Causal Discovery using Observations and Interventions

Gurpreet Singh
gs3056

Ritesh Baldva
rb3447

Abstract

In this project, we review a number of methods for structure discovery in causal models, particularly graphical causal models consistent with Pearl's notation of structural causal models [16, SCMs]. We formally define the problem statement for causal disvoery and present a broad category of papers which deal with this problem. We also explore the recent advances to better causal discovery using interventaional experiments. Finally, we propose an active causal learning algorithm that uses bayesian optimization to find the optimal intervention to maximize information gain. Our approach is an extension of the active learning algorithm proposed by Kügelgen et al. [12] which computes the information gain over all possible graphs. Our proposal is to sample the graphs using competitive graph sampling algorithms based on MCMC.

1 Introduction

Suppose if we have a thermometer showing the reading of the temperature of your bedroom. Let T be the random variable representing the actual temperature of the room (which is only observed through the thermometer) and let T' be the reading of the thermometer. We expect that $T = T'$ assuming the thermometer is not faulty. Now suppose you turn on a space heater which raises the overall temperature of the room, *i.e.*changes the value of T. We would expect the reading of the thermometer to change correspondingly as well. If we take a multitude of such samples, classical statistics would tell us that $T = T'$. However, suppose through manual intervention, someone changes the reading of your thermometer. This, obviously, would not affect the temperature of the room. However, classical statistics say that the temperature on the thermometer is supposed to be equal to the temperature of the room. Observations under manual interventions, therefore, are beyond the level of what statistics can deal with.

This is an example that Pearl [16] used to explain the three levels of causality. Causality is the field that deals with understanding how the data is generated rather than what data is generated. For the example above, if the statistician knew that the temperature of of the room causes the reading of the thermometer, she wouldn't mistakenly write the equation $T = T'$. In fact, Pearl [15, 16] introduced the notion of Structural Causal Models and Causal Graphs to tackle the idea of data generation from graphs.

Disocvering the way data is generated is, however, not an easy task and has many challenges. The most standard way to do this is to use domain knowledge to discover the causal structure. This is not always possible especially in fields like biology where we do not know everything about the effects in play. In this survey, we review some classical methods of causal discovery as well as discuss some of the state of the art methods of discovering the causal structure from observations alone. We also look at some of the recent advances on finding the causal structrue using data obtained from internventional experiments as opposed to only observations.

The remaining of the survey is structured as follows. In Section 2, we lay out the background work and discuss the concepts used throughout the survey. We also formally state the problem of Causal Discovery and some of the assumptions commonly used in most Causal Discovery methods, along with their implications. Then, we review some methods to Causal Discovery which try to find the causal structure using observations only in Section 3. In Section 4, we explore some methods to Causal discovery which are based on discovering the causal strucutre either using interventions alone, or using both observations and interventions. Lastly, we propose a method for Active Causal Discovery (defined in Section 2) which is based on finding an intervention at each timestep for optimal experimention and better discovery. We extend the approach described in [12], where instead of computing the utility over all graphs, we sample the graphs based on a well defined posterior using graph sampling techniques based on MCMC sampling methods [1].

2 Background

Causal discovery can be characterised as the degree to which one can rule out different plausible explanations for a causal effect. The way to rule out these competing explanations is addressed by the design of the study or experiment like random assignments or sampling bias and the set of assumptions. Before defining the problem of causal discovery, let's establish the standard terminology. We'll use Structural Causal Models (SCM) [16] as the framework for exploring the problem. An SCM, represented as a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \Pr(\mathbf{U}) \rangle$, is defined in terms of a *causal graph*, *structural equations* and a *probability distribution* over the exogenous variables $(\mathbf{U})$ in the graph which induces a distribution over the endogenous variables $(\mathbf{V})$.

Definition 1.1. Causal Graph [10]. *A causal graph $G = (V, E)$ represents casual effects between the variables, where V is the set of variables or nodes, and E is the set of directed edges containing directed edges $V_i \to V_j$, which represents a direct causal effect of V_i on V_j.*

We now define the terminology for the graphs. A *path* between X and Y in G, is a sequence of non-repeating directed edges, oriented in any fashion, starting from X and ending at Y. If every edge in the path points in the same direction, we have a *directed path*. Two nodes are *connected* in the graph if there exists a path between the two, otherwise they are *disconnected*. We make use of the standard relationship among the nodes, (e.g. *parents children*, *ancestors*, *descendants*). We can also classify the type nodes on a path. A node X_i is a *collider* on path P, if P contains $X_{i-1} \to X_i \leftarrow X_{i+1}$; it's a *mediator* if P contains $X_{i-1} \to X_i \to X_{i+1}$ (the other direction too) and it's a *common cause* if P contains $X_{i-1} \leftarrow X_i \to X_{i+1}$.

Definition 1.2. Structural Equations [16]. *Along with the causal graph, a set of equations of the form*

$$v_i = f_i(pa_i, u_i) \tag{1}$$

form the structural equations, where pa_i denotes the set of parent nodes of the variable v_i in the causal graph

and u_i denotes the "noise" or disturbances measured in observed variables.

