Multi-Environment Functional Causal Models using Gaussian Processes

Anonymous authors
Paper under double-blind review

Abstract

Causal learning plays an important role in decision-making, especially in fields such as healthcare wherein interventions depend on the underlying causal factors. For example, there are antiviral drugs for influenza but not other viral respiratory infections even though the same symptoms may manifest; highlighting the difference in interventions based on the causal etiology. Functional causal discovery methods can capture the underlying causal relations between random variables, in contrast to structure discovery methods which produce multiple causal graphs in the Markov equivalence class, failing to identify a unique graph. Functional Causal Models (FCMs) also can play an important role in *unsupervised domain adaptation* settings, by learning robust causal relationships from source environments for use in a target environment. Moreover, a Bayesian approach to functional causal discovery can help in quantifying associated uncertainties with causal relations given that uncertainty can arise due to distribution shift across environments. Accordingly, here we posit a multi-environment functional causal model which uses Gaussian processes to learn the functional causal form while exploiting the property that the residuals remain invariant across environments, to capture the true causal parents. We show initial results on multivariate synthetic data as well as real world cause-effect pairs.

1 Introduction

Healthcare data is often subject to differences across environments due to issues such as differences in measurement conventions across different provider or hospital sites, or changes in policies over time (Ghassemi et al., 2018). In the presence of such differences, learning from data across multiple environments plays a critical role to strengthen understanding of the underlying causal relations. For example, it has been shown that treatment policies for sepsis can be influenced by environment-specific factors such as severity which have been found to affect mortality rates, however this could be averted by being able to discover and treat the underlying cause in a way that is unbiased by environment specific policies (Esteban et al., 2007). Hence, it can be beneficial to incorporate information from multiple environments to learn causal relations. Given observational data from multiple source environments, we aim to learn the functional causal form which can then be used for improved decision making in a target environment. We first discuss the advances in structural and functional learning and then present an initial implementation of multi-environment functional causal models using Gaussian processes.

The healthcare community typically relies on knowledge graphs for understanding causal structure (Nordon et al., 2019; Rotmensch et al., 2017). The advance of electronic health records has enabled empirical learning of structural causal graphs; approaches generally focus on pruning edges in existing knowledge graphs based on the co-occurrence of symptoms and diseases (Nordon et al., 2019) and using parametric methods to learn causal relations under the assumption of a bipartite graph Chen et al. (2019). Other approaches to learning causal structure focus on conditional independence test based algorithms like PC[1] and FCI (Fast Causal Inference) (Spirtes et al., 2000) that recover the causal structure under a Markov Equivalence class. As Markov Equivalence classes contain multiple graphs these structure discovery algorithms may not be able to identify all the edges in the graph and cannot determine a single unique causal graph explaining the functional causal form between

[1] PC stands for Peter and Clark, named after Peter Spirtes and Clark Glymour

the random variables. However, FCMs can identify the causal direction between variables as well as capture the functional form in multivariate settings.

In the two variable setting, work has shown that the causal direction is identifiable under outcome-dependent selection bias with the constraint that the noise term is non-Gaussian as identified by Zhang et al. (2016). For the multivariate setting the functional causal model has been identified under linearity assumptions in methods such as LinGAM and using Additive Noise Models (ANM) (Shimizu et al., 2006; Peters et al., 2014). A deep learning approach to causal discovery compares the maximum mean discrepancy (MMD) between the observational data and that generated by a parametric generative network to learn the FCM under multivariate settings (Goudet et al., 2017). Since the search space is super-exponential in the number of random variables, the probable structures are restricted via knowledge of the skeleton structure of the causal graph. To-date, methods such as the above have not leveraged information from multiple environments, which limits the application of causal relations to an unknown target environment. Accordingly, here we propose an approach to learn the functional causal model from multiple environments. In order to prevent capture of any spurious correlations under multiple environments 1) we adopt a Bayesian approach, restricting the priors over the graph structure as well as their associated parameters (Heckerman, 1995; Heckerman et al., 2006); and 2) the decision to accept the causal parents of a random variable is based on the predictive variance of the residuals across multiple source environments. We accomplish this via a Gaussian process based method to capture the FCM via the causal parents of the outcome of interest. We show that proposed approach is able to capture the functional form under the multivariate non-linear setting as well as determine the direction between the cause and effect for two random variables. The initial work improves the marginal likelihood in the target environment over a non-causal model.

