POLITECNICO DI TORINO

MASTER's Degree in COMPUTER ENGINEERING

MASTER's Degree Thesis

Bridging Ananke with Pomegranate for Surrogate Experiments

Supervisor
Francesco VACCARINO

Candidate
Andrea PANUCCIO

DECEMBER 2025

Abstract

We delve into the statistical inference limitations, especially with respect to experimental results derivations, and how causal inference provides a formal approach that generalizes randomized controlled trials for the scientific inquiry. We also develop an interface to bridge between the well known frameworks of Ananke and Pomegranate, to perform asymptotically debiased effect estimations, from surrogate experiments, in absence of hidden confounders. Thus, we analyze how relevant errors of finite sample size are, to see if a pilot study may effectively be used to decide the best. Lastly we examine other causal tasks and limitations of current approaches and the one we propose, as well as possible future improvements. The code is available on GitHub, here.

Contents

1 Statistical Models and the IID Assumption Limitations

1.1 Statistical Learning Theory

Statistical inference is a branch of statistics that tries to build models, leveraging data and prior knowledge, such that the model is able to predict on unseen data. The two approaches, which differ by the way they model their assumptions, are the Frequentist one, that leverages Hypothesis Classes of functions, so to restrict the model to the space of a certain parametric function configurations, and the Bayesian one, which does not assume any functional form, but models dependencies by conditional probabilities, also relying on arbitrary priors. Yet both leverage the Central Limit and PAC Learning Theorems to provide at least theoretical guarantees to build asymptotically unbiased estimators.

1.2 IID Assumption Limitations

But the statistical learning theory builds many of its models on the assumptions that its training sample is independent and identically distributed. We can relax the independence assumption as far as the process is ergodic, thus quickly converging to a stationary distribution independent on the starting conditions, but here we follow a different, not mutually exclusive to it, path of generalization, which is based on different assumptions. If we can assume that the relationships between our model variables are stable per se, we can provide generalization on arbitrary shifts in the overall distributions, reassessing our priors. In other words, the causal framework provides a way to fit models with stable mechanisms, that perform domain adaptation by just finding the new priors with a small dataset.

2 Effect Estimation and Exogeneity Under RCTs

2.1 Estimating an Effect

Estimating an effect is essential for the scientific research since, no matter how beautiful a theory is, if it fails to predict evidence under certain system manipulations, it is simply untrustworthy an useless. Yet when we talk about effects we are making assumptions on the stability of such a relationship between the quantities we are measuring. Thus the will to limit, as possible, any bias the whole process may bring, which may lead to, maybe even repeatable tests, but that would fail outside of the test environment or the standard conditions.

2.2 Exogeneity

To grant that a variable influence on another is stable and unbiased, at least with an infinite sample size, causal frameworks require our controlling variables to be exogenous, that means that the variables which we assume to be the causes do not have to be influenced by anything that may depend on the experimental setup, our test environment. Talking about exogeneity without assuming causality, means limiting our experimentation to co-occurring events, which highly limits the scope of our models.

2.3 RCTs and Why They Grant Exogeneity

The so called golden standard for experiments are randomized controlled trials. We get a sample of our population (with no selection bias, in theory), then we randomly assign each subject to a certain group, each exposed to a fixed value of the controlled variables, so we see how differences in them lead to differences in other variables. In the causal setting we assume the firsts to be the causes and the seconds to be the effects. Such a randomization is what should grant exogeneity, since the controlled variables, being assigned randomly, shouldn't be affected by any other avaiable in the process. In practical design of experiments yet, we struggle to avoid biases, due to ethical, economical and logistical constraints.

3 A Formal Representation of Causal Assumption

3.1 Philosophical Assumptions of Causality and Science with respect to Pure Pragmatism and RL

The theoretical assumption statistical learning and, more in general, reinforcement learning is based on, follows the pragmatic view that an intelligent system can learn by building more and more abstract knowledge, leveraging concrete evidence to create a hierarchy through composition of progressively abstract concepts, all in order to optimize its capability to produce some value. Without such a given value, the pragmatic view tells us nothing on our chances of building up some knowledge. Causality adds to it the assumption that we can learn stable mechanisms even without any value. Such a view is still subjective, since both sensors and values are, and science attempts to devoid itself from them, by leveraging measurement items and standardizing measurement procedures. This science can claim to be intersubjective, at least in its testing. Such a step further, by the way, comes at the cost that, since there is no objective way to define value itself, any stable pattern we can claim to find can be irrelevant to anyone but us. Furthermore, since we never see the alternative outcome of anything, science is not willing to provide any explanation to phenomena, despite such practice led to highly useful models (think of Newton and gravitation, for example).

3.2 Pearlian Causal Framework and Markov Property

Out of the many causal mathematical frameworks that we have been developed (which differs by details on the assumptions of causality) the Pearlian is one of the more flexible ones. It is based on the Markov property, say the assumption that every variable is either exogenous or depending on only a subset of the other variables, and with a stable relationship. So the Causal Hierarchy Theorem states that we have three levels, called Rungs, of models, each rung adding assumptions and power of generalization at inference. Rung 1 is of models of pure correlations and can only make predictions assuming the distribution does not change. Rung 2 adds the Markov property and the power to make predictions of actions or assuming changes in the distribution; this according to most of science. Rung 3 adds priors on the hypothesis class of mechanisms, making counterfactual statements and inferring causes from effects; this goes out of the agreed scope of science, but delves into the philosophical one.

3.3 Nested Markov Models

A recent Rung 2 alternative, stemming from the Pearlian framework, is the one of Nested Markov Models, which is the one we will use. While the Pearlian framework leverages DAGs forced to model latent variables, NMM uses Acyclic Directed Mixed Graphs (ADMG), by collapsing latent nodes and their direct edges to bidirected edges. While the Pearlian framework provides a sound (with no hidden confounder) and complete algebra named do-calculus, leveraging three different mappings of $< DAG, formula >$ pairs, NMM provides a single mapping that leverages the general definition of Kernel and a Fixing operation iteratively. The causal graph can also be (sometimes partially) discovered from data, and the some time ago better constraint based discovery algorithm, the PC-stable, differed from the do-calculus and could not find come constraints, like the so called Verma constraint. NMM uses the same algorithm that uses for inference and finds all the so far known constraints, reducing the set of possible causal graphs to the theoretical minimum. Our choice, of using the NMM framework is so justified, in the design of experiments setting, for its elegance and maximal power in case of possible improvements regarding the graph discovery.

4 z-Identifiability and Surrogate Experiments

4.1 Identifiability Under no Hidden Confounder

Pearlian and NMM frameworks, as anticipated, have an algebra to derive effects bypassing their actual measurement, an algorithm that we use in this work so it is better to introduce it. The manipulation of the pair $< ADMG, formula >$ is called Identification and if we are able to go back to our observational ADMG, we say the quantity, say the query we wanted to derive, is Identifiable and the formula we have obtained by those manipulations is one Identification formula of such a query (there could be many, equivalent). Since such Identification algebras are complete but not closed with respect to the pair object it works on, Identification is not guaranteed to be always achieved, but we can say a query is not Identifiable if such algorithms cannot find it. It is important to note also that soundness is granted only in absence of hidden confounders that is, if we are not modeling variables affecting certain other ones in our models, we cannot guarantee any asymptotically unbiased estimation.

4.2 z-Identifiability and Generalization to Surrogate Experiments

Identification can be generalized by mapping, not from the query pair to the observational one, but to another one we want, so that, if we can intervene on a subset of our variables z, say perform an RCT manipulating the variables of z, we can compute such experiment, that we so call Surrogate Experiment, and say that the query is z-Identifiable for that z. So Identification can be seen as a special case of z-Identification, with $z = \emptyset$. A different experimental setup corresponds to a different causal graph and it is useful to think of it this way. There are cases in which we cannot manipulate certain quantities for ethical, economical, logistical reasons, but we can intervene on others: here surrogate experiments can be a good ally.

5 A Cost Estimation Framework to Drive Cost Effective Experiments

5.1 Different Identification Mechanisms Have Different Formulas

As we said, given a certain query, there may be multiple ways to identify it, that is, multiple formulas we could use, and that will be all asymptotically unbiased in absence of hidden confounders. But in real cases we will always work with a finite sample size, often willing to keep it as small as possible, to reduce costs and impact on the environment. This leads us to our question: can we derive insights on which experiment to run, given many possible ones, that are accurate enough to be reliable?

5.2 Estimating Sample Sizes and Providing Cost Functions

We assume we have been provided with a cost function of each experimental setup, function of the sample size, $cost_i : |sample_i| \to \mathbb{R}^+$, since the possible controllable variables are known a priori, thus the powerset of them is the set of all possible setups (the empty set being the observational one). It is possible, therefore, to collect an observational pilot study and estimate each setup, by simulating the postinterventional distributions of the various setups by interventions on the model. Once the sample sizes get retrieved the problem of the best cost seems straightforward, so we will assume $cost = |sample_i|$ for simplicity. What we want to know is how much reliable such procedure is and how much relying on surrogates in the first place.

6 Tool Practical Implementation

Our code was written in Python v3.11.13. The notebook can be found on GitHub here.

