of Bayesian network we call causal Bayesian network.

After introducing causal Bayesian networks, we present another more powerful formal causal model called structural model and present some interesting related concepts such as calculation of causal effect and whether causal effect can be estimated from observational probabilities.

[1] We adhere to Bayesian interpretation of probability. Conditional probabilities are not defined through joint probabilities as they are fundamental in Bayesian interpretation of probability where terms $P(X \mid Y)$ represent our degree of belief in X once we know Y.

2 Bayesian Networks

Bayesian networks are graphical probabilistic models. Joint probability is augmented with a DAG (directed acyclic graph) G. Nodes of G represent variables in V. Edges (arrows) reflect conditional dependence relations between variables of the model.

Before we give definition of Bayesian network we shall introduce a few concepts.

Let X, Y and Z be disjoint subsets of the set of variables V. We would like to know if X and Y are independent conditionally on Z. If they are then the following equation holds:

$$P(X, Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z) \qquad (1)$$

for all values of X, Y, Z. Equation (1) can be transformed to:

$$P(X \mid Z) = P(X \mid Y, Z).$$

Informal reading would be:

Once we know Z,
learning Y tells nothing new about X.

This ternary relation is called a **relation of conditional independence** and is usually denoted

$$X \perp\!\!\!\perp Y \mid Z \ ,$$

or sometimes $(X \perp\!\!\!\perp Y \mid Z)_P$ if we have to specify the related distribution P.

2.1 Graph construction

To construct a graph one must find sets of parents PA_i for each variable X_i. First, we have to fix some ordering of variables, let it be $X_1, \ldots X_n$. Then for each variable X_i we have to find a minimal set of predecessors $(X_1, \ldots, X_{i-1})$ that renders all other predecessors irrelevant.

Once we have sets PA_i we can draw arrows from each PA_i-variable to its child X_i. Finally, we have a DAG G.

We call PA_i a set of **Markovian parents** of X_i. Set PA_i renders all other predecessors irrelevant for X_i:

$$P(X_i \mid PA_i) = P(X_i \mid X_1, \ldots, X_{i-1}) \ .$$

We also say that set PA_i **screens off** all other predecessors (by making them irrelevant).

Example 1. *Suppose we know that variable X_1 affects X_3, but not if we know the value of X_2. Therefore X_2 screens off X_1 from X_3. X_2 in PA_i but not X_1. Here is a graph that illustrates it:*

$$X_1 \longrightarrow X_2 \longrightarrow X_3$$

Effect of X_1 on X_3 is mediated through X_2.

2.2 Factorization of P

Let P be a joint distribution on a set $V = \{X_1, \ldots, X_n\}$.

Bayes theorem for joint probabilities gives:

$$P(X_1, X_2) = P(X_1) \cdot P(X_2 \mid X_1) \ .$$

A few simple iterations yield the following factorization:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) \ . \quad (2)$$

Equation (2) holds for any distribution P in any ordering of variables.

Suppose now we have a graph G for P such that P and G form a Bayesian network. We can read PA_i from G as parent nodes. Now we can state reduced factorization that holds in Bayesian network:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid PA_i) \quad (3)$$

So if we have a graph, we have a factorization (and vice versa).

For graphs constructed as previously the following condition holds

Markov condition. *Every node is independent of all its non-descendants conditionally on its parents.*

Definition. *Bayesian networks are pairs (G, P) such that Markov condition holds.*

2.3 *d*-separation

Conditional independence has a graphical counterpart – a criterion of d-separation between nodes in a graph. d-separation is a ternary relation between (three) sets of edges on a graph. We carry this definition in three steps:

Step 1: Let p be a trail (a series of adjacent edges, arrow direction disregarded) in a graph between nodes X and Y. We say that p is d-separated by a set of nodes Z if (at least) one of the following holds:

(i) p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that middle node m is in Z;

(ii) p contains an inverted fork $i \to m \leftarrow j$ such that m or any of its descendants are in Z.

