Adaptive Causal Dimension Reduction

Hedong Yan
Computer Department
Hong Kong Baptist University
Hong Kong, China
herdonyan@life.hkbu.edu.hk

I. Motivation

Dimension reduction can mitigates curse of dimensionality and provide visualization approach to understand our data. It usually transforms the original high-dimensional data into low-dimensional data while minimize some information loss to preserve key properties. And we often work over low-dimensional space to leverage them, such as independence among dimensions, that is not significant in high-dimensional space. Also, we need to make a trade-off between redundancy and causal semantics in dimension reduction. However, causal information emerges in low dimensions space and labels space is rarely considered explicitly. If we preserved causal information among low-dimensional coders and labels, it would help us attain a quantitative understanding for causal effect of one dimension to another dimension. For example, we can generate low-dimensional intervenable and interactable counterfactual coders from medical images (chest X-ray images, etc.) for open scrutiny by experts to reveal potential bias of our model. Without such generating relationships over low-dimensional features and labels, it would be tremendously difficult for experts to understand the causal information rather than spurious correlation relationships of our low-dimensional representation.

We want to maximize the 'causal information' in low-dimensional data to find the most suitable dimension K. There are two challenges of adaptive causal dimension reduction. The first challenge is how to define 'causal information' of a low representation. It means that we must find some formulas to judge which dimension is better than others under causal semantics. The second challenge is the exponential increase of possible causal structures with the increase of representation's dimensions and labels number.

In this report, we introduce detailed aspects of causality and traditional dimension reduction techniques in section II and section III. Then we formulate our research questions and illustrate methods we plan to adopt in section IV. Finally, we present some obtained result and propose our future plan in section V and section VI.

II. Survey of causality

A. Notation from statistic

In 1923, Neyman published his master paper [1], but it was not been translated into English until 1990 [2]. He used the term "unknown potential yield" to indicate the missing "potential outcome" in his randomization experiment for evaluating crop varieties. The Rubin causal model was first named by Holland in 1986 [3]. In Rubin causal model, the first thing is to define interested estimand (potential outcome), and then design assignment mechanism before outcome are measured. Then build a model to do analysis.

There are some basic assumptions in potential outcome framework. Stable unit treatment value assumption (SUTVA) [4] states that unit/individual/sample should be independent from each other and the treatment effect for an individual is stable. Strong ignorablity [5] means that treatment assignments probability should be positive for every treatment value and every individual, and assignment mechanism should be independent of potential outcomes. Consistency requires subjects' response for specific treatment in experiment study is the same as outcome in observation study.

Recently, people are trying to find weaker assumption of strong ignorability, such as single strong ignorability [6] [7], sequential single strong ignorability [8]. Those assumptions requires the number of treatment should be more than one and make assumption of non-existence of multi cause hidden confounder.

Some other works focus on sensitivity analysis of causal inference to provide confidence interval [9]. For example, Rosenbaum's sensitivity parameter [10] Γ and Bahadur [11] efficiency were proposed. They try to separate analysis for exogenous factors from models.

B. Hypotheses from philosophy

Pearl proposed structure causal model and develop related theory, where variables were generated by endogenous variables and exogenous variables [12]. The uncertainty is only from exogenous variables which means the value of effect totally depended on his parents nodes if the value of noise term is fixed. He assumes there exist directed functions that every observable variable will be generated by something, whatever they can are observable or not. Form the view of potential outcome frameworks, Pearl's study of intervention is valid supplement when treatment assignment probability can be set as zero and one.

C. Intervention identification and approximation

Intervention identification can transform the query about the interest effect of given intervention to operational intervention and observable observation. If it is not identifiable, then we can use approximation methods to get a bound of causal effect.

Identification formulas for causal diagram were developed in the last 30 years based on the definition of Pearl's structure causal model. Back-door adjustment, front-door adjustment, and do-calculus for DAG (Directed Acyclic Graph) were named and the proof of those theorems were given in [13] [14] formally. However, the approach of such identification doesn't consider unobservable confounder and automatic identification algorithm. The completeness of such identification methods was also not given. In 2002, [15] proposed a complete criterion "c-factorization" for singleton treatment and singleton outcome. [16] and [17] proposed complete identification algorithms (Huang's algorithm and Shpiser's **ID** algorithm) to transform intervention query without condition variables into function of observation distribution automatically for multiple treatments and outcomes in Bayesian network with hidden variable and semi-Markovian model respectively. And [18] proposed **IDC** algorithm for intervention query with condition and proved the completeness. But all those identification methods doesn't consider the undirected edges (stable symmetric relationships). In 2019, [19] proposed complete identification algorithm for segregated graph to address on such patterns. Also, there are other identification algorithm for causal diagrams with loop [20].

However, it is also meaningful to not assume any intervention on those variables is impossible because active intervention will introduce information that observation can't give us. [21] defined z-identifiability and proposed complete $\mathbf{ID^z}$ algorithm to address on problem that any combination of experiments on $\mathbf{Z}$ can be performed and observable distribution is known for query without condition variables. [22] defined g-identifiability and proposed **gID** algorithm. It can factorize the original intervention query into expression of intervention distribution of $\mathbf{Z}$ and it doesn't need any observational data.

Recently, researchers start to notice it is not the only way to solve the identification problem from the view of SCM (structure causal model) directly. [23] revealed the connection between matrix theory and traditional identification. And they proposed an algorithm that leverage proxy-based methods and traditional methods. Neural identification was first been proposed and theoretically analysed in [24] and they also proved the completeness of their neural identification algorithm which use convergence of maximization and minimization of same neural network with intervention constrain as indicator. However, such neural identification need to retrain models if the assignment values of T and Y were changed.