6.1 The Benchmark suite

The main libraries we used are:

6.1.1 Microsoft CSuite

A cluster of synthetic processes built by Microsoft while working on the DECI paper [1]. It consists of 15 processes and both an observational and a postinterventional drawn distributions are accessible at the URL
`https://github.com/microsoft/csuite/releases/download/v0.1/csuite_{name}.zip`
by an HTTP request. The name field refers to the process name and we used all of them in our tests:

- lingauss
- linexp
- nonlingauss
- nonlin_simpson
- symprod_simpson
- large_backdoor
- weak_arrows
- cat_to_cts
- cts_to_cat
- mixed_simpson
- large_backdoor_binary_t
- weak_arrows_binary_t
- mixed_confounding
- cat_chain

- cat_collider

The object wrapped in the response also contains metadata like the causal graph representing the process. The full documentation can be found on GitHub here.

6.1.2 Ananke v0.5.0

Ananke is a causal discovery and inference framework based on the NMM framework and developed by the team of one of its main contributors, Ilya Shpitser[2]. It also provides linear SEM models to fit the data, but we only used the identification module. The documentation can be found here.

6.1.3 Pomegranate v1.1.2

Pomegranate is a framework providing probabilistic models with a wide data format support, from numpy to pytorch arrays, and a low level customization and access, facilitating customizations [5]. We leveraged its Bayesian Network model, with Categorical and ConditionalCategorical distributions. The documentation can be found here.

6.2 Bridging between Ananke and Pomegranate

To coherently connect the two frameworks, we automate the BN structure definition from the corresponding ADMG. Since Pomegranate assumes a DAG structure, not necessarily causal, we rely on the ADMG identification formula to prevent any possible anti-causal flow of information, and map bidirected edges to arbitrary directed ones, or none if a directed was already present.
To actually compute the causal query estimate, we have defined a symbolic tree of computation, made of Node class nodes, of three different types:

- SumNode: to handle marginalizations
- ProdNode: for products of probabilities
- CPDNode: implementing the fixing and conditioning operations

A static method is involved in parsing the Ananke identification formula to build the tree. Every node can call its version of the evaluate() method to compute the partial result, recursively calling the children ones. Lastly another static method reshapes the distribution result to leave the query variables and marginalize for all of the others.

To reduce error propagation due to finite precision computation, all the tensor are manipulated as log probabilities as much as possible, and logsumexp() is used for the marginalizations.

6.3 The (meta)Experiment

For each process, we have computed 4 measurements of the same query in different setups:

- Obs: Identification from observational BN
- BNsurr: z-Identification from observational BN with simulated interventions on the model
- Surr: z-Identification from postinterventional BN
- Exp: conditional probability from postinterventional BN

Where the z set is always the degenerate case of being the actual treatment variable. We automated preprocessing by binning each variable of a certain number of bins; all but two having width evenly split in $[-3\sigma, +3\sigma]$ of the observational marginal, the remaining two for the tails. Values are so encoded as categorical in BNs. Metadata of the observational binning are then used to keep the same process for the Surr BNs and the Exp BNs.
ADMGs corresponding to Obs, Surr and Exp setup are printed for clarity; the output can be found at G.
The BNsurr setup leverages the Obs BN but the Surr formula.
Once the queries are computed, we compare their means and variance with the Exp setup as a reference. Since the output can be multidimensional, we measure the variance as the Frobenius norm of the covariance matrix of the output variables joint.
We also measure an estimate of the required sample size, using each setup as an hypothetical pilot. We define confidence = 95% (so Z-score = 1.96), relative margin of error = 0.1%, then, to get the absolute margin, it gets multiplied by the bin variance, modeled as an uniform in the bin round interval. The binning error is also accounted as Type B error, summed to the covariance matrix in the formula (Q). The final formula used is:

$$n = \frac{Z^2 ||\Sigma + Q||_2}{(M\sqrt{\frac{\sum_i |bin_i|^2}{12}})^2}$$

6.4 Results and Analysis

The tables can be found at J.
From our results we see that:

- BNsurr results differ in general from Obs ones and are closer with respect to Exp both in mean and variance on average (assuming outcome distributed as normal). Typically, leveraging the knowledge on the ADMG helps.
- BNsurr results differ in general from Surr ones, performing the same on average. Surr have higher variance but less bias, even if they use the same formula, so due to the difference between simulated and actually postinterventional dataset.
- Surr don't always match Exp even if they should, in theory, since they both come from the postiterventional distribution. The only difference is that Exp uses knowledge of the treatment being exogenous, while Surr uses the Ananke formula, that should still work in this degenerate condition. We checked the formulas and they look correct, so the error comes from the algorithm run on the model.
- The most Obs, BNsurr and Surr are biased from Exp, the most they tend to exhibit low variance. This higher bias but lower variance behavior is typically shown on non Gaussian distributions. This may be due to the generalized heuristic of the binning process, that was well suited for normally distributed variables. We expect this phenomenon to vanish with an appropriate ad hoc binning or nonparametric, kernel based, models like GPs.
- The estimated required sample size tends to vary according to the fitting quality but, as far the estimation is performed assuming normally distributed variables, as often happens in the design of experiments, the closer is the estimate to the Exp setup, the closer the empirical estimate of the required sample size itself.

Thus, assuming that the knowledge on the ADMG is reliable and that there are no impactful biases, performing an observational pilot study and simulating different postinterventional policies, may lead to better sample size estimations that bare outcome observational statistics. Also, when those assumptions hold, sample sizes don't differ as we had expected, at least if we still assume typical estimation formulas. The go to would so be to always go for the cheapest per unit setup, as far as assumptions are trusted, or drop some of the assumptions on sample size estimation used today.

7 Assumptions on the Functional Form and Mediation Analysis of Experiments

7.1 Counterfactual Queries

There are causal queries NMM framework will not answer, that are the counterfactual ones. A counterfactual is a query of what "would have been", given a certain evidence, had something changed while all the rest being kept the same. Such questions are doubly problematic from a scientific perspective: they can't be falsified at inference time, neither at training time in the first place, since we always only see only the things that happen, not the alternative outcome. More than that, it seems like their applications may be very limited at first glance. The reason is that counterfactual queries shine only when we work at the meta level, guided by a value (say a reward, in RL, or objective function), or updating the value function itself. Since, as we pointed out, science aims to be value-free, it cannot make sense, neither find use of it. Still, we now make two examples of how it is quite useful. Say you have to have a multistep decision making problem. You find that there is one state that always leads to poor results, so you would like to ask what leads to such result, to prevent it. It becomes a Root Cause Analysis problem needed to update the value of such a state transition, which is a credit assignment problem. Let's now move to another example and say you want to make a surrogate experiment because you cannot afford the real one, but neither the surrogate can be properly performed following standard design of experiment practices, maybe because they violate human rights of subjects. Looking at the counterfactual when the subject choice is not compliant to its group, allows us to simulate fixing and even decompose the effect of a variable on another, like in Mediation Analysis.

7.2 Functional Form Assumptions and Bias

Mathematically, a counterfactual takes a piece of evidence, a realization, and asks how it would change by varying part of it. Since it works at the individual level, it needs our model to be able to shift the variables only for the exogenous factors, not for the stable mechanisms, so to distinguish between mechanisms and exogenous noise. This requires the model to make assumptions on the hypothesis class of the model. While more general architectures could potentially solve part of the bias, hidden confounders as well as bias in the sampling policy itself may compromise the model in unmeasurable ways. This suggests that alternatives able to model uncertainty (like Gaussian Processes and variations, despite they also make assumptions on Kernels) may

be preferable since they are at least able to quantify their confidence.

7.3 Mediation Analysis and Ethical Experiments

The only way we can decompose an effect at Rung 2 is the computation of the Controlled Direct Effect (CDE), which is the effect we would get by fixing the mediators to a value. The CDE is so a function of mediators fixed values:

$$\begin{aligned} CDE(m) = &P(Y = y|do(X = x), do(M = m)) \\ &- P(Y = y|do(X = x'), do(M = m)) \end{aligned}$$

But this computation hides an issue: to compute the CDE we intervene not only on the mediators but also on the treatment variable, so it is as feasible as RCTs (or estimating them). The Rung 3 extension to that is the Natural Direct Effect (NDE), that allows the mediators to get the values the would get observationally:

$$NDE(m) = P(Y_M|do(x)) - P(Y_M|do(x'))$$

.
While it may sound similar, the counterfactual on m simulates the realizations at the individual level, only affecting the ones that have a different value than the required one. Furthermore, it extends to the Natural Indirect Effect (NIE), that computes the contribution of the effect that passes through the mediators:

$$NIE(m) = P(Y_M|do(x)) - P(Y_{M'}|do(x))$$

.
NDE and NIE are *not* complementary in general.
For linear models it is guaranteed that the Total Effect (or Average Treatment Effect, ATE) is their sum:

$$\begin{aligned} ATE &= \mathbb{E}[P(Y = y|do(X = x)) - P(Y = y|do(X = x'))] \\ &= NDE + NIE \end{aligned}$$

.
The cases when such a relationship does not hold, in nonlinear cases, is due to a phenomenon called Moderation, when the treatment has a nonlinear interaction with at least one of its mediators, this fixing one or the other would cancel a contribution that will not sum up at the end. The counterfactual notation allows to perform ethical experiments more easily, but relies on the assumptions of the model for its soundness, obviously.