2 Background and Notation

We consider our data-generating process to be characterized by a Structural Causal Model (SCM) as introduced by Pearl et al. (2009) over a set of d real-valued observed variables $\mathbf{X} = \{X_1, X_2, ..., X_d\}$. We assume that the corresponding (data-generating) causal graph G is a directed acyclic graph (DAG) with the functional form between the variable X_i and its causal parents $\mathbf{Pa}_i^G$ captured by an additive noise model (ANM) (Hoyer et al., 2009) represented by Equation 1.

$$X_i = f_i(\mathbf{Pa}_i^G) + \epsilon_i, \quad i = (1, 2, .., d) \tag{1}$$

The causal parents $\mathbf{Pa}_i$ are factorized according to the causal graph G in accordance with the causal structure.

$$P(X_1, X_2, ..., X_d \mid G) = \prod_{i \in \{1,2,...,d\}} P(X_i \mid \mathbf{Pa}_i^G) \tag{2}$$

Here, we are interested in learning the structure of the local causal parents of an outcome of interest, the effect (E), as well as the functional form.

- **Structure modularity** The prior $P(G)$ can be written in the form: $P(G) \propto \prod_i \rho(X_i, \mathbf{Pa}_i^G)$

That is, the prior decomposes into a product with a term for each family (node and its causal parents) in G where $\rho(X_i, \mathbf{Pa}_i^G)$ represents the joint distribution of node X_i and it's causal parents $\mathbf{Pa}_i$. This implies that the choices of the causal parents for the different nodes is independent a priori. The functional form between X_i and its causal parents $\mathbf{Pa}_i^G$ represented by Equation 1 is used to define functional modularity which states that the functional form for a particular variable is only dependent on its causal parents and independent of any other nodes a priori.

- **Functional modularity** Let G and G' be two graphs in which $\mathbf{Pa}_i^G = \mathbf{Pa}_i^{G'}$ (we do not restrict the parents of X_j to be the same across the two graphs) then

$$f_i^G(X_i \mid \mathbf{Pa}_i^G) = f_i^{G'}(X_i \mid \mathbf{Pa}_i^{G'})$$

2.1 Causal Gaussian Processes

Consider a set of causal parent variables $\mathbf{Pa}_i$. We want to model a prior over the variable X_i which we believe to be a function of its causal parents $\mathbf{Pa}_i$. Formally, a stochastic process over $\mathbf{Pa}_i$ is a function that assigns to each $p \in \mathbf{Pa}_i$ a random variable $X_i(p)$. The process is termed as a Gaussian Process (GP) if for each finite set of values $\mathbf{Pa}_i$, the distribution over the corresponding random variables $X_i(p) = \{x_1, x_2, ...x_N)$ for N samples is a multivariate normal distribution.

We can write $f \sim GP(m, k)$ to denote that the process $f(\mathbf{x})$ is a GP with mean m and covariance function k. The joint distribution of X_i is therefore:

$$P(X_i \mid \mathbf{Pa}_i) = \frac{1}{Z}\exp\left(-\frac{1}{2}\left(X_i\right)^T\left(K + \sigma_i^2 I\right)\left(X_i\right)\right) \tag{3}$$

where K is the covariance matrix defined over the parents of the node X_i ($\mathbf{Pa}_i$), σ^2 represents the noise associated with X_i and Z is the normalizing constant. Further, we define the score; $\rho\left(X_i, \mathbf{Pa}_i\right)$ to be the conditional score of $P(X_i \mid \mathbf{Pa}_i^G)$ for a graph G. We thus want to maximize the score, in turn finding the local graph structure G for a node X_i and its parents $\mathbf{Pa}_i^G$ that maximizes the likelihood of the data $D = \{x_{i_{1:N}}, p_{i_{1:N}}\}$ where $p_i \subseteq \{X_k \mid k \neq i\}$; recovering the functional form. The score $\rho\left(X_i, \mathbf{Pa}_i\right)$ is defined as follows:

$$\rho\left(X_i, \mathbf{Pa}_i \mid D, G\right) = (2\pi)^{-\frac{N}{2}}|K_{Pa_i}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}X_i^T K_{Pa_i}^{-1} X_i\right) \tag{4}$$