Comparing with do-calculus based algorithms for structure causal model, po-calculus [25] with single world intervention graph (SWIG) [26] is useful complete identification methods in potential outcome framework.

For not identifiable cases, we can still give a bound to intervention query from observation data. For example, [27] gives the tightest bound to graph with instrument variables. Recently, [28] gives a more tight bound than natural bound for general DAG by utilizing observation data.

D. Transportability and data fusion

Transportability is trying to answer intervention query when population shifting occurred from the view of data generating mechanism. The distribution of observable variables maybe different and the data generating mechanism may be changed when we apply our causal conclusion to another domain. Generally, we will assume that corresponding population distribution is known rather than it need to be learned from sample. [29] formally studied "external validity" from the view of sharing causal diagram with assignment mechanism discrepancy of selected variable that is indicated by a variable set $\mathbf{S}$ and they proposed sID algorithm which is complete to solve this problem if joint distribution is known.

Data fusion was first proposed in [30]. The goal of data fusion is to answer the causal effect at a given population while the inputs are observational data, experimental data, selection biased data, and data from dissimilar population.

However, all those methods assume that superpopulation is known which means we doesn't need learn a model from limited data. This weakness is one of the largest obstacle for application of such identification-based learning methods.

E. Causal discovery and causal representation learning

Causal inference requires causal diagram of graphical model. However, the graphical model of real world is not presumed generally and we need to figure out the real graphs from the whole hypotheses space. Causal discovery is focusing on the how to learn causal diagrams or structure causal models from observational and interventional data. There are many algorithms to discovery the causal diagram or causal diagram class. For instance, PC [31], FCI [32] are independence based algorithm. LiGANM-based methods [33] assume mechanism is linear function with additive noise. Post-nonlinear based methods [34] will assume the mechanism satisfies the following function,

$$x_i = f_{i,2}(f_{i,1}(pa_i) + e_i), i = 1, ..., n \quad (1)$$

where pa_i is parents of x_i, $f_{i,1}$ is an nonlinear function, $f_{i,2}$ is invertible post-nonlinear function, and e_i is noise. However, causal discovery in high dimension space is still an open problem.

Causal representation learning is focusing on the find low dimension causal coder from high dimension data. Researches about causal representation learning can be seen in [35]. For example, CausalVAE [36] add a causal layer to learn linear SCM with additive noise and mask layer to do intervention on such coders to produce novel pictures comparing ConditionalVAE [37]. StructureDecoder [38] learn hierarchy coders in lower dimension to represent causal variables with topological order in structure causal model.

However, the core non-parametric methodology of causal inference 'identification' was not considered in those works now. There are still a lot of ignorance about the lower dimension representation for high dimension variables that will keep slightly invariant in causal information.

F. Neural networks for causality

Sum-product network was first proposed at [39]. There are important properties of sum-product network. The first is it can generate samples quickly and the second is it can calculate any marginal probability of joint distribution that is learning from joint data by one step forward propagation.

GFlowNet [40] [41] that was proposed recently also holds those two proprieties in some degree. Also, conditional sum-product [42] was applied in causal discovery [43] and causal estimation [44] by intervention data.

G. Applications

Causal inference can be widely used in machine learning and other situations for application.

For image recognition, feature disentanglement works, such as stable learning [45] [46] [47], counterfactual attention learning [48], and other causal inspired paper appears in recent years.

For treatment effect estimation, [49] uses precision in estimation of heterogeneous effects (PEHE) and build a dataset IHDP to measure response effect of treatment. [50] uses adjustment formula in their observational study about the effect of maternal smoking to children's autism. [51] uses text as covariate to help estimate treatment effect.

For natural language processing, [52] shows differenct application situation of causal inference in NLP, such as text as outcome, treatment, and confounders.

For reinforement learning, it can be used as sample-efficient data augmentation method [53].

H. Experiment platforms

1) Dataset: The promotion of large dataset to research is significant and this has been proven by ImageNet. Benchmarking on dataset can help us to evaluate hypothesis, algorithms, and models. However, there are little large datasets collected from reality for causal learning and reasoning task comparing with computer vision and natural language processing. There are two challenges to benchmark causal algorithms and models that is totally different from traditional correlation data benchmarking. On the one hand, evaluate interventions often cost far more time and money than prediction for algorithms and models. Sometime interventions are even immoral. For example, we can't encourage or force someone to smoke. On the other hand, counterfactual data can never be collected theoretically and there is lacking of credible methodology and enough representative researches to transform the reality dataset into counterfactual dataset. Table I will give some datasets that may be useful for causal tasks.

2) Packages: Another prospective for building experiment platform is maintain unified packages in causal toolbox. It can help researchers to propose and test novel ideas quickly, thus promote the development of causal science. There are many packages that implement pipeline of causal learning or reasoning. Some of them will provide standard and state-of-art learning and reasoning algorithms, such as causal-learn. Related work about causal packages are illustrated in Table II.

III. Survey of Dimension Reduction

PCA [72] [73] is a linear dimension reduction technique. It use orthogonal transformation to attain uncorrected principal components. Auto-encoder [74] [75] is a kind of representative nonlinear dimension reduction technique. It usually use neural networks and gradient-based optimization to learn the parameters for efficient computation. The reconstruction error is an important part of loss function in auto-encoders.

Recently, researchers start to notice the potential benefits if we introduce causality into our low-dimensional representation. CausalVAE [36] introduce causality by labeled data and prior distribution of labels. The reason they can learn the DAG over labels is the difference of distributions between causal-direction and anti-causal direction. However, the dimension number of their causal layer is presumed because the information of causality in their low-dimensional representation is from labels directly. So they can not give a criterion to decide how much dimension we need in our low-dimensional representation for high-dimensional data.