8 A Guide to Choose between Causal Models

8.1 What Query We Have to Answer

The first question one has to answer, to decide which model to choose, is the required power of inference. If the aim is to estimate the effect of an action and to work on policy making or domain adaptation, a Rung 2 model is enough. If the goal is to derive the cause from the effect, say to perform root cause analysis, or mediation analysis, there is the need for a Rung 3 model. The first Rung 2 model has been a Causal BN (CBN) and it is still the safest option. Many alternatives, based on the Frequentist approach and Influence Functions (IF) have been proposed (like DR-Learner and DML). They were born to be more sample efficient estimators for systems involving wide spaces, but a big drawback is that they often cannot grant to have asymptotically bounded variance in Conditional ATE (CATE), being so limited to ATE. Similarly, for the Rung 3 models the first one has been a linear Structural Equation Model (SEM), which gets generalized to the nonlinear case as SCM. Today, building an SCM where edges are modeled with Multi Layer Perceptrons (MLPs) is the safest option.

8.2 Which Assumptions We Can Make

For scientific inquiry, the Rung 2 models fit the best. They need no assumption but for the ADMG structure and are intrinsically limited to describe what happens, not allowing any counterfactual statement. They also need relatively less data, so are more aligned with real experimental settings, when we want to limit our impact on the environment and our costs. Rung 3 models, instead, need assumptions on the functionals form and the noise one. If we restrict the hypothesis class too much, we end up biasing our results ending up with the wrong conclusions; if we make it flexible enough, we need more data to fit them.

8.3 MLPs on SCMs and the Sample Size We Have at Disposal

Despite SCMs still have some limitations, leveraging MLPs is enough to make them relatively data hungry. This leads us to restrict their potential usage on settings where we can afford cheap and safe trainings, like in simulated environments. More generally, SCMs still assume a sum of nonlinear functions of each cause, to the effect ($Y = \sum_i^{Pa(Y)} F_i(i)$), while the most general case would be entirely nonlinear ($Y = f(Pa(Y))$).

9 Finding the Graph if We Lack Domain Knowledge

9.1 Markov Equivalence Class and Observational Equivalence

We previously cited the Causal Discovery process as the learning of the causal graph structure from data. Here we briefly discuss it more in detail. Given a joint probability distribution, there are many possible processes that could have generated it, that may differ in their structure. So sampling from a process forgets part of the information, that cannot be retrieved by that distribution alone. The set of all plausible graphs is called Markov Equivalence Class (MEC) and can be seen as an analogous to hypothesis classes. In particular, graphs of the MEC given by the observational distribution are said to have Observational Equivalence.

9.2 Interventional Markov Equivalence Class and Interventional Equivalence

If the MEC can easily be a non singleton set, by looking at different distributions drawn from the same system under different manipulations, we can highly reduce the plausible candidates. When we can look at interventional distributions, we talk about i-MEC as the set of graphs having Interventional Equivalence, which is a subset of the MEC. more importantly, once we can restrict the set by all manipulations we can expect to face, we may not care to find "the true" process, since we have found some behaving the same way for all intents and purposes.

9.3 Choosing the Least Cost Set of Interventions Under Assumptions of No Hidden Confounder

In order to limit blind and potentially harmful interventions on an unknown system, a process that has been proposed, leverages a first discovery from purely observational data, then prescribing, through an algorithm, the least cost interventions we may need, to unveil some dependencies of interest. The limitation of such approach is that it still assumes the absence of hidden confounders, since it assumes that its prescriptions would be sufficient to provide the needed information. It cannot so prescribe, obviously, interventions on variables we ourselves are not taking into account to begin with.

10 Current Models Limitations and Research Directions

10.1 Causal Representation Learning and Latent Markov Related Variables for RL

Making deep learning causal is today still an open problem. The attempt to perform causal discovery and mechanisms fit at once as well as the problem of extracting latent causal variables from a raw space where features are not Markov related, could widen causal models real world applications.

10.2 Finding Better Rung 3 Models Having Agnostic Nonlinear Functions of Mechanisms

As introduced, SCMs model nonlinear mechanisms of each $< cause, effect >$ pair, but all partial contributions to the same effect get summed up. This means that even an "OR" statement ($\vee$) cannot be modeled if we lack explicit modeling of involved mediators. This couples with the previous point. A further addition is that SCM need a predefined structure, so a good agnostic generalization cannot afford to model the causal structure explicitly. A reasonable solution is to learn a latent SCM.

10.3 Bounded Unbiased Estimation Under Hidden Confounders

Recent studies are focusing on unbiased estimation even in presence of hidden confounders. They provide boundaries instead of point solutions and are another useful tool to improve the power of design of experiments.

11 Conclusion

Causality needs trustworthy knowledge to be reliable, and that knowledge has to come from somewhere, let it be data, or domain expertise, or previous experiments, but it can increase the span of experiments we can perform ethically as well as helping formalizing assumptions, by being used as a formal language for experiments, even for RCTs. These two aspects could help a lot soft sciences to find a way to be formal and run repeatable experiments, maybe surrogate but reliable since, once a mechanism gets identified with a valid set of its parents, and we mean Markov Parents, it is expected to be consistent across different studies, even out of specific laboratory controlled conditions.

References

[1] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Deep End-to-end Causal Inference, June 2022. arXiv:2202.02195.

[2] Jaron J. R. Lee, Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Ananke: A Python Package For Causal Inference Using Graphical Models, January 2023. arXiv:2301.11477.

[3] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2000.

[4] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs, September 2023. arXiv:1701.06686.

[5] Jacob Schreiber. Pomegranate: fast and flexible probabilistic modeling in python, February 2018. arXiv:1711.00137.

[6] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, New York, NY, USA, 2014.

A Primer on Probability / Graphs for Causality

Definition (Joint Probability Function)**.** *Given:*

- *a sample space* Ω
- *a* σ*-algebra of events* $\mathcal{F}$
- *a probability measure* P
- *a probability space* $(\Omega, \mathcal{F}, P)$ *on them*

For a set of random variables $X_1, X_2, \ldots, X_n$*, the joint probability distribution of these variables is defined as the probability of the event* $\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$*. The joint probability function is written as:*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}),$$

and represents the probability of the simultaneous occurrence of the events $X_1 = x_1$*,* $X_2 = x_2$*, ..., and* $X_n = x_n$*.*

Definition (Conditioning)**.** *Given:*

- *a sample space* Ω
- *a* σ*-algebra of events* $\mathcal{F}$ *and events* A, B *in it*
- *a probability measure* P
- *a probability space* $(\Omega, \mathcal{F}, P)$ *on them*

For $P(B) > 0$*, the conditional probability of* A *given* B *is defined as:*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Definition (Graph)**.** *A graph is a pair* $G = (V, E)$*, where* V *is a set whose elements are called vertices, and* E *is a set of pairs* $\{v_1, v_2\}$ *of vertices, whose elements are called edges.*

Definition (Acyclic Graph)**.** *An acyclic graph is a graph* $G = (V, E)$*, if it does not contain any cycles, meaning there is no path in the graph that starts and ends at the same vertex without repeating any edge.*

Definition (Directed Graph)**.** *A directed graph is a graph $G = (V, E)$, where E is a set of ordered pairs (v_1, v_2) of vertices, whose elements are called directed edges.*

Definition (Directed Acyclic Graph, DAG)**.** *A directed acyclic graph, DAG is a directed graph $G = (V, E)$, which is also an acyclic graph.*

Definition (ADMG)**.** *[4, sec 2.1] The acyclic directed mixed graph, ADMG of a graph $G(V \dot{\cup} L)$, is the graph $\mathcal{G}(V) = \sigma_L(G(V \dot{\cup} L))$ obtained by applying the latent projection operation to it.*

Definition (CADMG)**.** *[4, sec 2.2] A conditional ADMG, CADMG is an ADMG $\mathcal{G}(V, W)$ with $V \cap W = \emptyset$, V called the set of random vertices, W the set of fixed ones, and the constraint that $\forall w \in W, pa(w) = \emptyset \wedge bi(w) = \{w \leftrightarrow a, a \in V \cup W\} = \emptyset$, that is, fixed vertices have no parent or bidirected edges.*

Definition (Latent Projection)**.** *[4, sec A.3] Given a graph $G(V \dot{\cup} L$, so that V is the set of observable vertices and L the set of latent vertices, the latent projection operation of G on L, written as $\mathcal{G}(V) = \sigma_L(G(V \dot{\cup} L))$, is defined such that, for every $a, b \in V$:*

- *if there exists a directed path $a \to ... \to b$ and all nodes except a and b are in L, the edge $a \to b \in \mathcal{G}(V)$*
- *if there exists a path between a and b such that all nodes except a and b are non-colliders in L and both a and b have arrows pointing at them, the edge $a \leftrightarrow b \in \mathcal{G}(V)$*