Definition 1.3. Causal Discovery [6]. *Causal discovery is the inference task from the joint observational distribution $\Pr(\mathbf{V})$ to the causal graph G.*

One important feature that we can get from a probability distribution is the set of (conditional) dependences and (conditional) independences. Before analyzing the similarities between distributions and graphs we first establish the the independence relations that exist in the the graph, which is termed as *d-separation*.

Definition 1.4. d-separation [16]. *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, denoted as $\left(X \perp\!\!\!\perp Y \mid Z\right)_G$. This can happen if a node in Z is a mediator or common cause in the path p, or if no node in Z is a collider in the path and no descendants of the collider are in Z.*

The correspondence between the [in]dependences in graphs and [in]dependences in the distribution is thus established by the Markov condition and Faithful condition. Note, that the faithfulness assumption implies the Markov condition.

Definition 1.5. Faithfulness [10]. *Conditional independence between a pair of variables $X_i \perp X_j | Z$, can be estimated from the distribution if and only if Z d-separates X_i and X_j in the causal graph G.*

Definition 1.6. Markov Condition [16; 6]. *Every node X in the causal graph is probabilistically independent of all it's non-descendants given it's parents, denoted as $\left(X \perp\!\!\!\perp nd(X) \mid pa(X)\right)_P$.*

Definition 1.7. Markov Equivalence Class (MEC) [8]. *The set of causal graphs that have the same d-separation properties and thus, imply the same (conditional) independences are called Markov equivalent and belong to the same Markov Equivalence Class (MEC).*

Note that a brute force method comes out of the above to do causal discovery, where we generate all possible graphs from the set of variables and check if the corresponding independences hold, given the assumptions to find the MEC of the causal structure. We can see in Figure 2 of Eberhardt [6] the different equivalence classes for all three node DAGs. Note that, there might be scenarios where there exists only one structure, which leads to a simple solution directly. However, as the number of possible causal graphs within the equivalence class can grow exponentially with the number of vertices, it becomes infeasible to do so, especially with big-data where the number of features (nodes) is high.

Now, given the data for a problem, we have an *observational distribution* over the variables. Note, that from the above structural equations or the causal graph it's possible to transition to the interventional setting. From pearl's *do*-operator, setting the treatment accordingly in the equations, will give us the causal effect on the outcome. Similarly, we can modify the causal graph by removing the edges incoming into the treatment variable to see the remaining active causal pathways in G.The distribution of the outcome variable, obtained by intervening on the treatment is called the *interventional distribution* denoted as $\Pr(y|do(x))$. Note that it is possible to generate the interventional distribution from this new graph too, since the causal graph describe a stochastic way of generating the distribution.

Due to the above, we also have to take notice of what kind of data is being collected and analysed. In a lot of literature, another assumption of *causal sufficiency* is made. This is to make sure that the problem of *confounding bias* does not arise. However, note that this assumption is very strong and does not manifest in

many real life situations.

Definition 1.8. Causal Sufficiency [6]. *Causal Sufficiency implies that there are no unmeasured common causes among any pair of variables of SCM.*

Definition 1.9. Confounding Bias [10]. *A confounding bias exists for the causal effect $x \rightarrow y$ if and only if the observational conditional probability distribution is not always equivalent to the interventional distribution, i.e* $\Pr(y|x) \neq \Pr(y|\, do(x))$

Definition 1.10. Back-door Criterion [16]. *A set of variables Z satisfies backdoor criterion for an ordered pair of variables (X_i, X_j) in a DAG G if no node in Z is a descendant of X_i and Z blocks every path between X_i and X_j with arrow into X_i.*

The confounding bias could exist because of unmeasured common causes, since they open up a back-door path from the outcome to the treatment or there could be selection bias that is present in the data [2]. One way for causal identification, is to find the set which satisifies the back-door criterion, and do the adjustment by conditioning on that particuluar set.

Definition 1.11. Back-door Adjustment [16]. *If the set of variables Z satisfies the backdoor criterion relative to (X, Y), then we can identify the causal effect of X on Y as,*

$$\mathbb{P}\left[\, y \,\middle|\, do(X) \,\right] = \sum_{z} \mathbb{P}\left[\, y \,\middle|\, x, z \,\right] \mathbb{P}\left[\, z \,\right]$$

The above interventional notation will be useful in the context of active causal learning [11]. Tong et al [21] defines active learning where the learner is able to select instances based off on experiments to guide itself to more accurate models. For causal discovery, one way to think about that is to do interventional experiments.

Definition 1.12. Active Causal Discovery [11]. *Active Causal Discovery refers to first finding the Markov equivalence class from the observational distribution and then orienting the edges with the help of interventions, or by using both interventions and observations.*

3 Causal Discovery using Observations

3.1 Constraint Based Causal Discovery

Algorithms which use the conditional independence tests to figure out the Markov equivalence classes of causal structures fall under this category. Note that these algorithms might not always give the complete outputs with directionality on each edge, i.e a graph may be output with an undirected edge which in reality would correspond to a particular orientation.

3.1.1. SGS Algorithm

If we take note of all the four assumptions namely, Markov, faithfulness, DAG structure and causal sufficiency, the SGS algorithm [19] follows by using an elimination strategy. The algorithm works as follows,

1. Consider a complete undirected graph G, with edges among all the V variables.
2. For each set of conditional independences that hold in the distribution, $X \perp Y|Z$, remove

the edge connecting X and Y, since there can not be a direct causal effect.