IV. Research Questions and Methods

A. Research questions

1) : Without loss of generality, given encoders $X_{L_1} = E_1(X_H)$ and $X_{L_2} = E_2(X_H)$, how to compare the causal information in X_{L_1} and X_{L_2} so that we can make a trade-off between causal information losing and redundancy? Specifically, if we had an encoder $X_L = E_L(X_H)$, how to calculate causal information $CI(X_L|E_L, X_H)$ and representation size $L = H(E_L|X_H)$ to get the adaptive representation size L^* and encoder E_{L^*}?

2) : The computation of causal information is highly probable to be exponential scale due to potential causal structures number. How to compute and find the optimum scale of low-dimensional representation efficiently?

B. Methods

1) Causal information: In causality, many algorithms based on causal sufficiency assumption, causal faithfulness assumption, and causal Markov assumption. However, those assumption was not always satisfied. For causal information calculation in our low dimension representation, we decide to use deduction methods from all non-parametric causal model using modern computing device and hypotheses testing methods to gain a measurement to decide the dimension number K.

Specifically, we will introduce identifiable structure bias for our low dimensional representation. Identifiable bias means we only search our optimum dimension K in the space of identifiable models.

2) Model: We will use linear non-Gaussian encoding model for us primary experiments and theoretical analysis. Then post-nonlinear encoding models will be considered.

TABLE I: Causal Dataset

Type	Name	Introduction	website
Benchmark	Causeme [54]	time-series	https://causeme.uv.es/
Benchmark	JustCause [55]	support IHDP, ACIC etc.	https://justcause.readthedocs.io/en/latest/
Benchmark	e-CARE [56]	reasoning and explanation for NLP	https://scir-sp.github.io
Dataset	IHDP [49]	home visits and IQ testing	https://www.icpsr.umich.edu/web/HMCA/studies/9795
Dataset	Twins [57]	birth weight and mortality	\
Dataset	Jobs [58]	real world data	\
Dataset	ACIC2019	conference challenge	https://sites.google.com/view/acic2019datachallenge/home

TABLE II: Causal Packages

Motivation	Toolbox	Support Team	Introduction
Causal Learning	causal-learn	CMU, DMIR, Gong Mingming team, Shouhei Shimizu team	python version of Tetrad
	Tetrad [59]	CMU	Java
	CausalDiscoveryToolbox [60]	FenTechSolutions	python, DAG/Pair, dataset, independence, structure learning, metrics
	gCastle	Huawei Noah	python, data generation and process, causal structure learning, metrics
	tigramite	Jakob Runge	python, learning from time-series data
Causal Reasoning	Ananke [61] [62] [63]	Ilya Shpitser team	python, support do-calculus
	EconML [64]	Microsoft	python, Econometrics
	dowhy [65]	Microsoft	python
	causalml [66]	Uber	python, campaign target optimization, personalized engagement
	CausalImpact	Google	R, time-series, adertisement and click
	WhyNot	John Miller	python, simulator and environment
	Causal-Curve [67]	Kobrosly, R.W.	python, continuous variable such as price, time and income
	grf [68]	grf-lab of Standford	R
	dosearch [69]	Santtu Tikka	R
	causaleffect [70]	Santtu Tikka	R
	dagitty [71]	\	R, support adjustment formula
End-to-End	causalnex	QuantumBlack	python, 0.11.0, structure learning, domain knowledge, estimation

V. Results obtained

A. Implementation of Shpitser's ID algorithm

In order to estimate the causal effect among different dimensions to calculate causal information to get optimum L^*, we implemented the Shpitser's complete identification algorithm. We did this because we did not find correct open-source codes (including causaleffect, Ananke, dowhy, dagitty [71]) to provide the complete identified mathematical expression of Shpitser's ID algorithm. The algorithm was implemented based on python. The input is a causal diagram, and the output is a mathematical expression using latex language.

B. The function of identification in causal effect estimation

From the identification result, we can train the prediction model and compute causal effect following the factorization results. However, we wondered what would happen if we did not do identification but just prediction. For example, the identification result of figure 1 is $P(C|do(S)) = \frac{\sum_d P(d)P(S,C|d,B)}{\sum_d P(d)P(S|d,B)}$. We choose $C^* = \arg_c \max P(c|do(S))$ as prediction value. The pure Bayesian prediction is $E(C|S,D,B)$. The average prediction is $E(C)$. In the following, we use X_1 denote dopamine, X_2 denote brain, T denote smoking, and Y denote lung cancer.

The experimental properties we are interested in about our model and algorithm after identification is OOD generalization under parametric interventions from correct identification comparing with pure prediction. It can be measured in two aspects: OOD unbiasness and variance. If the estimand is $E(Y_i(1) - Y_i(0))$, then we can use ATE and PEHE as unbiasness and variance measurement respectively.

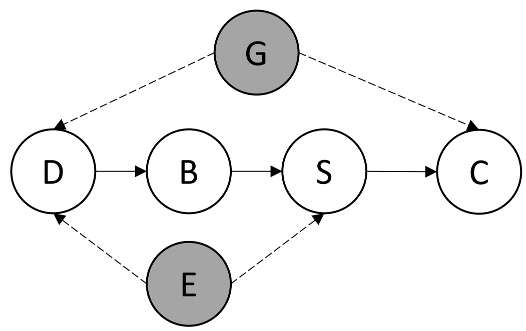

Fig. 1: Example of four variables. D means dopamine; B means senior brain activity (frontal lobe); G means unobserved gene/physique; E means social environment not easy to measure. S means smoking behaviour, and C means cancer. For example, $E \to D$ may represent some life pressures, and $E \to S$ may be unconscious mimic nature.

