A Gentle Introduction into Structural Causal Models

Daniel Saggau · daniel.saggau@campus.lmu.de
Department of Statistics, Ludwig Maximilian University Munich, Germany

June 2021

Abstract The interest in understanding relationships of variables beyond co-occurrence has increased the popularity of causal modelling. Probabilistic specifications cast a model based on conditional probabilities. SCMs cast a model based on assymetric assignments and extend probabilistic models by specifying the entire data generating process rather than solely utilizing conditional probabilities. Another difference between these models is their ability to address different queries such as *predictions*, *interventions* and *counterfactuals*. These queries are part of Pearl's causal hierarchy (2009). Pearl matches these queries with their respective actions namely *observing*, *doing* and *imagining*. I compare the feasibility of addressing these queries and undertaking respective actions for both specifications. To contextualize SCMs within the field of causality, I also discuss the role of time in causality. This paper uses various directed acyclic graphs to highlight the differences in these modelling approaches. The insights of this paper can be used as a baseline for subsequent research on structural causal models.

Contents

1 Introduction

For many research problems, we want to understand the relationship between variables beyond co-occurrence. The most popular causal model is the structural causal model or in short SCM. The SCM is the non-parametric counterpart to structural equation models. Researchers on SCMs tried to distance themselves from common practices for structural equation models (Pearl 2009a; Peters, Janzing, and Schölkopf 2017). The geneticist and statistican Sewall Wright introduced the first ancestor of the SCM, the path analysis (Pearl 2009a). Path analysis now falls into the broader class of structural equation models. Path analysis is a structural equation model with one variable per indicator.

The SCM is an expressive simulator to estimate causal relationships, accounting for latent factors. Latent factors are unobserved factors (Pearl 2009a). SCMs entail endogenous and exogenous variables. Exogenous variables should not be confused with latent variables. For reference, Peters, Janzing, and Schölkopf (2017) define endogenous and exogenous as follows: *Endogeneous variables are those that the modeler tries to understand, while exogenous ones are determined by factors outside the model, and are taken as given.* The basis of these SCMs is a set of equations, or more precisely assignments, providing functions to derive the conditional probabilities for our model (Hardt and Recht 2021). These assignments describe our variables in our model. Compared to probabilistic models, where we only specify conditional probabilities, the SCM actually enables the combination of latent variables and observational data. Conditional probabilities cannot represent latent variables because there is no conditional probability in our observational data for unobserved variables (Pearl 2009a).

To accommodate existing literature, this paper provides a gentle introduction to SCMs, focusing on the underlying assumptions we make when building a structural causal model. The aim of this paper is to provide an example based, inutitive understanding of different concepts within causality using manifold elements from core literature on causality.

The rest of the paper is structured as follows: Section 2 introduces the assumptions in causal modelling. Section 3 addresses Pearls Causal Hierachy. Section 4 focuses on the intersection of SCMs and the perception of time. Section 5 concludes.

2 Assumptions in Structural Causal Models

The structural causal model consists of a set of equations. These equations are asymmetric assignments, because they are not bi-directional. We cannot change the sides of the equation as one can do with regular equations (Pearl 2012a). (Peters, Janzing, and Schölkopf 2017) define a SCM as follows:

Definition 1: Structural Causal Model:

An SCM $\mathbb{C}$ with graph $C \to E$ consists of two assignments

$$C := N_C$$
$$E := f_E(C, N_E)$$

where $N_F \perp\!\!\!\perp N_C$ that is N_F is independent of N_C

In their definition, C is the cause and E is the effect.

We can set up an epidemiological example with three variables, looking at the impact of problem behavior and genetic code on lung cancer. Problem behavior is a latent variable, but we can use observational data to characterize problem behavior. One example of problem behavior is smoking. Henceforth,to estimate problem behaviour, we can for instance ask participants how frequently they smoke. For genetic code, we could look for specific genes and and examine whether the presence of specific genes has an impact on getting cancer. As we can see, these functions are driven by underlying latent variables. These latent factors are the foundation of the structural causal model (Hardt and Recht 2021; Pearl 2009a, 2012a).

$$S := f_S(U_S) \tag{1}$$

$$G := f_G(U_G) \tag{2}$$

$$C := f_C(S, G, U_C) \tag{3}$$

where: $\{S\}$ - Frequency of smoking $\{G\}$ - Presence of specific genes $\{C\}$ - Lung cancer

Every structural causal model contains an underlying graphical model (Hardt and Recht 2021). This is one important feature that differentiates SCMs from other frameworks[1].