Step 2: Nodes X and Y are d-separated by set Z if all trails between X and Y are d-separated.

Step 3: Let X, Y and Z be disjoint sets of nodes. We say that sets X and Y are d-separated by Z if all trails between nodes in X and Y are d-separated.

We write

$$(X \perp\!\!\!\perp Y \mid Z)_G \ ,$$

where G is a graph we're working with.

Theorem 1. *In Bayesian network d-separation implies conditional independence. If sets of nodes X, Y are d-separated by set Z, then X and Y are independent conditionally on Z.*

It is possible to have a graph that is too big, i.e. a graph with unnecessary edges. For example, a complete graph on V always forms If this is the case, then there are instances where conditional dependence holds but can not be deduced from d-separation. It is desirable to close this gap, i.e. to have a minimal graph.

2.4 Structure learning

Learning the structure (graph) of Bayesian networks is not in the focus of this paper and so I would like to note just a few points. Certain features of the graph structure can be deduced from the data alone, i.e. observational distribution of the system.

We are concerned here with graphs whose structure is at least partially derived from outside (expert) knowledge – knowledge of prior causal relations or physical laws, time order etc.

While it is attractive/tempting to think about arrows and paths as some causal/functional relations between nodes. Their meaning in the model is a bit more complicated. Direct link $A \to B$ *does* mean that A affects B, as expected. But, any undirected path of at least two steps between nodes A and B doesn't signify anything without additional information. With d-separation we can see that A and B can be dependent or independent dependent on the context (what is Z).

3 Causal Bayesian Networks

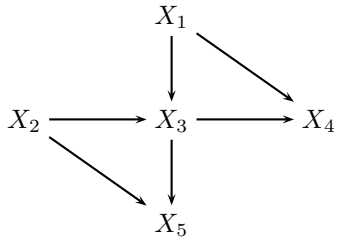
How is causal knowledge acquired? How does a child come to understand systems before learning language?

To acquire deep knowledge of an unknown system agents have to resort to experimentation. Observation alone is not enough: no amount of observation can substitute for experimentation. This reduces to well known mantra "correlation does not imply causation".

On the other hand, once we *have* a deep understanding of a system, we should know how it behaves even in unusual situations – when some of its mechanisms are replaced or broken. For example, we have an idea of what to expect if someone replaces car's wheels with bicycle wheels or pours water instead of gasoline in its tank. Notice how knowledge doesn't have to come from past experience.

3.1 Stability and Intervention

Suppose we have a Bayesian network whose edges represent stable mechanisms. Let's illustrate our point with an example Bayesian network:



We would like to intervene in this model and force a variable X_3 to take on a value x_3. This action will be written as

$$do(X_3 = x_3) \ ,$$

or just $do(x_3)$ for short.

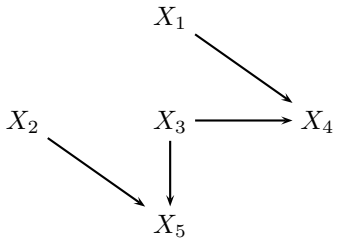
Can we calculate the effects on other variables in the model, i.e. what is the probability of $P(X_5 = x_5)$ after $do(x_3)$ intervention, written as

$$P(x_5 \mid do(x_3)) = ?$$

Definition. *Given two disjoint sets of variables X and Y, **causal effect** of X on Y is*

$$P(y \mid do(x)) \ .$$

To calculate $P(x_5 \mid do(x_3))$ we must build a new Bayesian network that reflects the effects of our action. All mechanisms affecting X_3 (represented by incoming arrows) are invalidated. Therefore we can delete X_3-incoming arrows from the graph. We get a new graph G':



Now we need a new probability distribution P' to fit G'. We get P' as "truncated factorization" of P with appropriate substitutions of $P'(x'_3)$ for $P(x'_3 \mid x_1, x_2)$:

$$P'(x'_3) = \begin{cases} 1, & x'_3 = x_3 \\ 0, & \text{else} \end{cases} \tag{4}$$

Other occurrences of x_3 in factored terms should be similarly substituted and adjusted.