B Agnostic PAC Learning

Definition (Empirical Risk Minimization)**.** *[6, sec. 2.2] Given:*

- *a population with distribution* $\mathcal{D} = \{< \mathcal{X}, \mathcal{Y} >\}$
- *a sample* S *drawn from* $\mathcal{D}$
- *a function* $\{ : \mathcal{X} \to \mathcal{Y}$
- *an Hp Class* $\mathcal{H}$ *so that* $h \in \mathcal{H}$ *and* $h_S : \mathcal{X} \to \mathcal{Y}$

The Empirical Error or Empirical Risk, on S *is defined as:*

$$L_s(h) = \frac{|\{i \in |S|, h(x_i) \neq y_i\}|}{|S|}$$

Definition (Uniform Convergence)**.** *[6, sec. 4.1] Given:*

- *a domain* Z *and a probability distribution* $\mathcal{D}$ *over* Z
- *a sample* S *drawn from* $\mathcal{D}$
- *an Hp Class* $\mathcal{H}$
- $\epsilon, \delta \in \{0, 1\}$

If always exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ *such that* $P(L_D(A(S)$ *and* $|S| \geq m_{\mathcal{H}}^{UC}$*, we say that* $\mathcal{H}$ *has the uniform convergence property on* Z*.*

Definition (Shattering and VC-dimension)**.** *[6, sec. 6.2] Given:*

- *a population with distribution* $\mathcal{D} = \{< \mathcal{X}, \mathcal{Y} >\}$
- *a sample* C *drawn from* $\mathcal{X}$
- *an Hp Class* $\mathcal{H}$

If $|\mathcal{H}| \geq 2^{|C|}$ *we say that* $\mathcal{H}$ *shatters* C*.*
The maximal C *that can be shattered, so that* $|\mathcal{H}| = 2^{|C|}$*, is the VC-dimension of* $\mathcal{H}$*.*

Theorem (Agnostic PAC Learning)**.** *[6, sec. 6.4] Given:*

- *an Hp Class* $\mathcal{H}$ *so that* $h \in \mathcal{H}$ *and* $h_S : \mathcal{X} \to \{\prime, \infty\}$
- $VCdim(\mathcal{H}) = d < \infty$

- *absolute constants* C_1, C_2

It always holds that $\mathcal{H}$ *has the uniform convergence property and is agnostic PAC learnable with sample size:*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

Theorem (No Free Lunch Theorem)**.** *[6, sec. 5.1] Given:*

- *a population with distribution* $\mathcal{D} = \{< \mathcal{X}, \{\prime, \infty\} >\}$
- *a sample* S *drawn from* $\mathcal{D}$ *such that* $S < |\mathcal{X}|/2$
- *a learning algorithm* A *over* $\mathcal{X}$

Exists a distribution $\mathcal{D}$ *where all the followings hold:*

1. *Exists* $f : \mathcal{X} \to \{0, 1\}$ *such that* $L_D(f) = 0$
2. $P(L_D(A(S) \geq 1/8) \geq 1/7$

C Rung 2 Definitions, Theorems and Properties

Definition (Conditional Independence)**.** *[3, sec. 1.1.5] Given:*

- *a finite set of variables* $V = \{V_1, V_2, \dots\}$
- *a joint probability function* $P(\cdot)$ *over* V
- *three disjoint subsets* X, Y, Z *in* V

The sets X *and* Y *are said to be conditionally independent, given* Z *if*

$$P(x \mid y, z) = P(x \mid z) \quad \textit{whenever } P(y, z) > 0$$

Property (Symmetry (of Conditional Independence))**.** *[3, sec. 1.1.5]*

$$(X \perp\!\!\!\perp Y \mid Z) \implies (Y \perp\!\!\!\perp X \mid Z)$$

Property (Decomposition (of Conditional Independence))**.** *[3, sec. 1.1.5]*

$$(X \perp\!\!\!\perp Y \cup W \mid Z) \implies (X \perp\!\!\!\perp Y \mid Z)$$

Property (Weak Union (of Conditional Independence))**.** *[3, sec. 1.1.5]*

$$(X \perp\!\!\!\perp Y \cup W \mid Z) \implies (X \perp\!\!\!\perp Y \mid Z \cup W)$$

Property (Contraction (of Conditional Independence))**.** *[3, sec. 1.1.5]*

$$(X \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp W \mid Z \cup Y) \implies (X \perp\!\!\!\perp Y \cup W \mid Z)$$

Property (Intersection (of Conditional Independence))**.** *[3, sec. 1.1.5]*

$$(X \perp\!\!\!\perp W \mid Z \cup Y) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \implies (X \perp\!\!\!\perp Y \cup W \mid Z)$$

Conditional independence tells us that, when we apply the chain rule of probability, we don't need to account for all of the variables that precede the ith one, but only the ones when we actually see an association. For these variables we indeed have a definition.

Definition (Markovian Parents)**.** *[3, sec. 1.2.2] Given:*

- *a finite and ordered set of variables* $V = \{X_1, \dots, X_n\}$

- *a joint probability function $P(\cdot)$ over V*

A set of variables pa_j is said to be the Markovian parents of X_j if pa_j is a minimal set of predecessors (with respect to the order) of X_j that renders X_j independent of all its other predecessors. In other words, pa_j is any subset of $\{X_1, \ldots, X_{j-1}\}$ such that

$$P(x_j \mid \mathrm{pa}_j) = P(x_j \mid x_1, \ldots, x_{j-1})$$

and no proper subset of pa_j satisfies C.

Note that so far the relationship is still associative, furthermore, more different orderings are allowed, coherently with our understanding of the inherent symmetry of association. We will now see better why and how to go further, but, for this purpose, we need to introduce graphs.

Definition (d-Separation)**.** *[3, sec. 1.2.3] Given:*

- *a graph G with X, Y disjoint subsets of the nodes of G*
- *a path p from a node in X to a node in Y*
- *a set of nodes Z in G*

p is said to be d-separated (or blocked) by Z if and only if one of the following cases occurs:

- *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z*
- *p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and no descendant of m is in Z.*

And a set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

So probability spaces model events and DAG do this too, joint probability functions have some events "screening off" the relationships of the ancestors to their children and directed graphs do this with d-separation.

Definition (Markov Compatibility)**.** *[3, sec. 1.2.2] Given:*

- *a DAG G*
- *a joint probability function P*

If P admits the factorization of equation C

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | pa_i)$$

relative to G, we equivalently say that:

- *G represents P*
- *G and P are compatible*
- *P is Markov relative to G*

Definition (Bayesian Network)**.** *[3, sec. 1.2.2] Given:*

- *a finite set of variables V*
- *a joint probability function $P(v)$ over V*
- *a DAG G*

G is said to be a Bayesian Network compatible with P if and only if $P(v)$ is Markov relative to G.

Property (Causal Markov Condition)**.** *[3, sec. 1.2.3] Given:*

- *a finite set of variables V*
- *a joint probability function $P(v)$ in V*
- *a DAG G*

$P(v)$ is said to have the Causal Markov Condition (or local Markov property), if it satisfies that every variable in V is independent of all its non descendants (in G), conditional on its parents.

The following theorem finally bridges the gap between d-Separation and Markov Compatibility.

Theorem. *[3, sec. 1.2.3] Given:*

- *a DAG G with three disjoint subsets of its nodes X, Y, Z*

If sets X and Y are d-separated by Z, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are not d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in at least one distribution compatible with G.

The theorem is unidirectional, from DAG to probability function. The opposite task is more complex, especially with purely observational data, since in that case, different DAG structures can be observationally equivalent, i.e. they are all possible, given the observed data.

Theorem. *[3, sec. 1.2.3] Given:*

- *a joint probability function P*
- *a DAG G*

A necessary and sufficient condition for P to be Markov relative to G is that, conditional on its parents in G, each variable be independent of all its predecessors in some ordering of the variables that agrees with the arrows of G.

So, stating that the probability function is Markov relative to the DAG is equivalent to saying that the order induced by the DAG is a valid one of C.

Theorem. *[3, sec. 1.2.3] Given:*

- *a joint probability function P*
- *a DAG G*

A necessary and sufficient condition for P to be Markov relative to G is for the Causal Markov Condition to hold.

Definition (Intervention)**.** *[3, sec. 1.3.1] Given:*

- *a sample space Ω*
- *a σ-algebra of events $\mathcal{F}$ and A in it*
- *a probability measure P*
- *a probability space $(\Omega, \mathcal{F}, P)$ over them*

For a random variable X and event $A \in \mathcal{F}$, the interventional probability using the do-operator, denoted $P(A \mid \mathrm{do}(X = x))$, is defined as the probability of A after intervening to set X to x, severing its natural causes in the causal graph. Formally:

- *$P(A)$: probability of A in $M_\varnothing$ the observational model*

- $P_X(A)$: *probability of* A *in* M_X *the model modified cutting incoming edges in* X

$$P(A \mid \text{do}(X = x)) = P_X(A \mid X = x).$$

Fixing a variable to a specific value, $do(X = x)$ is also called a hard-intervention. Generalizing, a soft intervention defines a mapping function, a policy, by which the variable takes a value, based on the other ones, $do(X = g(pa_x))$, with pa_x potentially different from the preinterventional model. Thus we can clearly see how the intervention operator generalizes the conditioning one, which can be indeed seen as a special case, having as policy the so called natural, behavioral or observational one.

