Targeted Causal Elicitation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We look at the problem of learning causal structure for a fixed downstream causal effect optimization task. In contrast to previous work which often focuses on running interventional experiments, we consider an often overlooked source of information - a domain expert. In the Bayesian setting this amounts to augmenting the likelihood with a user model whose parameters account for possible biases of the expert. Such a model allows for active elicitation in a manner that is most informative to the optimization task at hand.

1 Introduction

Causal models are an essential component in decision making, providing a clear line between which types of queries we can answer under different assumptions [Pearl and Mackenzie, 2018]. In this work we focus on causal graphs, specifically Directed Acyclic Graphs (DAGs), which are required to reason about causal effects arising from interventions. Learning this graph from data, a problem also referred to as causal discovery or causal structure learning, is a challenging task [Glymour et al., 2019]. Even in the limit of infinite observational data, algorithms such as the PC algorithm Spirtes et al. [2000] and GES algorithm Chickering [2002] can provably recover only an equivalence class of graphs [Verma and Pearl, 1990] called the Markov Equivalence Class (MEC) which contain DAGs that encode the same conditional independence relations. While we could in principle generate all DAGs in an MEC, some MECs could be extremely large. For example, in the sparse setting this number is super-exponential in the size of the nodes [He et al., 2015].

Methods that further refine the causal structure rely on perturbing the system and gathering the resulting interventional data. The selection of which variables to intervene on and which values to set them to is an active area of research Agrawal et al. [2019], Gamella and Heinze-Deml [2020].

While learning the full causal graph may be useful to qualitatively analyse the system, it may be too comprehensive, especially considering one of the main reasons we would want the causal graph in the first place - identifying causal effects. A practitioner may only need to know a subset of the graph to select a cost-effective intervention to regulate a specific target X_t variable of interest. For example, in the medical setting, a doctor may wish to know which treatment $do(\boldsymbol{X_I^*} = \boldsymbol{x_I^*})$ to give to a patient in order to maximize the probability of their LDL cholestrol being lower than 3.0 mmol/l, $\mathbb{P}(X_t < 3.0 \mid do(\boldsymbol{X_I^*} = \boldsymbol{x_I^*}))$ while minimizing the cost of this treatment.

2 Problem setting

Let $\mathcal{M}^*$ be the true Structural Causal Model (SCM) with observable variables $\boldsymbol{X} = \{X_1, \cdots, X_{n-1}, X_t\}$ and X_t being the target variable which we wish to regulate, i.e. be within a certain range. Collections of variables and values are in bold and indexed by a set $I \subseteq \{1, \cdots, n-1, t\}$. We assume $\mathcal{M}^*$ is such that the associated causal graph $\mathcal{G}^*$ is a DAG. We also assume no hidden

Working paper. Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

confounders, and that the functional dependencies between the variables have an additive noise structure with independent noise variables ϵ_i:

$$X_i = f_i(\boldsymbol{PA}_i^{\mathcal{G}^*}, \boldsymbol{\theta}_i) + \epsilon_i \quad i \in \{1, \cdots, n-1, t\} \tag{1}$$

Here, each f_i is a deterministic function which we assume is parametrized by $\boldsymbol{\theta}_i$ and depends on the parents of X_i in $\mathcal{G}^*$, denoted by $\boldsymbol{PA}_i^{\mathcal{G}^*}$. When X_i has no parents, we set $f_i = 0$. We denote the binary adjacency matrix of $\mathcal{G}$ by $\boldsymbol{G}$.

We assume that we start with N samples of observational data $\boldsymbol{x}_k \in \mathbb{R}^n$, and assume a budget of M expert feedback samples $\boldsymbol{e}_k^{ij}$ whose dimensionality depends on the expert model. Here $\boldsymbol{e}_k^{ij}$ refers to the kth feedback sample about the relation between variables X_i and X_j. Importantly, we assume it is expensive/prohibitive to gather interventional data.

The overall goal is to query the expert in order to find the interventions $do(\boldsymbol{X_I^*} = \boldsymbol{x_I^*})$ which satisfy both of the following criteria:

- We maximize the probability of the target variable X_t being in some region $\mathcal{T}$
- We minimize the cost of intervention.