(Pearl 2009a) describes the SCM a process based tool, because it enables researchers to reflect on their underlying assumptions. The SCM requires more assumptions and thought. Even for a very minimalisitic SCM, we need to define an admissible set of variables, ensure the random noise terms are independent and corroborate that the underlying mechanisms are autonomous. By being forced to think about all these steps, SCMs help to avoid poorly specified probabilistic specifications. Various research has pointed out examples where modeling without DAGs lead to severe mistakes: Hirano and Imbens (2001) suggest a method for covariate selection that according to Pearl (2009b) favours bias-enhancing features in the propensity score. Further Bollen and Pearl (2013) (2013) argue that Rosenbaum (2002) and Rubin (2007) falsely declared that 'there is no reason to avoid adjustment for a variable describing subjects before treatment.'

We can craete graphical models based on various different algorithms.

[1] e.g. the Potential Outcome framework(Pearl 2009a)

The most popular graphical model is the directed acyclic graph or in short DAG. A causal graph for a SCM contains endogenous and exogenous variables. A DAG entails nodes and edges.

Nodes represent our different variables. Edges depict the assignment equations. All edges are directed in the DAG. An acyclic graph has no roots that cause itself (directly and indirectly) (Morgan and Winship 2014).

This acyclic structure is important for the conditional probabilities (Forré and Mooij 2020).

may not find unique solution if cyclic in equilbrium (Peters, Janzing, and Schölkopf 2017)

Variables have incoming paths from their parent nodes.

These DAGs can be built on theory. Another way to determine the DAG structure is using observational data. One of the more prominent algorithm to estimate the underlying dag structure is the pc-algorithm (Kalisch et al. 2012) .

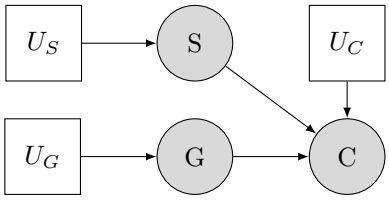

Figure 1: Structural Causal Model

Graphical models developed a mathematical langauge to assess certain conditions. One pivotal feature is the d-separation (Pearl 2010):

Definition 2 (d -separation) A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. If S blocks all paths from X to Y, it is said to " d -separate X and Y, " and then, X and Y are independent given S, written $X \perp\!\!\!\perp Y \mid S$.

Select subset of variables with backdoor criterion:

2.1 Independence

Definition: Independence *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions, In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution* (Peters, Janzing, and Schölkopf 2017).

If we specify the causal structure correctly:

(a) possible to undertake local intervention -> change f(x), regardless of f(y | x)

(b) these components are autonomous objects -> set of autonomous equations

Independence of Noise,

Noise independent so problematic behaviour and genetic code should be independent.

Mechanism

$$\begin{aligned} p(a,t) &= p(a \mid t)p(t) \\ &= p(t \mid a)p(a) \end{aligned} \tag{4}$$

$$p^{o}(a,t) = p(t \mid a)p^{o}(a) \text{ and } p^{s}(a,t) = p(t \mid a)p^{s}(a). \tag{5}$$

Definition: Causal Sufficiency

A set of variables X is usually said to be causally sufficient if there is no hidden common cause $C \notin X$ that is causing more than one variable in X (Peters, Janzing, and Schölkopf 2017; Spirtes 2010)

$$P(Cancer|Smoking) = P(Smoking|Cancer) \times P(Cancer)$$

Markov Condition

Pearl (2009a) defines the markov condition as follows:

Definition: Causal Markov Condition

Every Markovian causal model M induces a distribution P(x1,. . . , xn) that satisfies the parental Markov condition relative the causal diagram G associated with M; that is, each variable Xi is independent of all its nondescendants, given its parents PAi in G.

$$P_x(v) = \prod_{\{i|V_i \notin X\}} P\left(v_i \mid pa_i\right)$$
for all v consistent with x

2.2 Time

why time is ignored.