Definition (Causal Bayesian Network)**.** *[3, sec. 1.3.1] Given:*

- *a finite set of variables* V
- *a joint probability function* $P(v)$ *over* V
- *a joint probability function* $P_X(v)$ *resulting from the intervention* $do(X = x)$ *that sets a subset* X *of variables in* V *to constants* x
- *the set* P_* *of all interventional distributions* $P_X(v)$, $X \subseteq V$, *including* $P(v)$, *which represents no intervention (i.e.,* $X = \emptyset$*)*

A DAG G *is said to be a causal Bayesian network compatible with* P_* *if and only if the following conditions hold for every* $P_X \in P_*$*:*

- $P_X(v)$ *is Markov relative to* G
- $P(v) = 1$ *for all* $V_i \in X$ *whenever* V_i *is consistent with* $X = x$
- $P(v_i|pa_i) = P(v_i|pa_i)$ *for all* $V_i \notin X$ *whenever* pa_i *is consistent with* $X = x$

In other words, our Bayesian Network is causal if our DAG faithfully represents the underlying process that is, the encoded information is causal as well.

D Rung 3 Definitions, Theorems and Properties

Definition (Counterfactual)**.** *[3, sec. 1.4.4] Given:*

- *a sample space* Ω
- *a* σ*-algebra of events* $\mathcal{F}$ *and an event* A *in it*
- *a probability measure* P
- *a probability space* $(\Omega, \mathcal{F}, P)$ *over them*
- *some evidence* E

For a random variable X*, an intervention* $X = x$*, and the event* A*, the counterfactual probability, denoted* $P(A_X \mid E)$*, is defined as the probability of* A *had* X *been* x*, given evidence* E*. Formally:*

- $P(A)$*: probability of* A *in* $M_\varnothing$ *the observational model*
- $P(A_X)$*: probability of* A *in* M_X *the model modified cutting incoming edges in* X

$$P(A_X \mid E) = P(A \mid \mathrm{do}(X = x), E).$$

Here we highlight that, if we fix a counterfactual that goes according to our evidence, we are in practice performing a standard intervention, so the counterfactual operator generalizes the interventional one and, thus, also the conditioning one, by performing no intervention at all and only leveraging evidence.

Definition (Structural Equation)**.** *[3, sec. 1.4.1] Given:*

- *a variable* X_i
- *a set of variables* $PA \subset V^2$ *such that* $pa_i \in PA$*, markovian parents of* X_i
- *a set of i.r.v.* U_i
- *a mapping function* f_i

The equation

$$x_i =: f_i(pa_i, u_i)$$

is said to be structural under the constraints that we only compute x_i *from it and keeping the markov parental order while doing so.*

Definition (Structural Causal Model). *[3, sec. 7.1.1] Given:*

- *a set of background variables U, called exogenous*
- *a set of variables V, called endogenous*
- *a set of variables $PA \subset V^2$ such that $pa_i \in PA$, markovian parents of V_i*
- *a set of functions F, so that $\forall f_i \in F, f_i : U_i \cup pa_i \to V_i$*

A causal model is a triple $M = \langle U, V, F \rangle$ where each f_i in

$$v_i =: f_i(pa_i, u_i), \ i = 1, ..., n$$

assigns a value to V_i that depends on(the values of) a selected set of variables in $V \cup U$, and the entire set F has a unique solution $V(u)$.

As for Bayesian Networks we kept information on the stable conditional probabilities, in SCM we keep functions and an estimate of the noise terms, for each modeled variable.

Definition (Causal World). *[3, sec. 7.1.1] Given:*

- *a causal model M*
- *a particular realization of the background variables u*

A causal world w is a pair $\langle M, u \rangle$.

Definition (Causal Theory). *[3, sec. 7.1.1] A causal theory is a set of causal worlds.*

Thus, an SCM is referred to a causal model as well as a theory. More specifically, a single individual, out of all the deterministic causal models possible, once we have fixed the exogenous unmodeled factors, is referred to as a world. A sample of worlds is referred to as a theory and so each theory is an improper subset of the whole causal model, which is a theory as well; the theory containing all of the possible models allowed by the unspecified exogenous factors.

Definition (semi-Markovian Model). *[3, sec. 1.4.2] A causal model associated with a graph G is said to be semi-Markovian if G is acyclic.*

Definition (Markovian Model). *[3, sec. 1.4.2] A causal model, associated with a graph G and with background variables U, is said to be Markovian if it is semi-Markovian and $U_i \perp\!\!\!\perp U_j$ for all $U_i, U_j \in U$ and $i \neq j$.*

Definition (Probabilistic Causal Model)**.** *[3, sec. 7.1.1] Given:*

- *a causal model M*
- *a set of background variables U*

A probabilistic causal model is a pair$\langle M, P(u)\rangle$ where $P(u)$ is a probability function defined over the domain of U.

Theorem. *[3, sec. 1.4.2] Given:*

- *a Markovian model M*
- *a DAG G associated with M*

It always induces a distribution $P(x_1, ..., x_n)$ that satisfies the causal Markov condition relative to G.

Theorem. *[3, sec. 5.2.1] Given:*

- *a Markovian model M*
- *a DAG G associated with M*
- *three disjoint subsets X, Y, Z in the nodes of G*

If X and Y are d-separated by Z in G, then X is independent of Y in M. Conversely, if X and Y are not d-separated by Z in G, then X and Y are dependent conditional on Z in M but for some degenerate cases.

E Do-Calculus / Fixing & Kernels

E.1 Do-Calculus

Property (Identifiability). *[3, sec. 3.2.4] Given:*

- *any computable quantity $Q(M)$ of a Markovian model M*
- *a class of models $\boldsymbol{M}$*

We say that Q is identifiable in $\boldsymbol{M}$ if, for any pairs of models M_1 and M_2 from $\boldsymbol{M}$, $Q(M_1) = Q(M_2)$ whenever $P_{M_1}(v) = P_{M_2}(v)$. If we have unobserved and can only estimate a partial set F_M of features (of $P_M(v)$), Q is identifiable from F_M if $Q(M_1) = Q(M_2)$ whenever $F_{M_1} = F_{M_2}$.

So whenever our estimate has identifiability, given the topology of our DAG, we can get an asymptotically unbiased estimate of that quantity from observations.

Definition (Causal Effect). *[3, sec. 3.2.1] Given:*

- *two disjoint sets of variables X, Y*

The causal effect of X on Y, denoted either as $P(y|\hat{x})$ or as $P(y|do(x))$, is a function from X to the space of probability distributions on Y. For each realization x of X, $P(y|\hat{x})$ gives the probability of $Y = y$ induced by deleting from the model of D, all structural equations corresponding to variables in X and defining them as $X = x$ in the remaining structural equations.

Property (Causal Effect Identifiability). *[3, sec. 3.2.4] Given:*

- *two disjoint sets of variables X, Y*
- *a DAG G associated to a Markovian model M*

The causal effect of X on Y is identifiable from G if the quantity $P(y|\hat{x})$ can be computed unambiguously, that is, if $P_{M_1}(y|\hat{x}) = P_{M_2}(y|\hat{x})$ for every pair of models M_1 and M_2 with $P_{M_1}(v) = P_{M_2}(v) > 0$ and $G(M_1) = G(M_2) = G$.

Definition (Back-Door Criterion). *[3, sec. 3.3.1] Given:*

- *a joint probability function P*
- *a DAG G markov relative to P*
- *a disjoint sets of variables X, Y in P*

- *a set of variables* Z *in* $P \setminus \{X, Y\}$

Z satisfies the back-door criterion for X_i, Y_j in G if these conditions are both satisfied:

- *no node in* X_i *is a descendant of* X_i
- Z *blocks every path between* X_i *and* Y_j *that contains incoming edges towards* X_i

If Z satisfies these conditions for every $X_i \in X$ and every $Y_i \in Y$, it is said to satisfy the back-door criterion for X, Y in G.

Definition (Front-Door Criterion)**.** *[3, sec. 3.3.2] Given:*

- *a joint probability function* P
- *a DAG* G *markov relative to* P
- *a pair of variables* X, Y *in* P
- *a set of variables* Z *in* $P \setminus \{X, Y\}$

Z satisfies the back-door criterion for X_i, Y_j in G if these conditions are all satisfied:

- Z *intercepts all direct paths from* X *to* Y
- *there is no unblocked back-door path from* X *to* Z
- X *blocks all back-door paths from* Z *to* Y

The following two theorems work for Markovian models only and reflect how statisticians have typically adjusted for confounders in the XX century, with arguably different results since they lacked part of the theory involved in doing such operations.