Note that this intervention is not performed to narrow down the causal structure. It is the downstream intervention meant to be performed, or guide the expert given the SCM. The resulting optimization problem is given below, and is related to Causal Global Optimization (CGO) as defined in Aglietti et al. [2020].

$$\boldsymbol{X_I^*}, \boldsymbol{x_I^*} = \text{argmax}_{\boldsymbol{X_I} \in \mathcal{P}(\boldsymbol{X} \setminus X_t), \boldsymbol{x_I} \in D(\boldsymbol{X}_{\boldsymbol{I} \setminus t})} \frac{\mathbb{P}(X_t \in \mathcal{T} | do(\boldsymbol{X_I} = \boldsymbol{x_I}))}{\text{Cost}(do(\boldsymbol{X_I} = \boldsymbol{x_I}))} \tag{2}$$

Here, $\mathcal{P}(\boldsymbol{X})$ refers to the power set of $\boldsymbol{X}$ and $D(\boldsymbol{X_I})$ refers to the domain of values taken by the variables $\boldsymbol{X_I}$.

Since we do not know the true SCM $\mathcal{M}^*$ and it is required to compute causal effects, we must integrate it out in the above objective.

3 Methodology

The general model we consider is shown in Figure 1.

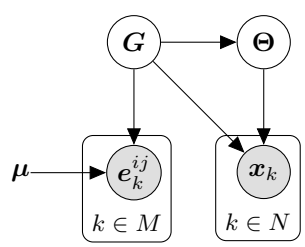

Figure 1: Augmented model including an expert likelihood term which consists of information about the graph structure.

Here, $\boldsymbol{\Theta}$ represents all parameters of the SCM functions. We consider feedback from the expert to be observed data and hence the resulting model incorporates an extra likelihood term which we call the expert model.

Expert model

We assume the expert feedback is i.i.d. The resulting expert model $\mathbb{P}(\boldsymbol{e}_k^{ij} | \boldsymbol{G}_{ij})$ is a likelihood for edge information of $\mathcal{G}$, specifically between variables X_i and X_j. The hyperparameter $\boldsymbol{\mu}$ models the

reliability of the expert, which we assume is fixed and uniform over all possible pairs of the variables. The dimensionality of this hyperparameter depends on the likelihood used. We consider 3 models of expert feedback:

- Bernoulli model: The expert gives feedback $\boldsymbol{e}_k^{ij} \in \{0, 1\}$, with 1 representing that they believe the edge $X_i \to X_j$ exists and 0 representing that this edge does not exist.
  - $\mathbb{P}(\boldsymbol{e}_k^{ij}|\boldsymbol{G}_{ij}) \sim Bernoulli(\mu \boldsymbol{G}_{ij} + (1-\mu)(1-\boldsymbol{G}_{ij}))$
- Beta model: The expert gives feedback $\boldsymbol{e}_k^{ij} \in [0, 1]$, representing the probability that they believe the edge $X_i \to X_j$ exists. This allows for the expert to give a feedback value of 0.5 if they do not know the answer.
  - $\mathbb{P}(\boldsymbol{e}_k^{ij}|\boldsymbol{G}_{ij}) \sim Beta(\alpha(\mu, \boldsymbol{G}_{ij}), \beta(\mu, \boldsymbol{G}_{ij}))$
- Dirichlet model: The expert gives feedback $\boldsymbol{e}_k^{ij} \in \Delta_2$, representing the probabilities for the 3 possible edge states between the pair of variables X_i, X_j, namely $\{X_i \to X_j, X_i\ X_j, X_j \to X_i\}$, assuming a fixed ordering for this representation. The expert can give feedback such as $(1/3, 1/3, 1/3)$ if they do not have any information on the direct causal relations between X_i and X_j, or $(1/2, 0, 1/2)$ to convey that they know there is a direct causal relationship but they do not know which direction it is in.
  - $\mathbb{P}(\boldsymbol{e}_k^{ij}|\boldsymbol{G}_{ij}) \sim Dirichlet(\alpha(\mu, \boldsymbol{G}_{ij}), \beta(\mu, \boldsymbol{G}_{ij}))$