3 Pearl's Causal Hierachy

Pearl (Bareinboim et al. 2020; Pearl 2009a) introduced the hierachy of causation to categorize different statistical and causal tools. The hierarchy of causation contains (see table 1) three levels. The higher the method on the hierarchy, the more information the method requires. This section discusses each of these methods and their respective advantages.

Table 1: Pearls Hierachy of Causation (2009)

Method	Action	Example	Usage
Association $P(a\|b)$	Co-occurrence	What happened…	(Un-)Supervised ML, BN, Reg.
Intervention $P(a\|do(b), c)$	Do-manipulation	What happens if …	CBN,MDP,RL
Counterfactual $P(a_b\|a', b')$	Hypotheticals	What would have happened if…	SCM ,PO

3.1 Association:

The first level, association, requires the least information. Association based methods are most prevalent and contain the largest class of methods. Standard statistical tools such as regression analysis, supervised and unsupervsied learning and bayesian networks all fall into this category.(Bareinboim et al. 2020) The underlying action for association is co-occurence. As prominently criticized by Bender et al. (2021), this reduces the number of questions we can answer, because methods are heavily dependent on the observational data. In the context of deep learning and the advancement of natural language processing, Bender et al. (2021) suggest that many association-based methods results in stochastic parrots as opposed to natural langauge understanding.

Association-based methods ignore external changes outside of our data. The interventional distribution has information on these external changes. Note, that the intervention distribution is only defined in high order methods.

Reiterating our lung cancer example, the full joint distribution looks as follows:

$$P(c, s, g) = P(s) \times P(c|s, g) \times P(g)$$

If there are any changes in our distribution, e.g. smoking changes its distribution from s to s_{new} we are unable to accomodate these changes (Peters, Janzing, and Schölkopf 2017). Association-based methods are not equipped to accodate such modifications.

Graph

One example is the bayesian network, which uses conditional probabilities instead of functions to describe the relationship between variables (Pearl 2009a) In the probabilistic representation, we ignore latent factors (Creager et al. 2020; Pearl 2009a).

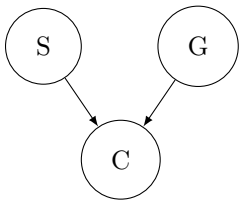

Figure 2: Probabilistic Model

3.2 Intervention:

Here we can use Pearl (2009a) do-calculus.

The do-calculus enables us to study the manipulation of parent nodes.

Instead of merely seeing the co-occurence of variables, we can actively manipulate the conditional distribution of one variable. Commonly, one could use this method to evaluate different policies (Creager et al. 2020). In our smoking example we can actively set the smoking behavior to a fixed value or a different conditional probability.

Joint distribution post intervention: $P_{S=s}(c, g) = P(c|S = s, g) \times P(g)$

There are various types of intervention.

I will illustrate atomic intervention and policy intervention based on our smoking example.

Suppose the government wants to examine, a different smoking behavior distribution will lower the number of lung cancer cases and medical expenses. This different smoking behaviour could be the result of increased taxes, or e.g. finesup to a $1000 (as in the case of Singapore) for smoking in prohibited areas. Alternative, the government could also undertake treatment by imposing higher tobacco taxes. In this case we solely focus on the effect of a different smoking behavior on lung cancer for the sake of simplicity.

In **atomic intervention,** we set a variable to a constant value. As one can see in figure 3, c is constant that is not dependent on the latent factor, because in atomic intervention we do not derive the value of c based on the function S.

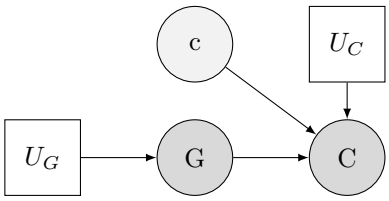

Figure 3: Atomic Intervention

Further, in mathematical notation the model would change as follows:

$$S := s \tag{6}$$

$$G := f_G(U_G) \tag{7}$$

$$C := f_C(S, G, U_C) \tag{8}$$

In **policy intervention** we specify a different conditional probability $(S = s)$ for an equation. We can derive s from S, because we include information on the intervention distribution. This information cannot be obtained, if we directly specify our model as conditional probabilities.