In our experiment, we use the linear model (same structure with figure 1) as a real-world model to generate data and test the out-of-distribution generalization ability. Each predictor of our association layer model is linear regression or classification model. To keep the consistency with X-learner, we also use two models for treatment and control group separately. We use random transformation and shifting of mechanisms as parametric intervention to test the robustness of our frame-

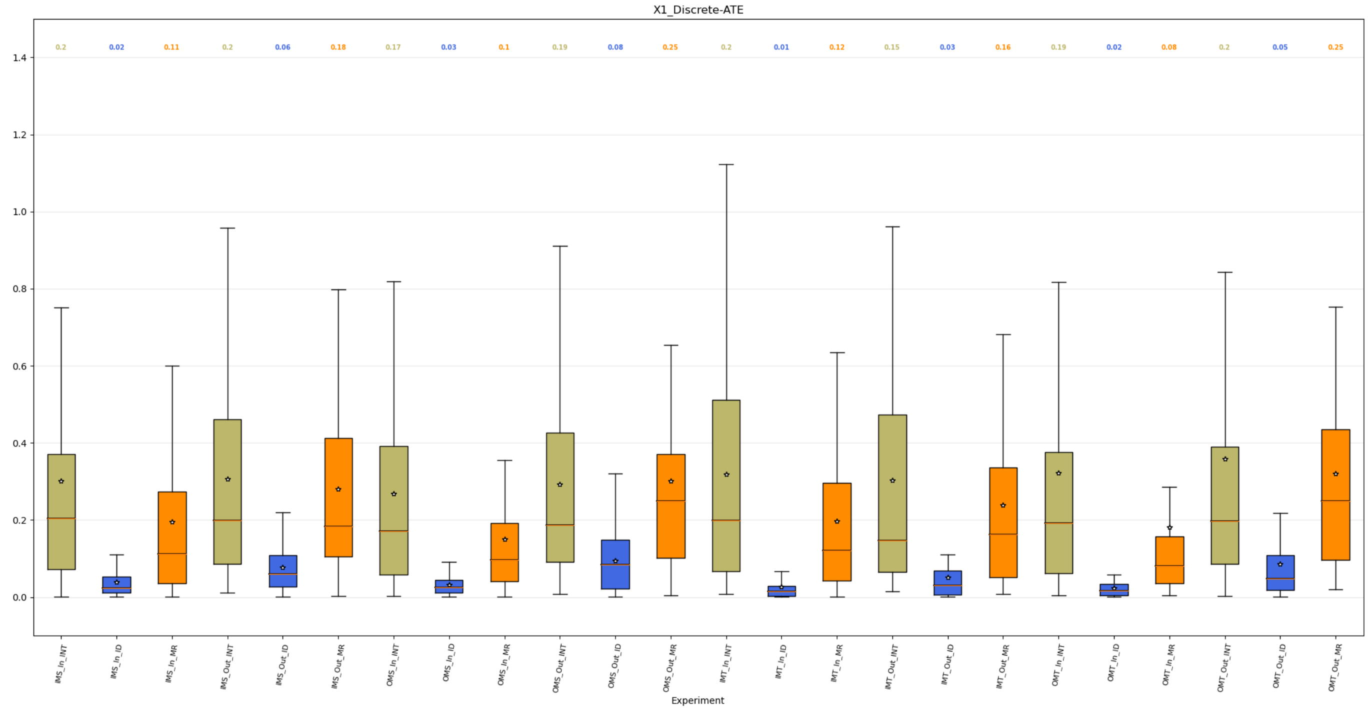

Fig. 2: Experiment error for ATE estimation where X1 is discrete. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