Theorem (Adjustment for Direct Causes)**.** *[3, sec. 3.2.3] Given:*

- *a set of variables* V
- *a set of variables* $PA \subset X^2$ *such that* $pa_i \in PA$*, markovian parents of* X_i
- *a set of variables disjoint of* $\{X_i \cup pa_i\}$*,* Y

The effect of the intervention $do(X_i = x_i')$ *on* Y *is given by*

$$P(y|\hat{x_i'}) = \sum_{pa_i} P(y|\hat{x_i'}, pa_i)P(pa_i)$$

where $P(y|\hat{x_i'}, pa_i)$ *and* $P(pa_i)$ *represent preintervention probabilities.*

Theorem. *[3, sec. 3.2.4] Given:*

- *a Markovian model* M
- *a DAG* G *associated with* M
- *a subset* V *of measured variables in* M
- *a set of variables* $pa \subset V$*, markovian parents of* X

The causal effect $P(y|\hat{x})$ *is identifiable whenever* $\{X \cup Y \cup pa_X \subseteq V\}$*, that is whenever* X*,* Y *and all parents of variables in* X *are in* V*. The expression for* $P(y|\hat{x})$ *is then obtained by adjusting for* pa_X*, as in E.1.*

Theorem (do-Calculus)**.** *[3, sec. 3.4.2] Given:*

- *an SCM* M *as in D*
- *a DAG* G *associated with* M
- *the joint probability function* $P(\cdot)$ *entailed by* M
- *three disjoint subsets of variables* X, Y, Z *in the variables of* M
- $G_{\overline{V_i}}$ *being* G *where incoming edges to* V_i *have been removed*
- $G_{\underline{V_i}}$ *being* G *where outgoing edges from* V_i *have been removed*
- $V(W)$ *being the subset of* V *not ancestors of the subset* W

These three rules always hold:

- ***Rule 1*** *(Insertion/deletion of observations):*

$$(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \implies P(y|\hat{x}, z, w) = P(y|\hat{x}, w)$$

- ***Rule 2*** *(Action/observation exchange):*

$$(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \implies P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w)$$

- ***Rule 3*** *(Insertion/deletion of actions):*

$$(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\overline{Z(W)}}} \implies P(y|\hat{x},\hat{z},w) = P(y|\hat{x},w)$$

where $Z(W)$ is the set of $Z-$ nodes that are not ancestors of any $W-$ node in $G_{\overline{X}}$.

do-calculus has been shown to be complete and, with no hidden confounders sound. The two following criteria are sound as well and can be derived by applying the do-calculus.

Theorem (Back-Door Adjustment)**.** *[3, sec. 3.3.1] If Z satisfies the back-door criterion relative to (X,Y), then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y|\hat{x}) = \sum_z P(y|x,z)P(z)$$

Theorem (Front-Door Adjustment)**.** *[3, sec. 3.3.2] If Z satisfies the front-door criterion relative to (X,Y) and if $P(x,z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x')$$

E.2 Fixing & Kernels

Definition (Kernel)**.** *[4, sec 2.3] Given*

- *a set of vertices V and a corresponding set of random variables X_v*
- *a domain $\mathcal{X}_W$*

A kernel is a function $q_V : x_W \to x_V$ written as $q_V(x_V|x_W)$, such that:

- *q_V is nonnegative*
- $$\sum_{x_V} q_V(x_V|x_W) = 1, \forall x_W \in \mathcal{X}_W$$

Definition (District)**.** *[4, sec 2.2] Given a CADMG $\mathcal{G}(V,W)$ and a vertex $v \in V$, the district of v, written as $dis_{\mathcal{G}}(v)$, is the maximal set made of all nodes bidirected connected to v.*

Definition (Markov Blanket). *[4, sec 2.8.2] Given a CADMG $\mathcal{G}(V, W)$ and a vertex $v \in V$, the Markov blanket of v in $\mathcal{G}$ is defined as:*

$$mb_{\mathcal{G}}(v) = pa_{\mathcal{G}}(dis_{\mathcal{G}}(v)) \cup (dis_{\mathcal{G}}(v) \setminus \{v\})$$

Definition (Fixable). *[4, sec 2.11] Given a CADMG $\mathcal{G}(V, W)$ and a vertex $v \in V$, v is fixable if there is no other vertex $v' \in V$ that is both a descendant of v and in its district. So the set of all fixable vertices is defined as:*

$$\mathbb{F}(\mathcal{G}) = \{v | de_{\mathcal{G}}(v) \cap dis_{\mathcal{G}}(v) = \{v\}\}$$

Definition (Fixing). *[4, sec 2.11] Given a CADMG $\mathcal{G}(V, W)$ and a kernel $q_V(x_V|x_W)$ associated to it, for every fixable vertex $r \in \mathbb{F}(\mathcal{G})$ the fixing operation is defined as:*

$$\phi_r(q_V(x_V|x_W), \mathcal{G}(V, W)) = (\frac{q_V(x_V|x_W)}{q_V(x_r|x_{mb_{\mathcal{G}}}(x_r))}, \mathcal{G}(V \setminus \{r\}, W \cup \{r\}))$$

Definition (Reachable). *[4, sec 2.13] Given an ADMG $\mathcal{G}(V \cup W)$ and a CADMG $\mathcal{G}'(V, W)$, $\mathcal{G}'$ is said to be reachable from $\mathcal{G}$ if exists a valid fixing sequence $\mathbf{w}$ of the vertices in W and $\mathcal{G}' = \phi_{\mathbf{W}}(\mathcal{G})$*

Definition (Intrinsic). *[4, sec 3.2] A district that is reachable in a graph $\mathcal{G}$ is said to be intrinsic. The set of all intrinsic sets of $\mathcal{G}$ is written as $\mathcal{I}(\mathcal{G})$.*

Theorem (Invariance of fixing ordering). *[4, sec 3.1] Given:*

- *an ADMG $\mathcal{G}(V)$*
- *a kernel $q_V(x_V|x_W)$ nested Markov to it, that is, corresponding to a $\mathcal{G}' \in \mathcal{I}(\mathcal{G})$*
- *$\mathbf{w_1}, \mathbf{w_2}$ both valid fixing sequences for the set $W \subseteq V$*

$$\phi_{\mathbf{w_1}}(q_V(x_V|x_W), \mathcal{G}(V, W)) = \phi_{\mathbf{w_2}}(q_V(x_V|x_W), \mathcal{G}(V, W))$$

Definition (Global Markov Property for CADMGs). *[4, sec 2.8.1] Given:*

- *a CADMG $\mathcal{G}(V, W)$ and a kernel $q_V(x_V|x_W)$ associated to it*
- *three disjoint sets of vertices $A, B, C \subseteq V$ where C can be empty*
- *$\mathcal{G}^{|W}$ being the subgraph of $\mathcal{G}$ after fixing W*

q_V has the global markov property for $\mathcal{G}(V, W)$ if:

$$\forall A, B, C,\ A\ m\text{-}separated\ from\ B\ given\ C\ in\ \mathcal{G}^{|W} \implies X_A \perp\!\!\!\perp X_B | X_C\ in\ q_V$$

Definition (Local Markov Property fot CADMGs)**.** *[4, sec 2.8.2] Given:*

- *a CADMG* $\mathcal{G}(V, W)$ *and a kernel* $q_V(x_V|x_W)$ *associated to it*
- *a vertex* $v \in V, ch_{\mathcal{G}}(v) = \emptyset$

q_V *has the local markov property for* $\mathcal{G}(V, W)$ *in* v *if:*

$$X_v \perp\!\!\!\perp X_{(V \cup W) \setminus (mb_{\mathcal{G}}(v) \cup \{v\})} | X_{mb_{\mathcal{G}}(v)} \text{ in } q_V$$

Theorem (ID Algorithm for NMMs)**.** *[4, sec 4.3] Given:*

- *a causal DAG* $G(V \cup L)$ *and its latent projection* $\mathcal{G}(V)$
- *two disjoint sets* $A \dot{\cup} Y \subseteq V$
- $Y^* = an_{\mathcal{G}(V)_{V \setminus A}}(Y)$ *ancestral graph of* Y *after intervening on* A
- $\mathcal{D}(\mathcal{G}(V)_{Y^*})$ *set of districts of* v *in the subgraph* $\mathcal{G}(V)_{Y^*}$

If and only if, $\mathcal{D}(\mathcal{G}(V)_{Y^*}) \subseteq \mathcal{I}(\mathcal{G})$*, the set is intrinsic in the observational graph, then* $p_Y(x_Y|do_{G(V \cup L}(x_A))$ *is identifiable and the following equation holds:*

$$p_Y(x_Y|do_{G(V \cup L}(x_A)) = \sum_{x_{Y^* \setminus Y}} \prod_{D \in \mathcal{D}(\mathcal{G}(V)_{Y^*})} \phi_{V \setminus D}(p_V(x_V), \mathcal{G}(V))$$

F Counterfactuals

Theorem. *[3, sec. 7.1.1] Given:*

- *a Probabilistic Markovian model* $\langle M, P(u)\rangle$

The conditional probability $P(Y_X|e)$ *of a counterfactual sentence "if it were* X *then* Y*," given evidence* e*, can be evaluated using the following three steps:*

1. ***Abduction*** *- Update* $P(u)$ *conditioned by the evidence* e *to obtain* $P(u|e)$*.*
2. ***Action*** *- Modify* M *by the action* $do(X)$*, where* X *is the antecedent of the counterfactual, to obtain the submodel* M_X*.*
3. ***Prediction*** *- Use the modified model* $\langle M_X, P(u|e)\rangle$ *to compute the probability of* Y*, the consequence of the counterfactual.*

Counterfactuals leverage evidence to prune our theory from all of the causal worlds that are not likely to have been happened. The purpose of the abduction step is this and the reason is that, when we make counterfactuals statements, we always assume a stability in time that is not sure to be the case. We are about to consider an alternative scenario, which we achieve in the action step, but we want it to occur in the "closest world" possible. The prediction step is self explanatory. One might ask what's the purpose of counterfactual reasoning.