3. For set of three variables X, Y, Z, figure out if Y is a collider by checking for the conditional dependence of X and Y given Z.

4. Recursively orient the remaining edges starting with the neighboring edges of the ones found in the above steps.

Note that the above scenario still has exponential runtime in worst case since the number of edges we need to run tests on is exponential. Also the output in case of non-singleton Markov Equivalence Classes will still be partially identified.

3.1.2. PC Algorithm

In PC algorithm [19] we optimize the edge elimination heuristic for step 2 that we were using before. After starting with the completely undirected graph,

1. Eliminate edges for which X and Y are unconditionally independent.

2. For each pair of variables X, Y that are connected with an edge, and for each variable Z that is still connected to either of them, eliminate the edge between X and Y if $X \perp Y|Z$.

3. For each pair of variables X, Y that are connected with an edge, and for each pair of variables T, Z that are connected to either of them, eliminate the edge between X and Y if $X \perp Y|\{T, Z\}$.

$\vdots$

After the final elimination, we proceed with finding colliders and performing edge orientation. This strategy thus avoid any unnecessary conditional independence tests that we might need to make and would still output the same result as the SGS algorithm. Note that, even more heuristics and wrappers for this PC algorithm have been designed in the literature. Another heuristic to speed up the edge elimination is that we only need to consider the nodes, that are adjacent to X and Y and not necessarily connected, because of the Markov property, which is referred to as the $PC^{\star}$ algorithm [19]. They also proposed the Fast Causal Inference algorithm which supports latent confounders under asymptotic correctness.

Also note that since most conditional independence tests assume gaussian or multinomial distributions, the above model is also restricted in sense of causal effects with arbitrary distributions. Among other variants, Colombo and Maathuis [4] noted that the output of the PC algorithm is order dependant and showed how it can lead to highly varied results in high-dimensional settings. In PC-Stable algorithm [4], the main difference lies in when the edges are deleted. For a step in the elimination heuristic of the original PC algorithm, we record which edge needs to be removed and remove it only when we proceed to the next level of sets to be conditioned on. Not only it makes the process order indpendent but also allows for each step to be parallelized, thereby improving runtime. The Parallel-PC algorithm [13] exploits

this caveat, but also keeps track that independence tests for a particular edge should always go to the same core. The authors were able to see an almost linear decrease in runtime with the number of cores when tested across different gene-expression datasets.

3.2 Scoring Based Algorithms

The major difference here is to use a fitting score like Bayesian Information Criterion (BIC) [**?**] instead of conditional independence tests used in the Constraint based algorithms.

$$BIC(\mathbf{X}, G') = \log P(\mathbf{X}|\hat{\mathbf{U}}, G') - \frac{M(G')}{2}\log(n) \tag{2}$$

where $M(G')$ represents the number of parameters estimated in the model, $\hat{\mathbf{U}}$ represents the MLE of the parameters, and n represents the sample size of the data, which would then ultimately pick the graph with the highest likelihood over the data. Note that process can also proceed in a Bayesian fashion, where we can define priors over the graph structure and use posteriors to get the scores. Note that the search happens over all possible graphs which is still infeasible.

3.2.1. Greedy Equivalence Search (GES)

GES [**?**] uses the above scoring criteria in the following two phase greedy fashion,

1. We start in an opposite fashion to the PC algorithm with a completely empty graph.
2. It then proceeds to add edges in the graph one by one in a greedy fashion, where BIC score is considered between the different models and choose the best one. The DAG is then mapped to the MEC before proceeding to adding the next edge.
3. Once all additions have taken place, the algorithms starts to delete the edges to arrive at the model with the maximum BIC score.

We can also use conjunction of Constraint-Based (CB) algorithms along with GES to come up with hybrid algorithms, where the skeleton of the graph is learned using the CB algorithms and orientation of edges can be decided in a greedy GES fashion.

3.3 Functional Causal Model Based Methods

While the above methods succeed in identifying the required MEC, they are held back by a lot of assumptions relating to the causal relation, namely they should be linear and have gaussian or multinomial parameterizations. In this subsection, we will weaken these assumptions and analyze causal discovery. These methods take advantage of the representational form as described in the Structural equations.

3.3.1. Linear Non-Gaussian Models

A linear causal model can be written in the following fashion,

$$x_i = \sum_{j \neq i} a_j x_j + u_i \tag{3}$$

where $u_i \sim$ non-Gaussian. As [6] explains, the above model forgoes assumptions not only about the gaussian parameterization, but also faithfulness. The author also lays out the argument for the two variable case, where the causal effect orientation is identifiable due to the Darmois-Skitovich theorem. It is noticed that if we *accidentally* assume the wrong direction for the causal effect and proceed to establish independences, we would easily be able to identify the mistake and correct the orientation. Such class of models are called LiNGAM.

Estimating such models from the observational data makes use of the Independent Component Analysis (ICA) algorithm. We can represent the above set of structural equations in the following form,

$$\mathbf{X} = \mathbf{A}\mathbf{X} + \mathbf{E} \tag{4}$$

$$\implies \mathbf{E} = (\mathbf{I} - \mathbf{A})\mathbf{X} \tag{5}$$

where, $\mathbf{E}$ represents the individual noise terms for each variable, which are assumed to be independent of each other and $\mathbf{X}$ represents the variables and $\mathbf{A}$ forms the matrix of coefficients.