To find the maximum score we need to iterate over all the possible graph structures which for d variables will be 2^{d-1}. After performing the search over all the possible graph structures $G \in \mathcal{G}$, the graph G with the maximum score represented by Equation 5 provides the potential causal parents $\mathbf{Pa}_i^G$ of the variable X_i.

$$\max_{G \in \mathcal{G}} \rho\left(X_i, \mathbf{Pa}_i \mid D, G\right) \tag{5}$$

3 Learning the functional form across multiple environments

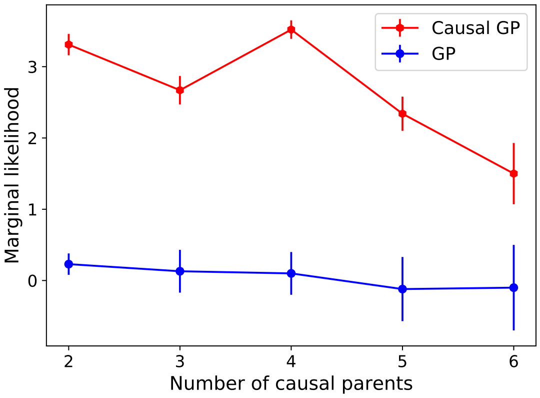
While Equation 5 provides the probable causal parents of E, we want to ensure that it is captures the functional form across multiple environments. The predictive variance of the residuals across the different environments $e_k = \{e_1, e_2, ...e_K\}$ is used to determine validity of the functional causal form (Ghassami et al., 2017). The functional form (f_i) between variables remains the same across the environments while the independent noises (ϵ_i) for variables X_i can change across the environments. The change in noise across environment can reflect differences in measurement of the variables as well as covariate shifts across the environments, common issues in areas such as healthcare (Mhasawade et al., 2019). Here we make a sound assumption that the noise is independent of the causal parents represented in Equation 1. Accordingly, $\mathbb{E}[(E - \bar{f}^i(\mathbf{C}))^2] = \mathbb{E}[(E - \bar{f}^j(\mathbf{C}))^2]$ where E represents the effect and $\mathbf{C}$ represents the causal parents of E. The predictive variance of the gaussian process for environment e^i will be equal to the predictive variance of the gaussian process for environment e^j whereas for the reverse direction this will not hold true, $\mathbb{E}[(\mathbf{C} - \bar{f}^i(E))^2] \neq \mathbb{E}[(\mathbf{C} - \bar{f}^j(E))^2]$. The predictive variance of the posterior of the Gaussian process for the environment e^i can be obtained from the squared residuals as $(y^i - \bar{f}(x^i))^2$ where $\bar{f^i}\,(x)$ is the posterior mean of the GP in environment e^i. For causal parents the GP conditional posterior will collapse to the true functional form between the variable and its causal parents, which will be reflected in the residual of the posterior predictive distribution. Residual variances across environments can then be compared using an F-test to determine the stable causal parents.[2]

4 Experiments and discussion

4.1 Real data: cause-effect pairs

First, we looked at cause-effect pairs from real datasets (`http://webdav.tuebingen.mpg.de/cause-effect/`). We selected datasets which are likely to suffer from bias according to commonsense or background knowledge as was done in (Zhang et al., 2016); we selected pairs 1, 2, 25 and 33. For each dataset we held out 10% as the target and divide the remaining into two source environments. We add noise to the target environment data to represent bias under

[2] The F-test can be used to compare two sample variances to determine if the differences between the two are due dataset shift or due to some inherent measurement noise.



(a) Marginal likelihood in the target environment with increasing number of causal parents reported over 10 runs.