To incorporate expert feedback in a principled manner, a full generative model over the unknown graph and SCM parameters is needed. We would also prefer the model to be amenable to efficient approximate inference, without compromising the ability to easily integrate expert feedback to the model. The DiBS model Lorch et al. [2021] satisfies these criteria. DiBS makes use of latent variables $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{k \times n}$ which generate the adjacency matrix via $\mathbb{P}(\boldsymbol{G}_{ij}|\boldsymbol{U}, \boldsymbol{V}) \sim Bernoulli(\sigma_\alpha(\boldsymbol{u}_i^T \boldsymbol{v}_j))$ if $i \neq j$ and $\boldsymbol{G}_{ij} = 0$ otherwise. Here $\sigma_\alpha(x) = 1/(1 + e^{-\alpha x})$. An important innovation in their work is that the prior over $\boldsymbol{U}, \boldsymbol{V}$, which is constructed such that the model tends to generate DAG adjacency matrices via the characterization introduced by Zheng et al. [2018]. Inference is done using Stein Variational Gradient Descent (SVGD) Liu and Wang [2016], where a fixed number of samples called particles are iteratively updated such that the final result represents samples from the posterior distribution of interest.

Preliminary results

In order to assess whether expert feedback has value or not, we run a toy experiment with an oracle Bernoulli expert model, whose feedback is always correct. We see that including expert feedback improves the average negative log-likelihood, where negative log-likelihood is divided by the sum of the observational and expert feedback data, $N + M$.

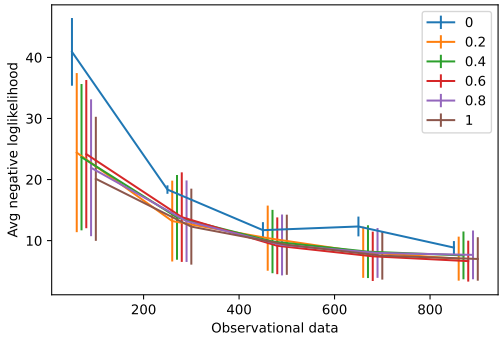

Figure 2: Average negative log-likelihood as a function of observational data samples N. Each curve represents the proportion $p \in \{0, 0.2, \cdots, 1\}$ of (oracle) expert feedback collected from all possible queries, which is d^2 where d is the number of variables. Here we fix $d = 12$. Error bars are over 5 seed values.

Bayesian Experimental Design (BED)

In order to make use of the expert's time in the most efficient manner, we propose using adaptive Bayesian Experimental Design (BED) to select queries. This involves specifying the following information:

- Designs ξ - This is what we control. In this setting, the design is chosen from set of all possible pairs of variables which we query the expert about.
- Observable outcomes y. This is what we observe after running the experiment with some design. In this setting, it is an element from the set of values the expert model can take.
- The quantity of interest - In this setting this is the intervention defined in Equation 2. To estimate this we require a model over the graph adjacency matrix $\boldsymbol{G}$ and the SCM function parameters $\boldsymbol{\Theta}$.

The BED framework in the adaptive setting selects the optimal design at each time step t by solving the following optimization problem

$$\xi^* = \text{argmax}_\xi \int U(y, \xi, h_{t-1}) \mathbb{P}(y \mid \xi, h_{t-1}) \; dy \tag{3}$$

Where $h_t = \{(y_i, \xi_i)\}_{i=1}^t$ with $h_0 = \emptyset$ and $U(y, \xi, h_{t-1})$ is a chosen utility of the outcome y generated from the design ξ. A common choice for U is the Information Gain (IG), defined as

$$IG(y, \xi, h_{t-1}) = \mathbb{H}[\mathbb{P}(\boldsymbol{G}, \boldsymbol{\Theta}|h_{t-1})] - \mathbb{H}[\mathbb{P}(\boldsymbol{G}, \boldsymbol{\Theta}|y, \xi, h_{t-1})] \tag{4}$$

The resulting BED objective below is called the Expected Information Gain (EIG).

$$EIG_t(\xi, h_{t-1}) = \mathbb{E}_{\mathbb{P}(y|\xi, h_{t-1})}\left[\mathbb{H}[\mathbb{P}(\boldsymbol{G}, \boldsymbol{\Theta}|h_{t-1})] - \mathbb{H}[\mathbb{P}(\boldsymbol{G}, \boldsymbol{\Theta}|y, \xi, h_{t-1})]\right] \tag{5}$$

The reason we need a model over the SCM above is similar to why we need a surrogate model in Bayesian Optimaization (BO). Cast in the framework of BED, the quantity of interest in BO is the global minimizer $x^* = \text{argmin}_x \, f(x)$ and a model over f is introduced to define the relevant utility functions.