Thus, we first use ICA to decompose the observational data matrix $\mathbf{D}$ as follows, where C represents the factor which has the independent components.

$$\mathbf{D} = \mathbf{C}\mathbf{X} \tag{6}$$

As laid out in [10], the goal to then estimate $\mathbf{A}$, then follows by first deriving an initial estimate as $\mathbf{I} - \mathbf{C}$. Then row permutations are applied on that to achieve a lower triangle matrix which represents the final estimate.

3.3.2. Non-Linear Additive Noise Models

The structural equations for non-linear additive noise models can be represented as follows [6],

$$x_j = f_j(pa(x_j)) + u_j \tag{7}$$

where $u_j \sim N(0, \sigma_j^2)$. The conditions for the above scenario have been established by Peters et al. [17]. They define a term called *causal minimality* and show the conditions on the observational distribution $\Pr(\mathbf{V})$ and on functions f_j to correctly identify the causal graph.

Definition 1.13. Causal Minimality. *The joint distribution induced by the SCM* $\Pr(\mathbf{V})$ *satisfies causal minimality if the functions in the structural equations* f_j *are not constant in any of their arguments, i.e*

$$\forall\, j\,,\, i \in pa(x_j)\ \exists\, \mathbf{x}_{pa_{x_j}\setminus\{i\}},\, x_i \neq x_i^{'} \quad f_j(\mathbf{x}_{pa_{x_j}\setminus\{i\}}, x_i) \neq f_j(\mathbf{x}_{pa_{x_j}\setminus\{i\}}, x_i^{'}) \tag{8}$$

Theorem 3.1. *Let* $\Pr(\mathbf{V})$ *be generated by a non-linear additive noise model, with a causal graph* G*, which satisfies the causal minimality condition, then* G *is identifiable from the joint observational distribution.*

Later, Mooji et al [20], also establish a model for orientation of the causal effect in a two variable model. It assumes that the latent variables are continuous along with the Markov condition but allows the model to remain non-parametric while still retaining identifiability. This is a pairwise model where $\Pr(\mathbf{X})\Pr(\mathbf{Y}|\mathbf{X})$ and $\Pr(\mathbf{Y})\Pr(\mathbf{X}|\mathbf{Y})$ are computed and the edge between X and Y is oriented corresponding to the better fit. For the directon $\mathbf{X} \to \mathbf{Y}$ the prior

on $\mathbf{X}$ is set as a mixture of gaussians and the functional mapping from $\mathbf{X}$ to $\mathbf{Y}$ is thus imposed a gaussian process prior.

In recent literature, we can use a graph neural network to represent the causal mechanism among the variables. Goudet et al. [9] propose the same, and use Causal Graph Neural Networks (CGNN) to identify the cause-effect relationship, identifying conditional independences and orienting edges within a graph. The causal mechanism in the form of 1 hidden-layer neural networks can be represented as follows,

$$x_i = \sum_{j=1}^{n_h} w_k^i \sigma \left(\sum_{k \in pa(x_i)} w_{kj}^i x_k + w_j^i u_i + b_j^i \right) + b^i \tag{9}$$

where n_h denote the number of hidden units, $\sigma(\cdot)$ is the activation function and $\left\{ w_k^i, w_{kj}^i, w_j^i, b_j^i, b^i \right\}$ represent the network parameters. The training of the networks follows from the observational distribution where the loss function used is based on Maximum Mean Discrepancy, which measures the distance between the means of two probability distributions, which would here be the observational distribution and distribution which the current parameters induce. They assume causal markov and faithfulness assumptions to hold and proceed with the structure identification based on scoring methods, by considering each edge in isolation and then use hill climbing approaches to optimize for the overall score.

Apart from the above methods, there are ways where you encode all possible information as constraints in propositional logic describing the underlying the causal graph structure. For e.g one way to describe a causal effect is by creating the respective literals as described in [6]. Thus, the causal discovery problem is transformed as an boolean satisfiability problem and translates the paradigm to a complete optimization based approach. SAT solvers can thus determine the corresponding solutions, and the graph can be reconstructed by using the truth assignments.

4 Causal Discovery using Interventions

When a human is tasked with understanding the reasoning behind an experiment, she firsts tries to build a reasoning model or a causal model using the observations available for the data. Based on future outcomes upon interventions and experiments, she continually learns and updates her causal model to reflect the new findings. This is the standard cognitive approach to understanding the world. This is usually a long process of experimentation (intervention) and updating the causal model.

Rottman and Keil [18] review multiple experiments which examine how people learn causal relationships. These experiments are based on active learning where each timestep entails sampling of multiple observations under an intervention. The authors claim that such experiments allow the participants to learn rather quicker than just showing differnt observations of the data. Active causal learning takes the same approach with the idea that observing data under intervention could allow us to learn more stable and better causal structures.

The motivation to use interventional experiments to learn data comes from the fact that it is often difficult to identify direction for all edges in a Markov Equivalence Graph for some observational data. Even though there are methods (discussed in Section 3) which estimate the direction, these are mostly based on functional

heuristics and can often infer incorrect edge directions. Allowing interventional observations, we can better identify the cause and effects and would be able to better identify the exact causal graph.