Property (Composition (of Counterfactuals))**.** *[3, sec. 7.3.1] Given:*

- *a Markovian model* M
- *three disjoint sets* W, X, Y *in the variables of* M

We have

$$W_x(u) = w \implies Y_{xw}(u) = Y_x(u).$$

Property (Effectiveness (of Counterfactuals))**.** *[3, sec. 7.3.1] Given:*

- *a Markovian model* M
- *three disjoint sets* W, X *in the variables of* M

We have

$$X_{xw}(u) = x.$$

Property (Reversibility (of Counterfactuals))**.** *[3, sec. 7.3.1] Given:*

- *a Markovian model M*
- *a set X in the variables of M*
- *two variables W, Y in the variables of M*

We have

$$(Y_{xw}(u) = y) \wedge (W_{xy}(u) = w) \implies Y_x(u) = y.$$

Property (Recursiveness (of Counterfactuals))**.** *[3, sec. 7.3.1] Given:*

- *a Markovian model M*
- *two variables X, Y in the variables of M*

We have Let $X \to Y$ stands for the inequality $Y_{xw}(u) \neq Y_w(u)$ for some values of x, w, and u. M is recursive if, for any sequence $X_1, X_2, ..., X_k$, we have

$$X_1 \to X_2, X_2 \to X_3, ..., X_{k-1} \to X_k \implies X_k \nrightarrow X_1$$

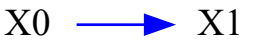
Theorem. *[3, sec. 7.3.1] Composition, effectiveness and reversibility are sound in structural model semantics; that is, they hold in all Markovian models.*

Theorem. *[3, sec. 7.3.1] Composition, effectiveness and reversibility are complete for all Markovian models.*

Theorem. *[3, sec. 7.3.1] Composition, effectiveness and recursiveness are complete for all Markovian models.*

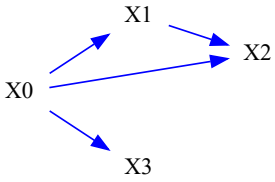
To clarify, reversibility is telling us that if we have a cycle but both variables determine each other consistently, it is legit to have it. The more appropriate interpretation is that, since they always have to occur together, they are in practice two faces of the same phenomenon, so we could even cut out one of them and merge the arrows of the other. We are in practice over-specifying a phenomenon, with two different, yet both valid thus consistent, variables. Recursiveness is instead telling us that, if we have cycles, their realizations become stationary and no positive feedback loop is created. So, to have guarantees, we would like our SCM to have a set of the properties defined earlier, such that we satisfy the theorems above.

G ADMGs of Obs Distributions

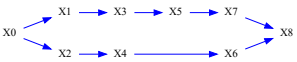


: Model 1

: Model 2



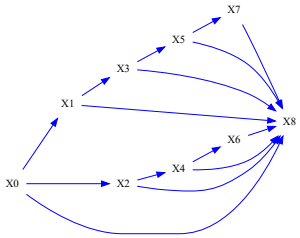
: Model 3



: Model 4

: Model 5

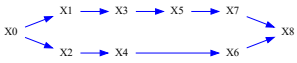
: Model 6

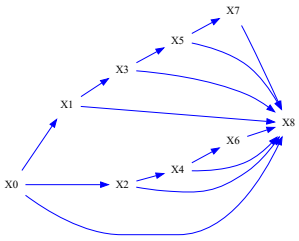


: Model 7

: Model 8

: Model 9

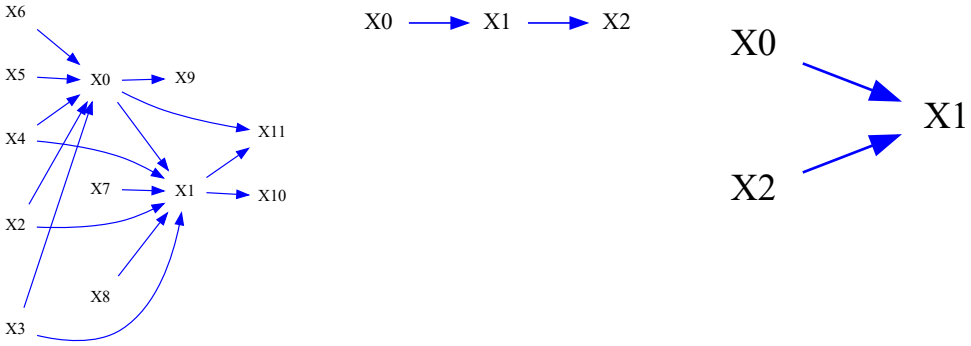




: Model 10

: Model 11

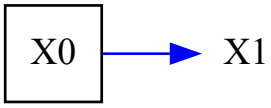
: Model 12

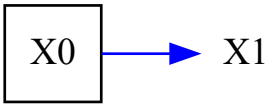


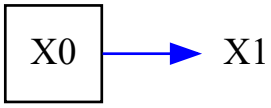
: Model 13 : Model 14 : Model 15

Figure 1: ADMGs of observational distributions.

H ADMGs of Surr Distributions



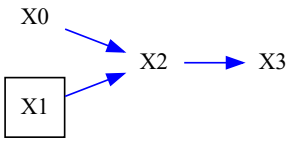


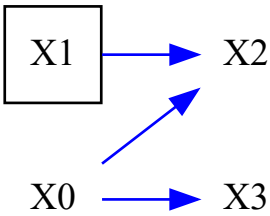


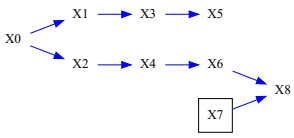
: Model 1

: Model 2

: Model 3



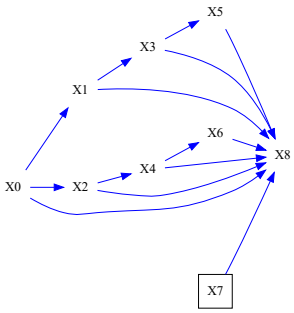




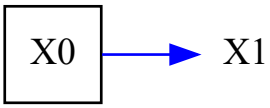
: Model 4

: Model 5

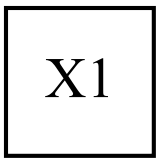
: Model 6



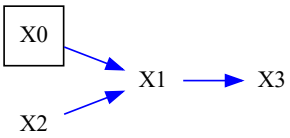
: Model 7



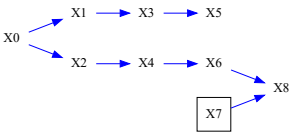
: Model 8



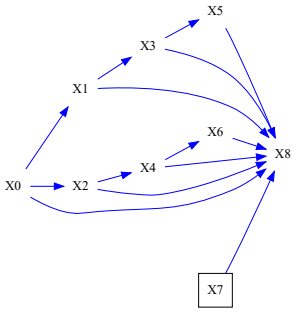
: Model 9



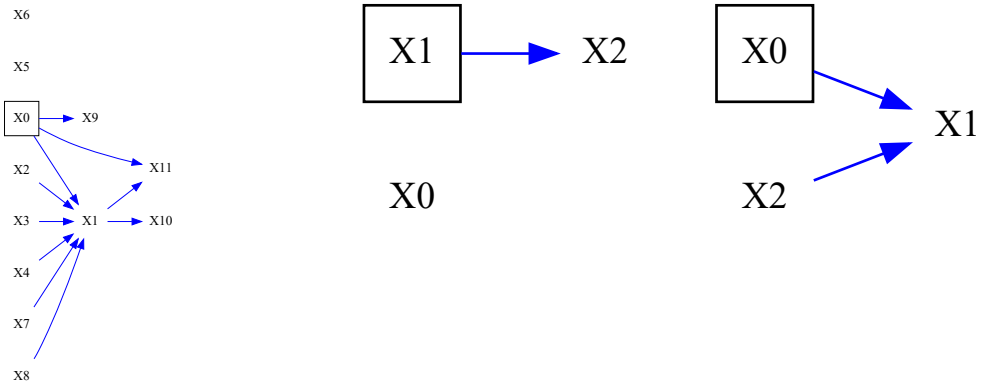
: Model 10



: Model 11



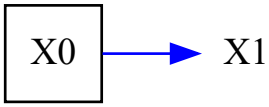
: Model 12

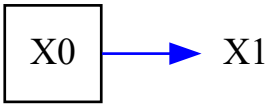


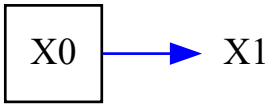
: Model 13 : Model 14 : Model 15

Figure 2: ADMGs of surrogate interventional distributions.