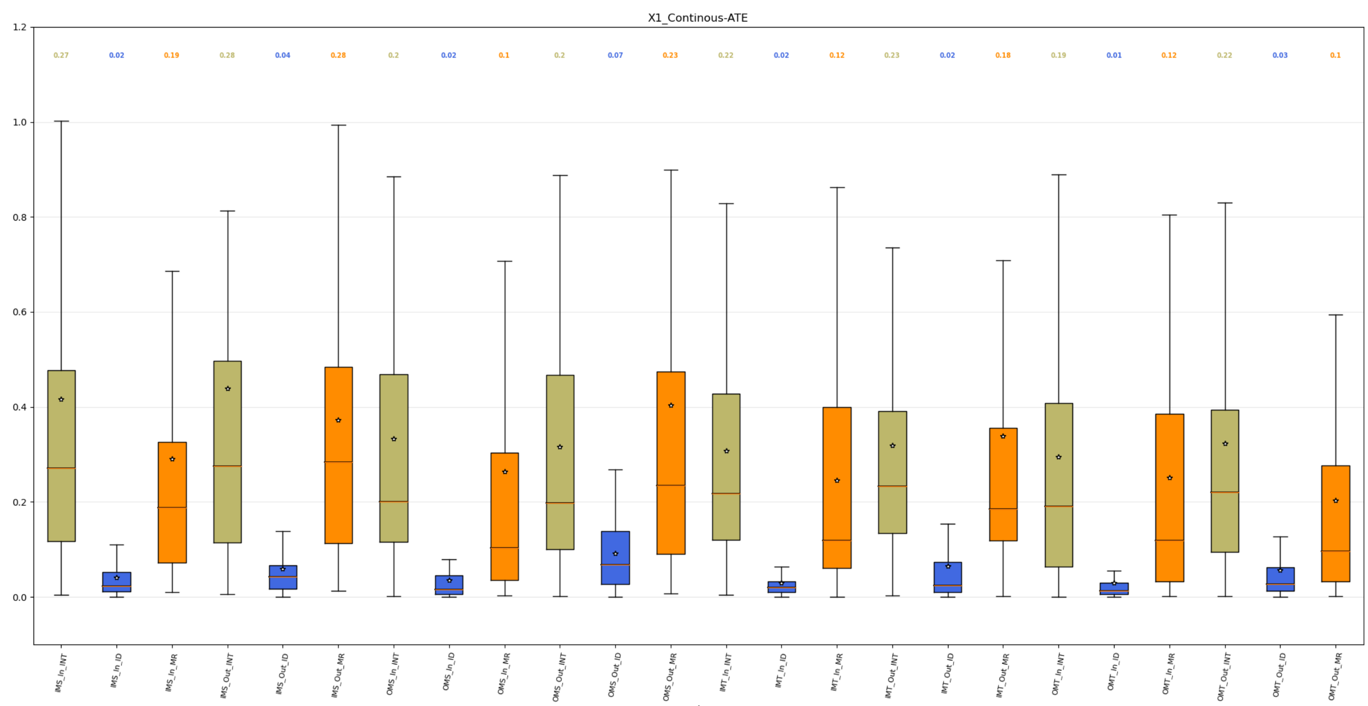

Fig. 3: Experiment error for ATE estimation where X1 is continuous. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

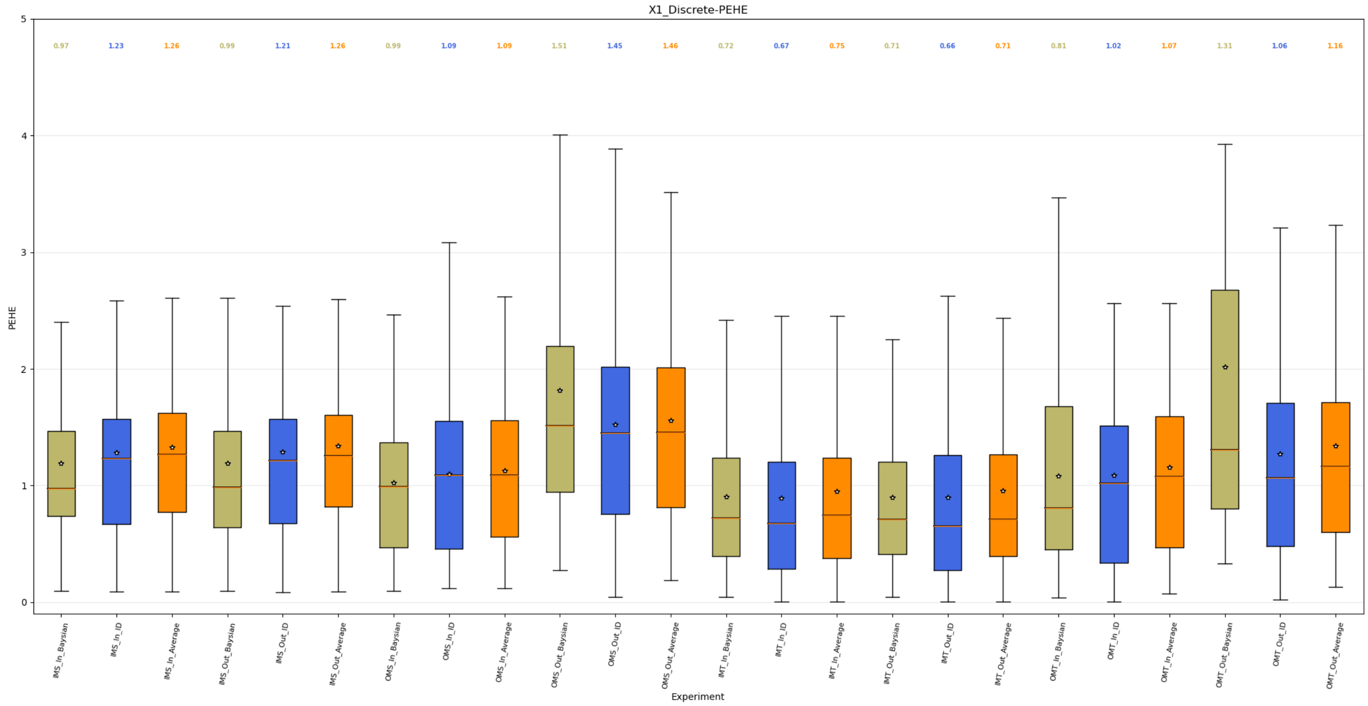

Fig. 4: Experiment error for PEHE estimation where X1 is discrete. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