The task of learning from interventions, however, is non-trivial. Even though, we as humans can often very expertly perform such tasks, it is difficult to teach a machine to do the same. Consider the following example SCM $\mathcal{M}$ which hopefully demonstrates the complexity of the task.

Define an SCM $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}))$.

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{X, Y, Z\} \\ \mathbf{U} = \{U_X, U_Y, U_Z\} \\ \mathcal{F} = \begin{cases} X \leftarrow U_X \\ Z \leftarrow U_Z \\ Y \leftarrow \begin{cases} Z & \text{if } X = 0 \\ U_Y & \text{else} \end{cases} \end{cases} \\ P(U_X = 1) = P(U_Y = 1) = P(U_Z = 1) = 0.5 \end{cases}$$

where all observed and unobserved variables are binary *i.e.* $X, Y, Z, U_X, U_Y, U_Z \in \{0, 1\}$.

In the above SCM, it is clear that X causes Y. However, consider an intervention $do(X = 1)$. In this case, the observations would be such that $P(Y = 1 \,|\, do(X = 0)) = 0.5$. Therefore, based on this, we might say that X does not cause Y since an intervention on X did not change the distribution of Y. This is an incorrect implication and, therefore, we need to be careful when identifying effects on intervention. Rottman and Keil [18] state 3 rules which allow us to work on building a method for discovering causal structures using interventional data. These rules are stated as follows

1. A variable only changes state if it is directly intervened upon, or manipulated in some way, or if a variable that causes it (directly or indirectly) is manipulated. If it is not intervened upon, its state should remain stable.
2. If a variable X is intervened upon and Y changes state at the same time, then X causes Y. It is unlikely that Y happened to change due to some second unknown factor at the same time as X.
3. If a variable X is intervened upon and Y does not change state, then X probably does not cause Y. However, X may cause Y, but Y may not change after an intervention on X if Y is already at the state produced by an intervention on X.

Other than the challenge described above, there are additional statistical challenges that are not different from the challenges discussed in the previous section. To find independences between observational and interventional data we have to rely on statistical tests which can infer incorrect results if the data is not sufficient. Moreover, we need to make sure the order of interventions is optimal to ensure the most efficient scheme for causal discovery. We need to address these challenges while designing a scheme to perform active causal discovery.

Tong and Koller [21] first explored the idea of active learning for the efficient learning of bayesian networks. However, the difference between the objecitve described by Tong and Koller [21] is different from active

causal discovery since they were trying to infer the best parameters for the bayesian network which is fixed whereas we wish to find the best causal structure given the data.

In general, an active causal learning strategy would consist of the following broad steps at each timestep

1. Suppose if the set of causal models in consideration are $\mathcal{M}$.
2. Identify an optimal/random intervention for this timestep B.
3. Collect data D upon intervening consistent with B.
4. Update the set of models $\mathcal{M}$; or more generally the probabilities of each model in $\mathcal{M}$.
5. Repeat until needed.

For any algorithm, there are two steps to consider, finding the optimal intervention and updating the set of models based on an intervention. We formally define these steps with a probabilistic perspective as that allows us to present the idea more generally. Some of the equations and notations have been borrowed from Section 3 of [3] which describes in greater detail the objectives and some standard information theoretic approaches to active learning.

Choosing an Intervention Choosing the best intervention is important when each experiment is expensive and entails much more time or resources than choosing the best intervention itself.

The task of choosing an intervention, however, is non-trivial. Not only is choosing the best intervention a difficult and generally an intractable problem, it is also difficult to quantify the utility value itself. Different approaches use differnt notions of utility or gain in choosing an intervention (discussed in more detail later).

Suppose if we have a model (or SCM) $M \in \mathcal{M}$ which defines the causal graph and assumes a non-parameteric form of the functionals between different variables. Suppose if we define the set of interventions as $\mathcal{B}$, then we choose the best intervention b^* as

$$b^* \;=\; \underset{b \in \mathcal{B}}{\operatorname{argmax}} \underset{\mathbf{d} \in D_b}{\mathbb{E}} \left[\text{Gain} \left(M \,\middle|\, \mathbf{d}, b \right) \right] \tag{10}$$

where D_b represents the samples generated upon intervention b and the function Gain is the utility gain for a model given the sample $\mathbf{d}$ and intervention b.

Causal models inference We refer to the updating of the models set as inference. For any model $M \in \mathcal{M}$, we update the probability of M (posterior) as follows

$$\mathbb{P}\left[M \,\middle|\, D_b, b \right] \;\propto\; \mathbb{P}\left[D_b \,\middle|\, M, b \right] \mathbb{P}\left[M \right]$$

We use this idea to update the probabilities for the set $\mathcal{M}$ with the prior being the posterior computed from the previous timesteps. In case of non-bayesian approaches, the posterior over the set $\mathcal{M}$ is only non-zero for models which are part of the updated model set $\mathcal{M}$ and is uniform.

4.1 Constraint Based Active Causal Learning

Following the general algorithm for causal discovery, we now review some approaches to active causal discovery. One of the earlier attempts to discover causal structure using active learning was made in [11]. The main contribution by He and Geng [11] was to show that the learning of the causal structure

can be broken down into individually learning chain components which are cordal undirected graphs. We provide a summary of the main contributions in their work followed by some criticism of the opted approach.