I ADMGs of Exp Distributions



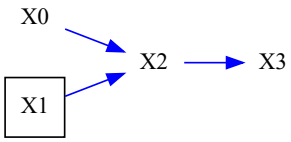


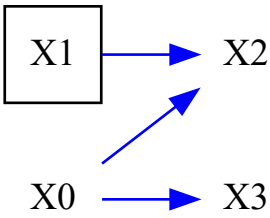


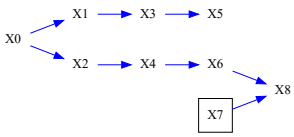
: Model 1

: Model 2

: Model 3



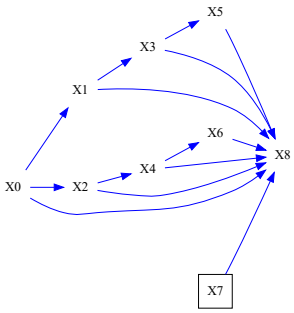




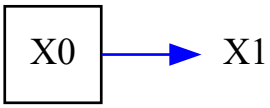
: Model 4

: Model 5

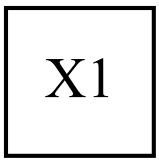
: Model 6



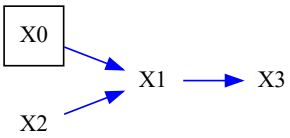
: Model 7



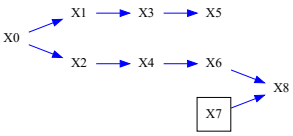
: Model 8



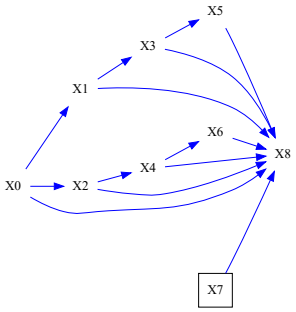
: Model 9



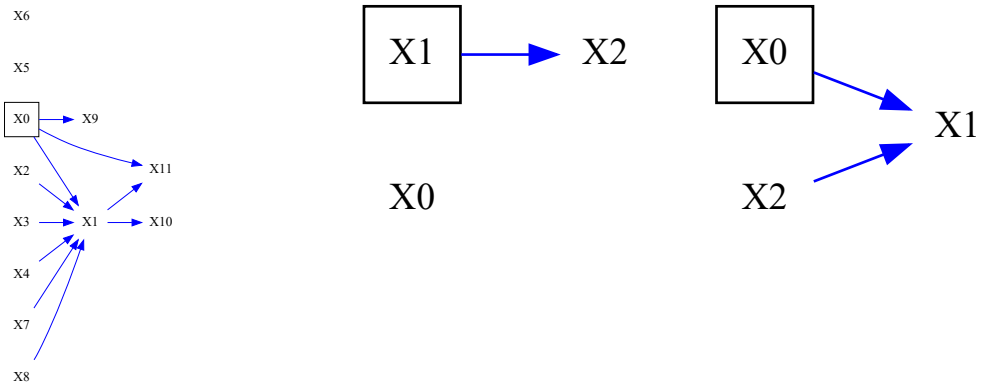
: Model 10



: Model 11



: Model 12



: Model 13 : Model 14 : Model 15

Figure 3: ADMGs of RCT interventional distributions.

J Tables of the Experiment

```
=== MODEL 1 : lingauss ===

Query: X0->X1
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-0.02204]	0.027505	0.759659	8.744198	10087
1	X0	BNsurr	[0.005465]	0.000000	0.759659	8.325579	9604
2	X0	Surr	[0.005465]	0.000000	0.759659	8.325579	9604
3	X0	Exp	[0.005465]	0.000000	0.759659	8.325579	9604

```
=== MODEL 2 : linexp ===

Query: X0->X1
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-0.049037]	0.054715	0.756558	8.600421	10085
1	X0	BNsurr	[0.005678]	0.000000	0.756558	8.190467	9605
2	X0	Surr	[0.005678]	0.000000	0.756558	8.190467	9605
3	X0	Exp	[0.005678]	0.000000	0.756558	8.190467	9605

```
=== MODEL 3 : nonlingauss ===

Query: X0->X1
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[1.35445]	0.069119	0.760907	8.248011	9453
1	X0	BNsurr	[1.423569]	0.000000	0.760907	8.380440	9604
2	X0	Surr	[1.423569]	0.000000	0.760907	8.380441	9604
3	X0	Exp	[1.423569]	0.000000	0.760907	8.380441	9604

```
=== MODEL 4 : nonlin_simpson ===

Query: X1->X2
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-0.484844]	0.675698	0.792519	17.795263	17330
1	X1	BNsurr	[-0.224345]	0.936197	0.792519	14.629851	14247
2	X1	Surr	[-0.744105]	0.416437	0.792519	17.283159	16831
3	X1	Exp	[-1.160542]	0.000000	0.792519	9.862323	9605

```
=== MODEL 5 : symprod_simpson ===

Query: X1->X2
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[0.635223]	0.098929	0.741050	8.685083	11064
1	X1	BNsurr	[0.32041]	0.413742	0.741050	8.171227	10409
2	X1	Surr	[0.766986]	0.032834	0.741050	9.159257	11668
3	X1	Exp	[0.734152]	0.000000	0.741050	7.539303	9604

```
=== MODEL 6 : large_backdoor ===

Query: X7->X8
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[0.190026]	0.102367	0.718239	9.847823	14216
1	X7	BNsurr	[0.223308]	0.069084	0.718239	9.879086	14262
2	X7	Surr	[0.71796]	0.425568	0.718239	7.232352	10441
3	X7	Exp	[0.292392]	0.000000	0.718239	6.652965	9605

```
=== MODEL 7 : weak_arrows ===

Query: X7->X8
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-0.652733]	0.000000	0.769686	8.773941	9605
1	X7	BNsurr	[-0.652733]	0.000000	0.769686	8.773942	9605
2	X7	Surr	[-0.652733]	0.000000	0.769686	8.773942	9605
3	X7	Exp	[-0.652733]	0.000000	0.769686	8.773942	9605

```
=== MODEL 8 : cat_to_cts ===
```

```
Query: X0->X1
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-0.318378]	0.326108	0.764024	8.842992	9970
1	X0	BNsurr	[0.00773]	0.000000	0.764024	8.518600	9605
2	X0	Surr	[0.00773]	0.000000	0.764024	8.518600	9605
3	X0	Exp	[0.00773]	0.000000	0.764024	8.518600	9605

```
=== MODEL 9 : cts_to_cat ===
```

```
Query: X1->X0
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[0.018936]	0.000000	0.757601	8.235737	9605
1	X1	BNsurr	[0.018936]	0.000000	0.757601	8.235737	9605
2	X1	Surr	[0.018936]	0.000000	0.757601	8.235737	9605
3	X1	Exp	[0.018936]	0.000000	0.757601	8.235737	9605

```
=== MODEL 10 : mixed_simpson ===
```

```
Query: X0->X1
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[1.370524]	1.330444	0.761476	12.747465	14566
1	X0	BNsurr	[0.751814]	0.711733	0.761476	13.551882	15485
2	X0	Surr	[2.012014]	1.971934	0.761476	11.136943	12725
3	X0	Exp	[0.04008]	0.000000	0.761476	8.405528	9604

```
=== MODEL 11 : large_backdoor_binary_t ===
```

```
Query: X7->X8
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-1.082203]	1.247377	0.714492	6.372758	9394
1	X7	BNsurr	[-1.082203]	1.247377	0.714492	6.372756	9394
2	X7	Surr	[-0.895875]	1.061049	0.714492	6.014855	8867
3	X7	Exp	[0.165174]	0.000000	0.714492	6.515229	9605

```
=== MODEL 12 : weak_arrows_binary_t ===
```

```
Query: X7->X8
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-0.778845]	0.000000	0.700325	6.013651	9605
1	X7	BNsurr	[-0.778845]	0.000000	0.700325	6.013651	9605
2	X7	Surr	[-0.778845]	0.000000	0.700325	6.013651	9605
3	X7	Exp	[-0.778845]	0.000000	0.700325	6.013651	9605

```
=== MODEL 13 : mixed_confounding ===
```

```
Query: X0->X1
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[-4.591528]	4.483054	0.928895	0.747067	386
1	X0	BNsurr	[-4.591526]	4.483052	0.928895	0.747080	386
2	X0	Surr	[-4.59195]	4.483477	0.928895	0.744541	385
3	X0	Exp	[-0.108474]	0.000000	0.928895	11.912068	6147

```
=== MODEL 14 : cat_chain ===
```

```
Query: X1->X0,X2
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[1.26725 0.734231]	0.562794	0.744097	5.900937	7395
1	X1	BNsurr	[1.26725 0.45175]	0.280312	0.744097	5.956745	7465
2	X1	Surr	[1.267249 0.430648]	0.259211	0.744097	6.018162	7542
3	X1	Exp	[1.26725 0.171437]	0.000000	0.744097	6.114711	7663

```
=== MODEL 15 : cat_collider ===
```

```
Query: X0->X1
```

	z-set	Type	Mean	MSE of Mean	Bin Size	Variance	Sample Size
0	∅	Obs	[0.631243]	0.272257	0.690677	5.867237	9905
1	X0	BNsurr	[1.308193]	0.404693	0.690677	5.155112	8703
2	X0	Surr	[0.555847]	0.347653	0.690677	5.736534	9685
3	X0	Exp	[0.9035]	0.000000	0.690677	5.689065	9605