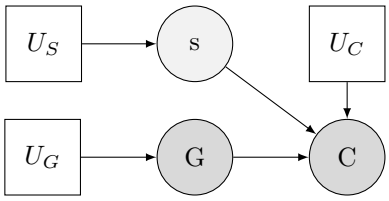

Figure 4: Policy Intervention

where s is our new conditional probability.

One extension of policy intervention is off-policy intervention. **Off-policy intervention** models different intervention that is not in our historical data. To accommodate the lack of data, we can estimate the intervention either based on model-based (e.g. regression),model-free estimates (propensity scores), or a mixture of these methods. For further information, see e.g. Oberst and Sontag (2019) or (Creager et al. 2020).

$do(t)$ -> replace function $T := f_T(\pi)$ with different conditional probability (or constant) (Pearl 2009a).

Counterfactuals:

Process is described as follows:

(a) Abduction: Cast probability $P(u)$ as conditional probability $P(u|\epsilon)$
(b) Action: Exchange $(X = x)$
(c) Prediction: Compute $(Y = y)$

Stable Unit Treatment Value Assumption {SUTVA} 'The treatment that one unit receives does not change the effect of treatment for any other unit.'

Consistency The outcome Y agrees with the potential outcome corresponding to the treatment indicator.'

Ignorability The potential outcomes are conditionally independent of treatment given some set of deconfounding variables. As suggested by (Hardt and Recht 2021), this condition ensures that we are dealing with a perfect randomized controlled trial.

- First two hold for Counterfactuals in SCM
- third not testable but can check via backdoor criterion in SCM
- Source: (Hardt and Recht 2021)

4 SCMs and Time

largely ingore time

time in mechanical modeling crucial (Peters, Janzing, and Schölkopf 2017).

- Time in Social Sciences: Often Vague
- Time in Physical Sciences: Mechanical via **Differential equations**
- dependence on prior time point and change in time contribute to the value at time point

Initial Value:

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Derivative of function x with respect to time t:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$$

Value of Function at time t + dt:

$$\mathbf{x}(t + dt) = \mathbf{x}(t) + dt \cdot f(\mathbf{x}(t))$$

model	IID setting	changing distributions	counter-factual questions	physical insight
mechanistic model	Y	Y	Y	Y
structural causal model	Y	Y	Y	N
causal graphical model	Y	Y	N	N
statistical model	Y	N	N	N

Table: Source: Peters, Janzing, and Schölkopf (2017)

5 Conclusion

Structural causal models are flexible simulators to disentangle causality for manifold different queries. There are many advantages of structural causal models: (1) We are able to model latent fators forcing us to reconsider existing assumptions about the relationship in our data. (2) Further, we get a underlying graphical representation including a mathematical language for this graphical systems to test causal assumptions that are otherwise untestable. (3) Additionally, one is able to model queries beyond mere association going as far as dealing with hyptoethical situations. Simultaneously,

6 References

Bareinboim, Elias, JD Correa, Duligur Ibeling, and Thomas Icard. 2020. "On Pearl's Hierarchy and the Foundations of Causal Inference." *ACM Special Volume in Honor of Judea Pearl (Provisional Title)*.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23.

Bollen, Kenneth A, and Judea Pearl. 2013. "Eight Myths about Causality and Structural Equation Models." In *Handbook of Causal Analysis for Social Research*, 301–28. Springer.

Chiappa, Silvia, and William S. Isaac. 2019. "A Causal Bayesian Networks Viewpoint on Fairness." *arXiv:1907.06430 [Cs, Stat]* 547: 3–20. https://doi.org/10.1007/978-3-030-16744-8_1.

Cortés, Efrén Cruz, and Debashis Ghosh. 2019. "A Simulation Based Dynamic Evaluation Framework for System-Wide Algorithmic Fairness." *arXiv:1903.09209 [Cs]*, March. http://arxiv.org/abs/1903.09209.

Creager, Elliot, David Madras, Toniann Pitassi, and Richard Zemel. 2020. "Causal Modeling for Fairness in Dynamical Systems." In *Proceedings of the 37th International Conference on Machine Learning*, edited by Hal Daumé III and Aarti Singh, 119:2185–95. Proceedings of Machine Learning Research. PMLR.http://proceedings.mlr.press/v119/creager20a.html.