Definition 1.14. Chain Component [11]. *Chain component of a graph is a connected undirected graph obtained by removing all directed edges from the chain graph.*

The chain graph is the PDAG which represents the Markov Equivalence Class of the causal graphs which can be obtained from the observations alone. As mentioned earlier, He and Geng [11] showed that finding the causal structure of the complete graph is equivalent to finding the causal structure of each chain component individually. However, an additional constraint over each chain component is that the undirected graphs are chordal [1]

Interventional Experiment and Updating the Model Class For any choice of intervention, one variable is chosen for a chain component and the intervention is performed for all values in its domain. These values are randomly assigned to experiments. For each intervention value, we have certain sample data points. The post-intervention independences are then estimated using these sample data points and used to reduce the equivalence class correspondingly.

Only the models that fit the updated equivalence class are considered for further experimentation. Note that this is done in a non-probabilistic fashion as we are simply selecting models which fit our updated beliefs.

Choosing the intervention Even though He and Geng [11] present different approaches for randomized experimentation as well quasi-experimentation, we exclude the latter from this review, however all criticisms we present are applicable to both.

The choice of intervention is made by maximizing entropy which is defined as follows. For some intervention $b \in \mathcal{B}$, the entropy is defined as

$$H_b = -\sum_{i=1}^{M} \frac{l_i}{L} \log \frac{l_i}{L}$$

where l_i denotes the number of possible DAGs of the chain component with the ith orientation among all possible orientations and $L = \sum_i l_i$.

This ensures that we choose the intervention which would allow us to minimize the variation between different orientations and, therefore, allow us to reduce the most number of models.

The original work shows detailed examples of the approach along with diagrams of causal graphs and we suggest the reader to go through them for a better understanding of the approach. Before we provide our criticisms, we describe another similar approach for choosing an optimal intervention proposed by Eberhardt [5]. The idea described in [5] only differs in how the optimal intervention is chosen. Instead of finding chain components, all maximal cliques in the graph are identified and the variable which is part of the most number of cliques is chosen for intervention. If there are multiple

[1] An undirected graph is chordal if every cycle of length larger than or equal to 4 has a chord

such variables, then any variable out of these can be chosen randomly or based on domain knowledge (for example we would pick the variable which is easier to manipulate).

4.1.1. Criticism

The first thing to notice is that both the approaches described above are based on constricting the Markov Equivalence Class obtained from a standard causal discovery algorithm such as the PC Algorithm. Despite this being a classic approach, the PC algorithm can be often inefficient at finding skeletons for causal structures. Even though the PC algorithm is complete, given a limited number of samples, the independences obtained from the observations, especially heavily conditioned independences, are innacurate and therefore the MEC obtained is not suitable for the causal structure.

Moreover, both the experiments require interventional experiments to run for all possible domain values in order to identify the independences. This requires a large number of samples to be obtained for each timestep and therefore is not suitable if the number of domain values for any variable is large or if the variable is continuous.

For the approach described in [11], the entropy is computed over all orientations and therefore requires us to iterate over all possible graphs. This can be exponential and therefore extremely slow for even small graphs or chain components. Similarly for the approach in [5], finding the maximal cliques itself is a NP-hard problem and therefore simply choosing an intervention is non-trivial in both the approaches. In the next subsection, we explore another recently proposed strategy for choosing an intervention which addresses some of these challenges by considering a complete bayesian setting.

4.2 Optimal Intervention with Gaussian Process Networks

As pointed out earlier, a constraint based aproach can often be unsuitable for active causal learning. Kügelgen et al. [12] propose an approach based on finding the optimal intervention for a functional model with each graphical equation modeled using a Gaussian Process. The relationship between each variable X_i and it's causal parents $pa(X_i)$ is represented as an additive noise model (ANM) and is given as follows

$$X_i \;=\; f_i(pa(X_i)) + \epsilon_i \tag{11}$$

The above suggests that we are implicitly assuming each variable X_i to be continuous. This is one of the caveats of this approach which is also assumed in our proposed method. We leave the extension to discrete variables as a future step.

Assuming a gaussian process network entails that the functions f_i are sampled from a gaussian process. This allows us let the functions be non-parameteric while retaining the nice properties of Gassian processes and, therefore, affording closed form marginals.

Suppose for a model $M \in \mathcal{M}$, we represent $\boldsymbol{\theta}_M$ as the functional parameters for the corresponding graph structure. For a gaussian process network, these parameters describe the functions between the causal nodes.

The marginal likelihood probability of the data is then defined as follows

$$\mathbb{P}\left[\, D \,\middle|\, M \,\right] \;=\; \int_{\boldsymbol{\theta}_M} \mathbb{P}\left[\, D \,\middle|\, \boldsymbol{\theta}_M, M \,\right] \mathbb{P}\left[\, \boldsymbol{\theta}_M \,\middle|\, M \,\right] \, \mathrm{d}\boldsymbol{\theta}_M \tag{12}$$

The corresponding posterior over the models in the model class $\mathcal{M}$ is defined as follows

$$\mathbb{P}\left[\, M \,\middle|\, D \,\right] \;\propto\; \mathbb{P}\left[\, M \,\right] \mathbb{P}\left[\, D \,\middle|\, G \,\right] \tag{13}$$

The advantage of defining the posterior in this manner allows us to leverage even limited samples of data points without the need to find independence constraints which typically require large amounts of data. This idea is similar to score based methods, such as GES which uses the bayesian information criterion to compute the score. The idea, therefore, is to set a higher probability value for causal graphs which on an average explain the observations better. Even though this is not explicitly defined in the work by Kügelgen et al. [12], we later on extend this to consider interventional data as well.