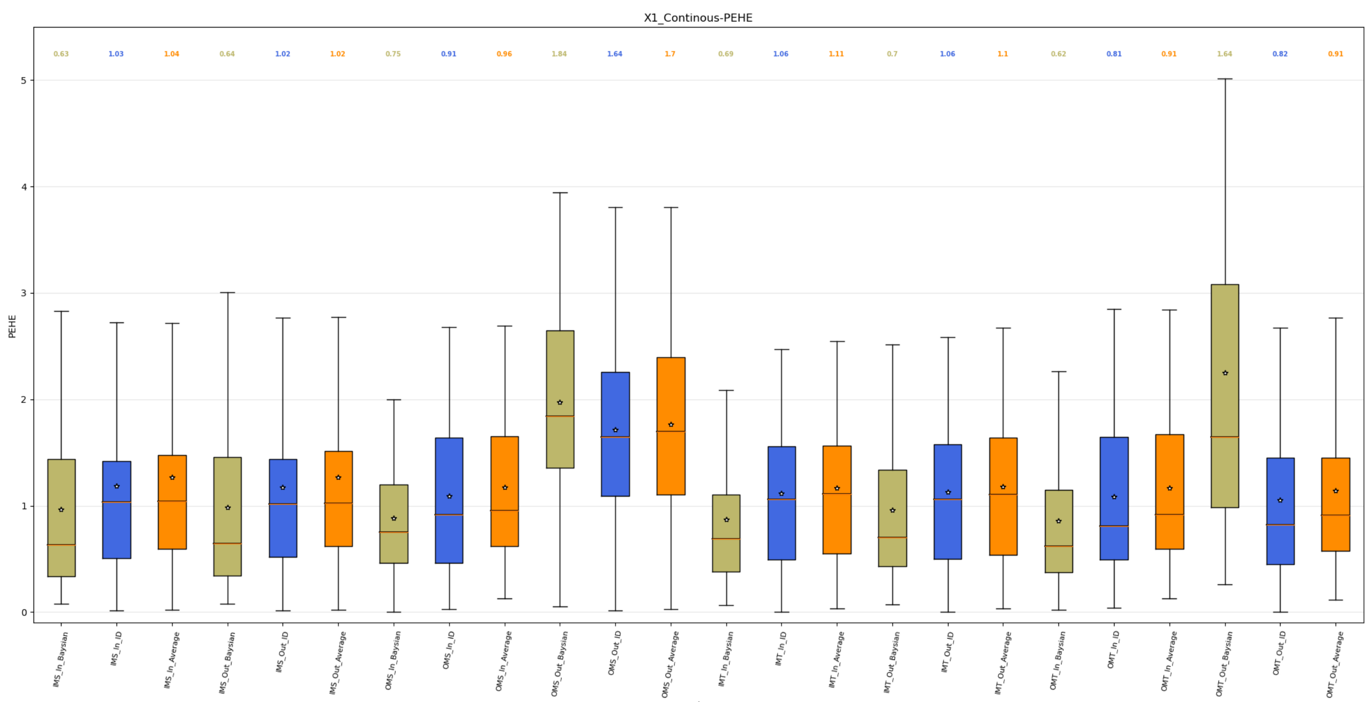

Fig. 5: Experiment error for PEHE estimation where X1 is continuous. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

work. For every setting, we run 50 independent experiments to evaluate the result where there are 1000 samples totally in each experiment.

The train sample number is 800, and the train/valid splitting is 640:160. The test sample number is 200. In algorithm 2 and 3, the sampling numbers of X_1 and (Y, T) are both 100. The dimension of every variable is 1. In optimization, the max epoch is 100000, and we will stop if there is no decrease of loss above 20 and 100 epochs for continuous and discrete testing, respectively. The loss function is MSE loss for regression and Cross Entropy loss for classification; the learning rate is 0.001. When positivity is not satisfied or the joint distribution is zero, we will resample data. The T are discrete variables and X_2 and Y are continuous variables. X_1 can be continuous or discrete variable. We don't use variational method to fitting function of error variance, and use prior noted in the paper directly due to convenience. All the experiment are independent. Figure 6 shows some continuous data. In those figures, left part is train data, and right part is testing data. Yellow and purple means different treatment assignments. And z-axis is value of Y.

Although nonlinear model is not used in our experiments, it can still work if there are nonlinear predictors and environments.

Figure 2, 3, 4, and 5 show the experiment results. We should notice that in-sample testing is not only IID testing due to the missing counterfactual data, and our out-sample testing is under those parametric interventions. In unbiasness testing, estimations after identification are more unbiased than MR [76] and INT [77] from ATE estimation result in both discrete and continuous cases. Considering estimation variance, it got better performance when outer mechanisms are changed.

VI. Future Plan

In next months, we will introduce causal information measurement of low-dimensional representation based on causal effect calculation for deciding which dimension should we reduced to. And we will do both theoretical analysis and empirical studies of our adaptive causal dimension reduction algorithms.