	Residuals	
Dataset	Causal GP	LinGAM
Pair 1	0.014 ± 0.015	1.462 ± 0.733
Pair 2	0.081 ± 0.001	3.977 ± 1.010
Pair 25	0.104 ± 0.009	4.535 ± 0.945
Pair 33	0.062 ± 0.001	2.061 ± 1.700
	Marginal Likelihood (Diff)	
Pair 1	5.440 ± 0.981	-
Pair 2	3.542 ± 0.742	-
Pair 25	4.091 ± 1.002	-
Pair 33	2.862 ± 0.331	-

(b) Residual obtained from Causal GP and LinGAM methods, difference in marginal likelihood between the causal and non causal direction produced by Causal GP (cannot be evaluated for LinGAM as it is a parametric model) in target environment. Values are reported over 10 runs (avg ± std.dev).

Figure 1: Preliminary results on synthetic and real datasets.

a distribution shift. We then fit a Gaussian process in both directions and compare the residual invariances to determine the causal direction. For the true causal direction the marginal likelihood should be greater as compared to the non-causal direction. Also, residuals in the true causal direction should be lower than those in the non-causal direction. We confirmed the marginal likelihood was larger for the causal direction using our method (Table 1b) for each of the four datasets, and we also report residuals in the target environment for all the pairs along with a comparison to the standard LinGAM method (Shimizu et al., 2006).

4.2 Multivariate functional causal model learning

We explored performance of our method on graphs with increasing numbers of causal parents. Because increasing the number of causal parents exponentially increases the possible graph structures to be explored, based on computational time constraints we explored settings including up to six causal parents. However, the graph includes a total of eight nodes including the outcome of interest (E). We explored the multiple settings with increasing number of causal parents ($|\mathbf{Pa}_E| = \{i \mid 2 \leq i \leq 6\}$) and used the same procedure with multiple source environments explained in Section 4.1 for dividing the data into two source environments. In all settings *Causal GP* was able to capture the structural and functional form between effect variable E and its causal parents $\mathbf{Pa}_E$. We report the marginal likelihood of the target data in Figure 1a for *Causal GP* well as for Gaussian process (GP), which does not search over all the possible structures but instead uses all random variables other than E to capture the functional form. The proposed approach as expected performs consistently better than a simple Gaussian process (a higher marginal likelihood). As the number of causal parents increases the performance of Causal GP decreases but still remains higher than GP up to six causal parents.

5 Conclusion

We present an approach for determining causal relations across multiple environments using Gaussian processes. The proposed method captures complex causal relations resulting in 1) lower residuals on target environment data (better than a linear assumption model) and 2) larger marginal likelihood of target data by harnessing information from multiple environments in a multivariate setting up to six causal parents. In developing this work further, higher dimensional settings should be examined which are common in real-world scenarios and can aid in improved decision-making under distribution shifts.

REFERENCES

Irene Y Chen, Monica Agrawal, Steven Horng, and David Sontag. Robustly extracting medical knowledge from ehrs: A case study of learning a health knowledge graph. *arXiv preprint arXiv:1910.01116*, 2019.

Andrés Esteban, Fernando Frutos-Vivar, Niall D Ferguson, Oscar Peñuelas, José Ángel Lorente, Federico Gordo, Teresa Honrubia, Alejandro Algora, Alejandra Bustos, Gema García, et al. Sepsis incidence and outcome: contrasting the intensive care unit with the hospital ward. *Critical care medicine*, 35(5):1284–1289, 2007.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pp. 3011–3021, 2017.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *arXiv preprint arXiv:1806.00388*, 2018.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Causal generative neural networks. *arXiv preprint arXiv:1711.08936*, 2017.

David Heckerman. A bayesian approach to learning causal networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 285–295. Morgan Kaufmann Publishers Inc., 1995.

David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. In *Innovations in Machine Learning*, pp. 1–28. Springer, 2006.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

Vishwali Mhasawade, Nabeel Abdur Rehman, and Rumi Chunara. Population-aware hierarchical bayesian domain adaptation via multiple-component invariant learning. *arXiv preprint arXiv:1908.09222*, 2019.

Galia Nordon, Gideon Koren, Varda Shalev, Benny Kimelfeld, Uri Shalit, and Kira Radinsky. Building causal graphs from medical literature and electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1102–1109, 2019.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):5994, 2017.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

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

Kun Zhang, Jiji Zhang, Biwei Huang, Bernhard Schölkopf, and Clark Glymour. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *UAI*, 2016.