The main consideration and contribution of the paper is the proposal of a novel strategy to choose an optimal intervention for continuous variables.

Suppose if we intervene on the variable V as $do(V = v)$. The measure of an intervention is described by the information theoretic measure of information gain U defined as follows (we represent $\mathbb{P}\left[\, \cdot \,\middle|\, do(V = v) \,\right]$ as $\underset{V=v}{\mathbb{P}}\left[\, \cdot \,\right]$)

$$U(D_v, do(V = v)) \;=\; \sum_{M \in \mathcal{M}} \underset{V=v}{\mathbb{P}}\left[\, G \,\middle|\, D_v \,\right] \log \underset{V=v}{\mathbb{P}}\left[\, M \,\middle|\, D_v \,\right] - \sum_{M \in \mathcal{M}} \mathbb{P}\left[\, M \,\right] \log \mathbb{P}\left[\, M \,\right] \tag{14}$$

where D_v represents the data points sampled from the actual intervention distribution. Since the second term in the above equation is constant with respect to the intervention variable, we can disregard it for the maximization problem defined later.

The utility in Equation 14 corresponds to the expected change in entropy for the posterior over the causal models given samples from the intervention distribution. Now we define the maximization problem to choose the optimal intervention. Suppose the set $\mathcal{V}$ represents the set of variables we can intervene on and let $\mathcal{D}_V$ represent the domain for the variable V. Then, the intervention is chosen based on the following optimization objective

$$V^*, v^* \;=\; \underset{V \in \mathcal{V}, v \in \mathcal{D}_V}{\operatorname{argmax}} \int_{\mathbf{v}_{-V}} U(\mathbf{v}_{-V}, do(V = v)) \underset{V=v}{\mathbb{P}}\left[\, \mathbf{v}_{-V} \,\right] \, \mathrm{d}\mathbf{v}_{-V} \tag{15}$$

The above objective is based on Bayesian experiment design which is a decision theoretic approach for selecting and experiment aiming to maximize a utility function.

Upon combining Equations 14 and 15, the optimization objective decomposes into the following form

$$V^*, v^* \;=\; \underset{V \in \mathcal{V}, v \in \mathcal{D}_V}{\operatorname{argmax}} \sum_{M \in \mathcal{M}} \mathbb{P}\left[\, M \,\right] \int_{\mathbf{v}_{-V}} \underset{V=v}{\mathbb{P}}\left[\, \mathbf{v}_{-V} \,\middle|\, M \,\right] \log \underset{V=v}{\mathbb{P}}\left[\, M \,\middle|\, \mathbf{v}_{-V} \,\right] \, \mathrm{d}\mathbf{v}_{-V} \tag{16}$$

The integral can be computed using Monte Carlo estimation since sampling from the marginal of a gaussian process network is not difficult [12] as a closed form solution is available. The probability within the logarithm can be computed using Equation 13 by summing over all models in $\mathcal{M}$.

Since the optimization objecitve can be highly non-convex, Bayesian Optimization is a suitable strategy to find an optimal intervention.

Even though this approach is highly desirable since it allows us to choose an optimal intervention for continuous random variables without assuming parameteric functions to model causal effects. However, this approach has its own caveats.

Firstly, the summation over all graphs is non-optimal as the number of directed acyclic graphs for a set of variabes is super-exponential and therefore we would only be able to compute the information gain and utility for a very small number of variables. Morevoer, this summation is not only done once but for every sample from the causal model, the posterior needs to be computed which would mean that the complexity (assuming everything else to constant) is $\mathcal{O}\left(N \,|\, \mathcal{M} \,|^2\right)$ where N is the number of Monte Carlo samples used to estimate the integral and $|\, \mathcal{M} \,|$ is the size of the model class. This is explicitly stated by the authors as a limitation as well.

Secondly, the paper does not inform on how to update the graph prior as more intervention data is available. We propose to extend the paper in order to address these shortcomings based on graph sampling techniques and bayesian information criterion.

In the next subsection, we briefly look at techniques which can help us avoid summing over the complete model class and instead sample from a well defined posterior which is based on the observations available with and without interventions. This would allow us to scale up the approach to relatively larger number of variables by computing an estimate of the information gain for each variable and intervention.

4.3 Active Causal Structure Learning for Gaussian Process Networks

As discussed earlier, choosing an intervention by summing the information gain over all graphs is nonoptimal as the number of graphs can be super-exponential. Even though it is possible to reduce the search space by limiting the graphs to a smaller class such as the MEC derived using PC algorithm. This could be a feasible approach if we have large amounts of observations.

Kügelgen et al. [12] suggest using the approach described in [1] to sample graphs. The idea in this sampling approach is to sample orders (variable orders) instead of graphs as it offers better mixing of the MCMC chain [7]. This is, however, not optimal for active causal learning. This is because the approach described in [1] is based on independences obtained from observations. As described earlier, this is not suitable for interventions, especially with continuous variables as it is difficult to attribute independences from interventional observations.