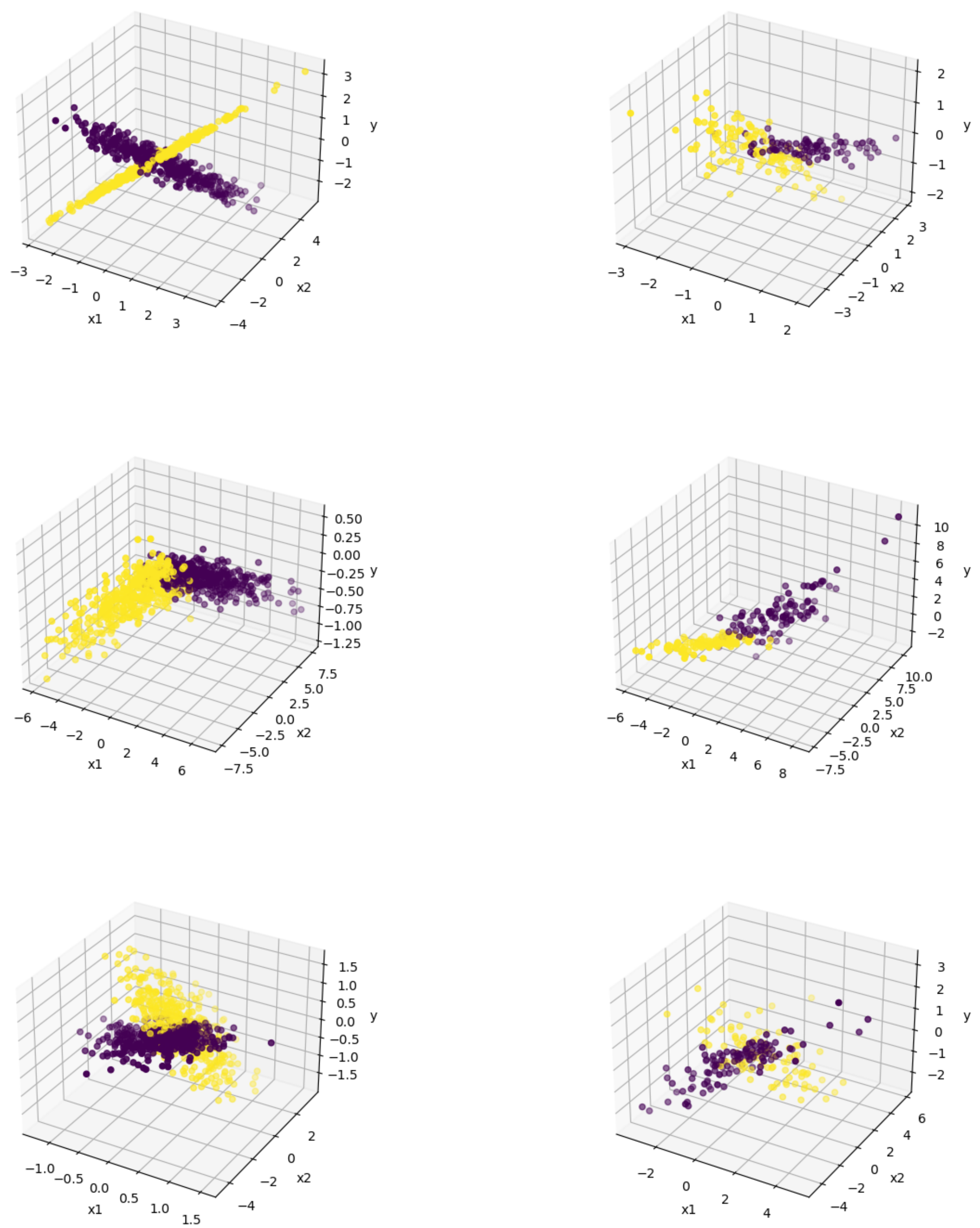

Fig. 6: Some samples of generated data

REFERENCES

[1] J. Neyman, "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, vol. 10, pp. 1–51, 1923.

[2] J. Splawa-Neyman, D. M. Dabrowska, and T. Speed, "On the application of probability theory to agricultural experiments. essay on principles. section 9." *Statistical Science*, pp. 465–472, 1990.

[3] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[4] D. B. Rubin, "Randomization analysis of experimental data: The fisher randomization test comment," *Journal of the American statistical association*, vol. 75, no. 371, pp. 591–593, 1980.

[5] P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.

[6] Y. Wang and D. M. Blei, "The blessings of multiple causes," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1574–1596, 2019.

[7] A. D'Amour, "On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives," *arXiv preprint arXiv:1902.10286*, 2019.

[8] I. Bica, A. Alaa, and M. Van Der Schaar, "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders," in *International Conference on Machine Learning*. PMLR, 2020, pp. 884–895.

[9] A. Franks, A. D'Amour, and A. Feller, "Flexible sensitivity analysis for observational studies without observable implications," *Journal of the American Statistical Association*, 2019.

[10] P. R. Rosenbaum, "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, vol. 74, no. 1, pp. 13–26, 1987.

[11] R. R. Bahadur, *Some limit theorems in statistics*. SIAM, 1971.

[12] J. Pearl, *Causality*. Cambridge university press, 2009.

[13] ——, "[bayesian analysis in expert systems]: comment: graphical models, causality and intervention," *Statistical Science*, vol. 8, no. 3, pp. 266–269, 1993.

[14] ——, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[15] J. Tian and J. Pearl, "A general identification condition for causal effects," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada*, R. Dechter, M. J. Kearns, and R. S. Sutton, Eds. AAAI Press / The MIT Press, 2002, pp. 567–573. [Online]. Available: http://www.aaai.org/Library/AAAI/2002/aaai02-085.php

[16] Y. Huang and M. Valtorta, "Pearl's calculus of intervention is complete," in *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*. AUAI Press, 2006.

[17] I. Shpitser and J. Pearl, "Identification of joint interventional distributions in recursive semi-markovian causal models," in *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*. AAAI Press, 2006, pp. 1219–1226. [Online]. Available: http://www.aaai.org/Library/AAAI/2006/aaai06-191.php

[18] ——, "Identification of conditional interventional distributions," in *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*. AUAI Press, 2006.

[19] E. Sherman and I. Shpitser, "Identification and estimation of causal effects from dependent data," *Advances in neural information processing systems*, vol. 31, 2018.

[20] P. Forré and J. M. Mooij, "Causal calculus in the presence of cycles, latent confounders and selection bias," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 71–80.

[21] E. Bareinboim and J. Pearl, "Causal inference by surrogate experiments: z-identifiability," in *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, N. Freitas and K. Murphy, Eds. Catalina Island, CA: AUAI Press, Aug 2012, pp. 113–120.

[22] S. Lee, J. Correa, and E. Bareinboim, "General identifiability with arbitrary surrogate experiments," in *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. Tel Aviv, Israel: AUAI Press, 2019.