Since we get fully specified P' we can calculate

$$P(x_5 \mid do(x_3)) = P'(x_3) \ .$$

P' is **interventional distribution** of action $do(x_3)$. Common shorthand for P' is P_{x_3} (or $P_{do(x_3)}$). We can extend simple actions to actions on sets of variables.

Two conditions hold

Theorem 2.

$$P(v_i \mid pa_i) = P_{pa_i}(v_i) \ .$$

Seeing and doing is equal for V_i if done to its Markov parents.

Theorem 3.

$$P_{pa_i,s}(v_i) = P_{pa_i}(v_i)$$

Once we control Markovian parents of V_i, no other intervention will affect probability of V_i.

Remark. *"Seeing" and "doing" should not be confused. Under certain conditions we can exchange action for observation but we have to be careful about it (see later).*

$$P(x_5 \mid do(x_3)) \neq P(x_5 \mid x_3)$$

Now we can give a formal definition of a causal Bayesian network.:

Definition. *Causal Bayesian Network (CBN) is a Bayesian network together with a set of interventional distributions P_* .*

Since post-interventional probability distributions are completely specified in CBN's, we can calculate causal effects.

4 Functional causal model

We now introduce a new formalism as an alternative to causal Bayesian networks.

A **functional causal model** consists of a set of equations:

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n \ .$$

f_i are unspecified functions, pa_i represent relevant selection of $x_1, \ldots, x_{i-1}$, and u_i are errors or disturbances due to omitted factors.

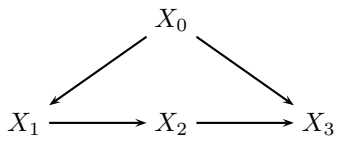
A set of equations where each equation is autonomous is called **structural model**.

With every functional model M we can construct a DAG as follows: if a variable x_j appears on the right hand side of i-th equation we draw an arrow from node X_j to node X_i. Resulting graph is called a **causal diagram** of M.

Example 2. *Let M be a model given by*

$$\begin{aligned} x_1 &= f_1(u_1) \\ x_2 &= f_2(x_2, u_2) \\ x_3 &= f_3(x_2, u_3) \\ x_4 &= f_4(x_1, x_3, u_4) \end{aligned}$$

M's associated causal diagram is:



Disturbances u_i are taken to be values of random variables U_i which we call **unobserved variables**. Disturbance variables are normally not included (drawn) in the graph.

If the associated causal diagram is a acyclic (DAG), then the corresponding model is called **semi-Markovian**. From now on we will only consider semi-Markovian models.

We can take x_i to be values of variables X_i. We call $V = \{X_i\}$ a set of **observed variables**. Probability distribution P_0 on unobserved variables $\{U_i\}$ induces a probability distribution P on a set V.

Remark. *Somewhat surprising might be the fact that functions f_i and distribution of unobserved variables U_i remain unspecified. But what we are looking for here is the structure of the model, i.e. its causal diagram. Specification of f_i and u_i is not needed for structure.*

4.1 Markovian models

We'll discuss a special type of semi-Markovian model first. If the unobserved variables U_i are independent, we have a **Markovian model**. In Markovian models induced distribution P along with associated causal diagram G satisfies Markov condition (hence the name). Also, factorization (3) for P holds on V.

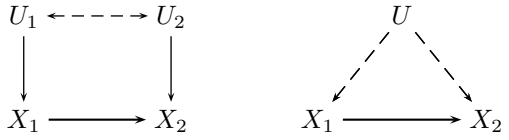
Associated diagram G and induced distribution P form a causal Bayesian network, and it is just as easy to convert a CBN to a Markovian structured model.

Every Markovian model is as descriptive[2] as a CBN. Also, all causal effects can be calculated from pre-intervention distribution.

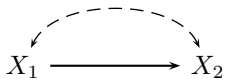
5 Semi-Markovian models

If we relax the condition of independence of unobserved variables which defines Markovian models we lose Markov condition (and factorization etc.)