Instead we focus on MCMC sampling based on the previous approach described in [7]. The original graph sampling via MCMC was proposed by Madigan et al. [14]. However the posteriors can be highly peaked at multiple local optimals and therefore the chain mixing is slow [7]. In an attempt to remedy this, Friedman and Koller [7] proposed a sampling approach based on topological orders of the nodes rather than complete graphs. They reported faster mixing and therefore better chain convergence for this case. For each order, the posterior can be decomposed over the variables [2].

We skip the details mentioned in the original work as it is straightforward to extend this approach for

[2] This assumes the Markov Condition to be satisfied

our use case. We now formally define our approach in Algorithm 1.

Algorithm 1: Active Causal Discovery with Optimal Intervention Selection

Input : Observation data D_O, Oracle $\mathcal{O}$ to generate intervention samples

Procedure :

1. Sample some graphs from the approach described in [7] using the data D_0 as $\mathcal{M}_0$.
2. For timestep $t = 0 \ldots$, do
   - Find optimal intervention as follows

$$\underset{V \in \mathcal{V}, v \in \mathcal{D}_V}{\operatorname{argmax}} \sum_{M \in \mathcal{M}_t} \int_{\mathbf{v}_{-V}} \mathop{\mathbb{P}}_{V=v} \left[\mathbf{v}_{-V} \,\middle|\, M \right] \log \mathop{\mathbb{P}}_{V=v} \left[M \,\middle|\, \mathbf{v}_{-V}, D_{0:t} \right] \, \mathrm{d}\mathbf{v}_{-V} \quad (17)$$

   where $D_{0:t}$ represents the data generated in all previous timesteps. The log probability can be estimated using the same graph samples, however, admittedly that would add to the bias of the estimate.
   - Sample data for the intervention as D_{t+1}
   - Using the data $D_{0:t+1}$, sample models $\mathcal{M}_{t+1}$ using the MCMC approach

There are still some things which are not optimal about the algorithm. The first is that for each timestep and for each sampling step, the probability over the complete data distribution is needed to be computed which could be potentially slow. Since we are only sampling an order and marginalizing over the actual graph, we still have an overhead of exponential steps, although this is much more controlled and can be potentially reduced using domain knowledge [7].

Using the same approach, it is not trivial to include discrete variables. We hope to look into this in more detail in the future.

Another extension that could be viable is to look into chain components instead of the complete graph. That would allow us to run simultaneous interventions and potentially reduce the number of timesteps as well as improve convergence. Multiple interventions is not possible at the moment since we are using Bayesian Optimization over a single objective. Looking at chain components would also allow us to run the optimization strategy independently and parallelly over the chain components to find optimal interventions for each chain component.

One major drawback of all of the active learning methods we discussed is that none of these suppose latent variables. This makes the approach much less likely to be applicable in widespread domains since ubobserved confounders are very common in real world experiments. Another challenge would be to address missing data. We can look into marginalizing the data over missing variables however given the high complexity of the approach already, this might not be optimal.

The above mentioned challenges are some of the problems we hope to look into in the future and further extend the approach to wider applications.

References

[1] Raj Agrawal, Tamara Broderick, and Caroline Uhler. Minimal i-map mcmc for scalable structure discovery in causal dag models, 2018.

[2] Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[3] Neil Bramley. Constructing the world: Active causal learning in cognition. In ., 02 2017.

[4] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[5] Frederick Eberhardt. Almost optimal intervention sets for causal discovery. *CoRR*, abs/1206.3250, 2012. URL `http://arxiv.org/abs/1206.3250`.

[6] Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, 2017.

[7] Nir Friedman and Daphne Koller. Being bayesian about network structure: A bayesian approach to structure discovery in bayesian networks. *Mach Learn*, 50, 07 2001. doi: 10.1023/A:1020249912095.

[8] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL `https://www.frontiersin.org/article/10.3389/fgene.2019.00524`.

[9] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018.

[10] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2018.

[11] Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

[12] Julius Kügelgen, Paul K Rubenstein, Bernhard Schölkopf, and Adrian Weller. Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks, 2019.

[13] Thuc Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM transactions on computational biology and bioinformatics*, 2016.

[14] David Madigan, Jeremy York, and Denis Allard. Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique*, 63(2):215–232, 1995. ISSN 03067734, 17515823. URL `http://www.jstor.org/stable/1403615`.

[15] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444. URL http://www.jstor.org/stable/2337329.

[16] Judea Pearl. *Causality*. Cambridge University Press, 2000. doi: 10.1017/CBO9780511803161.

[17] Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal Discovery with Continuous Additive Noise Models. *arXiv e-prints*, art. arXiv:1309.6779, September 2013.

[18] Benjamin M. Rottman and Frank C. Keil. Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, 64(1):93 – 125, 2012. ISSN 0010-0285. doi: https://doi.org/10.1016/j.cogpsych.2011.10.003. URL http://www.sciencedirect.com/science/article/pii/S0010028511000879.

[19] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[20] Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in neural information processing systems*, pages 1687–1695, 2010.

[21] Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In *Advances in neural information processing systems*, pages 647–653, 2001.