4 Related Work

Incorporating expert information

In Cano et al. [2011], the authors propose to learn the structure of Bayesian Networks by querying an expert about the existence of edges. Queries are chosen such that the posterior entropy of the parent sets are reduced. This assumes an ordering given apriori, and importance sampling is used to approximate the distributions over the parent sets. A more sophisticated iteration is presented in Masegosa and Moral [2013], which drops the assumption of needing the causal ordering. Our work differs from this as we give causal semantics to the graph, and we do not aim to learn the whole graph. The Bayesian treatment also makes it possible for more flexible feedback as well.

Sequential experimental design for causal queries

In Toth et al. [2022], the authors propose an approach which couples together experimental design for gathering interventional data and inferring a causal query. A causal query is in principle a function of the SCM, which is modelled using DiBS. This is in contrast to our work where we assume gathering interventional data is too costly, and the only intervention to be performed is the downstream task which the practitioner is interested in.

Our optimization objective in Equation 2 is the same as the Causal Global Optimization (CGO) objective introduced in Causal Bayesian Optimization (CBO) Aglietti et al. [2020] if all interventions have uniform cost and assuming the true graph $\mathcal{G}^*$ is known. The CGO objective is

$$\boldsymbol{X_I^*}, \boldsymbol{x_I^*} = \text{argmin}_{\boldsymbol{X_I} \in \mathcal{P}(\boldsymbol{X} \setminus X_t), \boldsymbol{x_I} \in D(\boldsymbol{X_{I \setminus t}})} \mathbb{E}_{\mathbb{P}(X_t | do(\boldsymbol{X_I} = \boldsymbol{x_I}), \mathcal{G}^*)}[X_t] \tag{6}$$

If we let $\mathbf{1}_{X_t \in \mathcal{T}}$ be the indicator random variable for the event $\{X_t \in \mathcal{T}\}$ we see that

$$\mathbb{E}_{\mathbb{P}(X_t|do(\boldsymbol{X_I}=\boldsymbol{x_I}),\mathcal{G}^*)}[\mathbf{1}_{X_t \in \mathcal{T}}] = \mathbb{P}(X_t \in \mathcal{T}|do(\boldsymbol{X_I} = \boldsymbol{x_I}), \mathcal{G}^*) \tag{7}$$

Thus when $-\mathbf{1}_{X_t \in \mathcal{T}}$ is the observable outcome in CBO, Equation 6 above becomes our optimization objective in Equation 2 with the assumption that the true causal graph is known. This assumption in CBO has been relaxed recently Branchini et al. [2022].

Selection of optimal interventions to identify causal effects

Recent work has also focused on finding optimal interventions which allow for causal effects to be identified, taking an algorithmic approach. In Kandasamy et al. [2019], the authors provide an algorithm to find the set of interventions required to identify all causal effects in a DAG. They refer to this set of interventions as the Minimum Intervention Cover (MIC) of the graph. In Akbari et al. [2022] the authors show that finding the minimum cost interventions to identify a causal effect of interest is NP-hard, and provide heuristic algorithms to overcome this complexity. This is in contrast to this work, where the focus is on optimizing a pre-specified causal effect rather than its identifiability.

5 Discussion and open problems

In contrast to running experiments to gather interventional data, which is a process which identifies edge structure with no bias albeit with a high variance, our approach directly gathers information about the edge structure, at the cost of potential biases.

Some open problems related to this work include the following.

Incorporating more than pairwise knowledge

Currently, we only consider edge information between pairs of variables. However, there are other forms of expert knowledge such as existence of paths between variables, non-existence of some variables in certain paths etc. which could be addressed as well Constantinou et al. [2022]. The inclusion of more possibilities could help narrow down causal structure significantly in systems with many variables where interventional data is hard to come by.

Probabilistic models for SCMs and their inference

Our experiments suggest that given oracle expert feedback, the DiBS model with SVGD has an easier time recovering the true graph with a single particle rather than multiple particles. This could be for many reasons, for e.g. due to the repulsive force between the particles pushing the particles away from the true graph. Model misspecification is also a problem. For example even if our posterior samples of the graphs are all from the same MEC, we could have different edge directions giving different results for estimates of causal effects in the BED step.