If some U_i's are dependent, then this assumption must enter the model. We model this with **latent** variables. Latent variables are unmeasured variables that have exactly two observed children. We also say that affected observed variables are **confounded**.



As with Markovian models we exclude unobserved variables from the graph, but we mark confounding (background dependencies) with bi-directional arcs. Latent variables are not drawn (they hide behind the arcs).



As before, we would like to express causal effect in terms of pre-interventional probability distribution. In semi-Markovian models this is not always possible. If causal effect of X on Y can be calculated, we say that it is **identifiable**.

5.1 Calculus of Intervention

Let $G_{\underline{X}}$ denote graph in which all X-incoming arrows are deleted and similarly let $G_{\overline{X}}$ denote graph in which all X-outgoing arrows are deleted.

$P(y \mid do(x), z)$ denotes "probability that $Y = y$ after we see $Z = z$ when we $do(X = x)$".

Theorem 4 (Rules of *do* Calculus)**.** *Let G be a DAG associated to a causal model, and let $P(\cdot)$ be a probability distribution induced by that model.*

For disjoint sets of variables X, Y, Z, W we have the following rules:

Rule 1 *(Insertion/deletion of observations)*

$$P(y \mid do(x), z, w) = P(y \mid do(x), w)$$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$.

Rule 2 *(Action/observation exchange)*

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w)$$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$.

Rule 3 *(Insertion/deletion of actions)*

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), w)$$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X},\overline{Z(W)}}}$ *, where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.*

[2] It turns out that structural models unlike causal Bayesian networks allow counterfactual analysis (see [1]).

Right hand sides of rules 1–3 simplify $P(\cdot)$ expressions as they eliminate do operator. d-separation on modified graphs dictate when a rule applies. Explanation and proofs can be found in [1].

Theorem 5 (Completeness)**.** *Calculus of intervention is complete. It is sufficient for deriving all identifiable causal effects.*

5.2 Identifiability Criterion

Now that we know which syntactic transformations are sufficient, we should answer when such transformations yield an answer. We want to know when causal effect can be identified.

Here we present one simple criterion for identifiability.

Theorem 6 (Tian and Pearl, [3])**.** *A sufficient condition for identifying the causal effect $P(y \mid do(x))$ is that there exists no bi-directed path between X and any of its children.*

So, whenever there exist no bi-directed paths between X and any of its children we know that causal effect can be identified. Otherwise, when such a path exists, we don't know.

A complete criterion (called *hedge criterion*) is presented in [5]. As the statement of this result is significantly more complex, we omit it here.

6 Conclusion

I'd like to share a few notes on the metaphysics of causation. A question "What does it mean for (an event) A to cause B?" has taunted philosophers from ancient times. A satisfying answer has not been found yet. Contemporary philosophers nowadays increasingly believe that we must be pluralists about causality – that it means many different things to us. Therefore, there are several explanations.

Structural models don't define causality and don't rely on any particular definition of causality. We can drop the word "causal" from the terminology. A crucial point is: structural models consist of stable mechanisms between variables; changes are local, not widespread (replacing one mechanism leaves other mechanisms intact).

On the other hand, if we wish to reason about such systems and describe how they respond to change – we might as well use causal language because it fits well and we are accustomed to it.

References

[1] J. Pearl: Causality: Models, Reasoning and Inference, 2nd edition, Cambridge University Press, New York, 2009.

[2] J. Tian, J. Pearl: A New characterization of the Experimental Implications of Causal Bayesian Networks, 2002.

[3] J. Tian, J. Pearl: A General Identification Condition for Causal Effects, National Conference on Artificial Intelligence, 2002.

[4] P. Spirtes, C. Glymour, R. Scheines: Causality, Prediction and Search, 2nd edition, MIT Press, 2001.

[5] I. Shpitser: Complete Identification Methods for Causal Inference, Doctoral Thesis, 2008.

[6] J. Williamson: Bayesian Nets and Causality, Oxford University Press, USA, 2005.