Forré, Patrick, and Joris M Mooij. 2020. "Causal Calculus in the Presence of Cycles, Latent Confounders and Selection Bias." In *Uncertainty in Artificial Intelligence*, 71–80. PMLR.

Hardt, Moritz, and Benjamin Recht. 2021. *Patterns, Predictions, and Actions: A Story about Machine Learning*. https://mlstory.org. http://arxiv.org/abs/2102.05242.

Hernan, M. A. 2015. "Invited Commentary: Agent-Based Models for Causal Inference–Reweighting Data and Theory in Epidemiology." *American Journal of Epidemiology* 181 (2): 103–5. https://doi.org/10.1093/aje/kwu272.

Hernán, Miguel A, and James M Robins. n.d. "Causal Inference: What If," 311.

Hernán, Miguel A, and Sarah L Taubman. 2008. "Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions." *International Journal of Obesity* 32 (3): S8–14.

Hirano, Keisuke, and Guido W Imbens. 2001. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research Methodology* 2 (3): 259–78.

Kalisch, Markus, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. 2012. "Causal Inference Using Graphical Models with the r Package Pcalg." *Journal of Statistical Software* 47 (11): 1–26.

Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. 2020. "Algorithmic Recourse: From Counterfactual Explanations to Interventions." *arXiv:2002.06278 [Cs, Stat]*, October. http://arxiv.org/abs/2002.06278.

Kilbertus, Niki, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2018. "Avoiding Discrimination Through Causal Reasoning." *arXiv:1706.02744 [Cs, Stat]*, January. http://arxiv.org/abs/1706.02744.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2018. "Counterfactual Fairness." *arXiv:1703.06856 [Cs, Stat]*, March. http://arxiv.org/abs/1703.06856.

Kügelgen, Julius von, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2021. "On the Fairness of Causal Algorithmic Recourse." *arXiv:2010.06529 [Cs, Stat]*, February. http://arxiv.org/abs/2010.06529.

Mooij, Joris M, Dominik Janzing, and Bernhard Schölkopf. 2013. "From Ordinary Differential Equations to Structural Causal Models: The Deterministic Case." *arXiv Preprint arXiv:1304.7920*.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Analytical Methods for Social Research. Cambridge University Press. https://doi.org/10.1017/CBO9781107587991.

Ness, Robert Osazuwa, Kaushal Paneri, and Olga Vitek. 2019. "Integrating Markov Processes with Structural Causal Modeling Enables Counterfactual Inference in Complex Systems." *arXiv:1911.02175 [Cs, q-Bio, Stat]*, November. http://arxiv.org/abs/1911.02175.

Oberst, Michael, and David Sontag. 2019. "Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models." In *International Conference on Machine Learning*, 4881–90. PMLR.

Pearl, Judea. 2009a. *Causality*. Cambridge university press.

———. 2009b. "Myth, Confusion, and Science in Causal Analysis."

———. 2010. "An Introduction to Causal Inference." *The International Journal of Biostatistics* 6 (2): 1–62.

———. 2012a. "The Causal Foundations of Structural Equation Modeling." California Univ Los Angeles Dept of Computer Science.

———. 2012b. "The Do-Calculus Revisited." *arXiv Preprint arXiv:1210.4852*.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. Cambridge, Massachuestts:

The MIT Press.

Rosenbaum, Paul R. 2002. "Overt Bias in Observational Studies." In *Observational Studies*, 71–104. Springer.

Rubin, Donald B. 2007. "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36.

Spirtes, Peter. 2010. "Introduction to Causal Inference." *Journal of Machine Learning Research* 11 (5).

Tarka, Piotr. 2018. "An Overview of Structural Equation Modeling: Its Beginnings, Historical Development, Usefulness and Controversies in the Social Sciences." *Quality & Quantity* 52 (1): 313–54.

Turbin, Mark S, Richard Jessor, and Frances M Costa. 2000. "Adolescent Cigarette Smoking: Health-Related Behavior or Normative Transgression?" *Prevention Science* 1 (3): 115–24.