Expert Model parameters

The hyperparameters of the expert model can be interpreted as for example how reliable they are. Realistically, an expert's knowledge over a causal graph is not uniform over all nodes and edges. One suggestion to learn this is to use interventional data to check for the consistency between the expert's beliefs and what the interventional data suggest.

Experimental design loop

In the current formulation, the design space is discrete, and grows quadratically with the number of variables in the system. When this number is large we may need to resort to heuristics to optimize the EIG. Moreover, the current approach to select queries is myopic. This is suboptimal compared to an experimental design approach that takes into account the available budget. The current approach is also susceptible to time lags in between querying the expert, which may lead to challenges in deploying such a system in the real world. There is recent work which has addressed the latter 2 issues by using a neural network to learn a design policy Foster et al. [2021].

References

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, page 3155–3164. PMLR, Jun 2020. URL https://proceedings.mlr.press/v108/aglietti20a.html.

Raj Agrawal, Chandler Squires, Karren Yang, Karthik Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. (arXiv:1902.10347), Feb 2019. URL http://arxiv.org/abs/1902.10347. number: arXiv:1902.10347 arXiv:1902.10347 [stat].

Sina Akbari, Jalal Etesami, and Negar Kiyavash. Minimum cost intervention design for causal effect identification. In *Proceedings of the 39th International Conference on Machine Learning*, page 258–289. PMLR, Jun 2022. URL https://proceedings.mlr.press/v162/akbari22a.html.

Nicola Branchini, Virginia Aglietti, Neil Dhir, and Theodoros Damoulas. Causal entropy optimization. (arXiv:2208.10981), Aug 2022. URL http://arxiv.org/abs/2208.10981. arXiv:2208.10981 [cs, stat].

A. Cano, A. R. Masegosa, and S. Moral. A method for integrating expert knowledge when learning bayesian networks from data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(5):1382–1394, Oct 2011. ISSN 1083-4419, 1941-0492. doi: 10.1109/TSMCB.2011.2148197.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, Mar 2002. ISSN 1532-4435. doi: 10.1162/153244302760200696.

Anthony C. Constantinou, Zhigao Guo, and Neville K. Kitson. The impact of prior knowledge on causal structure learning. (arXiv:2102.00473), Apr 2022. URL http://arxiv.org/abs/2102.00473. number: arXiv:2102.00473 arXiv:2102.00473 [cs].

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, page 3384–3395. PMLR, Jul 2021. URL https://proceedings.mlr.press/v139/foster21a.html.

Juan L. Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. In *Advances in Neural Information Processing Systems*, volume 33, page 15464–15475. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b197ffdef2ddc3308584dce7afa3661b-Abstract.html.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, Jun 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524.

Yangbo He, Jinzhu Jia, and Bin Yu. Counting and exploring sizes of markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 16(79):2589–2609, 2015. ISSN 1533-7928.

Saravanan Kandasamy, Arnab Bhattacharyya, and Vasant G. Honavar. Minimum intervention cover of a causal graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:2876–2885, Jul 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33012876.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/b3ba8f1bee1238a2f37603d90b58898d-Abstract.html.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. In *Advances in Neural Information Processing Systems*, volume 34, page 24111–24123. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ca6ab34959489659f8c3776aaf1f8efd-Abstract.html.

Andrés R. Masegosa and Serafín Moral. An interactive approach for bayesian network learning using domain/expert knowledge. *International Journal of Approximate Reasoning*, 54(8):1168–1181, Oct 2013. ISSN 0888613X. doi: 10.1016/j.ijar.2013.03.009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 0-465-09760-X.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius von Kügelgen. Active bayesian causal inference. (arXiv:2206.02063), Jun 2022. URL `http://arxiv.org/abs/2206.02063`. arXiv:2206.02063 [cs, stat].

Ts Verma and Judea Pearl. *Equivalence and Synthesis of Causal Models*, page 221–236. ACM, New York, NY, USA, 1 edition, 1990. ISBN 978-1-4503-9586-1. doi: 10.1145/3501714.3501732. URL `https://dl.acm.org/doi/10.1145/3501714.3501732`.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html`.