[23] S. Lee and E. Bareinboim, "Causal identification with matrix equations," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[24] K. Xia, K. Lee, E. Bengio, and E. Bareinboim, "The causal-neural connection: Expressiveness, learnability, and inference," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[25] D. Malinsky, I. Shpitser, and T. Richardson, "A potential outcomes calculus for identifying conditional path-specific effects," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3080–3088.

[26] T. S. Richardson and J. M. Robins, "Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality," *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, vol. 128, no. 30, p. 2013, 2013.

[27] A. Balke and J. Pearl, "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 1171–1176, 1997.

[28] J. Zhang and E. Bareinboim, "Non-parametric methods for partial identification of causal effects," Technical Report Technical Report R-72, Columbia University, Department of . . . , Tech. Rep., 2021.

[29] E. Bareinboim and J. Pearl, "Transportability of causal effects: Completeness results," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*, J. Hoffmann and B. Selman, Eds. AAAI Press, 2012. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5188

[30] ——, "Causal inference and the data-fusion problem," in *Proceedings of the National Academy of Sciences*, R. M. Shiffrin, Ed., vol. 113. National Academy of Sciences, 2016, pp. 7345–7352.

[31] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.

[32] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 499–506.

[33] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[34] K. Zhang and A. Hyvarinen, "On the identifiability of the post-nonlinear causal model," *arXiv preprint arXiv:1205.2599*, 2012.

[35] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, 2021. [Online]. Available: https://doi.org/10.1109/JPROC.2021.3058954

[36] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "Causalvae: Disentangled representation learning via neural structural causal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9593–9602.

[37] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[38] F. Leeb, G. Lanzillotta, Y. Annadani, M. Besserve, S. Bauer, and B. Schölkopf, "Structure by architecture: Disentangled representations without regularization," *arXiv preprint arXiv:2006.07796*, 2020.

[39] H. Poon and P. Domingos, "Sum-product networks: A new deep architecture," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 689–690.

[40] Y. Bengio, T. Deleu, E. J. Hu, S. Lahlou, M. Tiwari, and E. Bengio, "Gflownet foundations," *arXiv preprint arXiv:2111.09266*, 2021.

[41] E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio, "Flow network based generative models for non-iterative diverse candidate generation," *arXiv preprint arXiv:2106.04399*, 2021.

[42] X. Shao, A. Molina, A. Vergari, K. Stelzner, R. Peharz, T. Liebig, and K. Kersting, "Conditional sum-product networks: Imposing structure on deep probabilistic architectures," in *International Conference on Probabilistic Graphical Models*. PMLR, 2020, pp. 401–412.

[43] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, "Differentiable causal discovery from interventional data," *arXiv preprint arXiv:2007.01754*, 2020.

[44] M. Zečević, D. S. Dhami, A. Karanam, S. Natarajan, and K. Kersting, "Interventional sum-product networks: Causal inference with tractable probabilistic models," *arXiv preprint arXiv:2102.10440*, 2021.

[45] Z. Shen, P. Cui, T. Zhang, and K. Kunag, "Stable learning via sample reweighting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5692–5699.

[46] R. Xu, P. Cui, Z. Shen, X. Zhang, and T. Zhang, "Why stable learning works? a theory of covariate shift generalization," *arXiv preprint arXiv:2111.02355*, 2021.

[47] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5372–5382.

[48] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1025–1034.

[49] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[50] D. Caramaschi, A. E. Taylor, R. C. Richmond, K. A. Havdahl, J. Golding, C. L. Relton, M. R. Munafò, G. Davey Smith, and D. Rai, "Maternal smoking during pregnancy and autism: using causal inference methods in a birth cohort study," *Translational psychiatry*, vol. 8, no. 1, pp. 1–10, 2018.

[51] L. Yao, S. Li, Y. Li, H. Xue, J. Gao, and A. Zhang, "On the estimation of treatment effect with text covariates," in *International Joint Conference on Artificial Intelligence*, 2019.

[52] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts *et al.*, "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond," *arXiv preprint arXiv:2109.00725*, 2021.

[53] C. Lu, B. Huang, K. Wang, J. M. Hernández-Lobato, K. Zhang, and B. Schölkopf, "Sample-efficient reinforcement learning via counterfactual-based data augmentation," *arXiv preprint arXiv:2012.09092*, 2020.

[54] "Causeme: An online system for benchmarking causal discovery methods."

[55] T. Hawkins and A. Kim, "Just cause," in *Just War Theory and Literary Studies*. Springer, 2021, pp. 55–83.

[56] L. Du, X. Ding, K. Xiong, T. Liu, and B. Qin, "e-care: a new dataset for exploring explainable causal reasoning," *arXiv preprint arXiv:2205.05849*, 2022.

[57] D. Almond, K. Y. Chay, and D. S. Lee, "The costs of low birth weight," *The Quarterly Journal of Economics*, vol. 120, no. 3, pp. 1031–1083, 2005.

[58] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, pp. 604–620, 1986.

[59] J. D. Ramsey, K. Zhang, M. Glymour, R. S. Romero, B. Huang, I. Ebert-Uphoff, S. Samarasinghe, E. A. Barnes, and C. Glymour, "Tetrad—a toolbox for causal discovery," in *8th International Workshop on Climate Informatics*, 2018.

[60] D. Kalainathan and O. Goudet, "Causal discovery toolbox: Uncover causal relationships in python," *arXiv preprint arXiv:1903.02278*, 2019.

[61] R. Nabi, R. Bhattacharya, and I. Shpitser, "Full law identification in graphical models of missing data: Completeness results," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7153–7163.

[62] J. J. Lee and I. Shpitser, "Identification methods with arbitrary interventional distributions as inputs," *arXiv preprint arXiv:2004.01157*, 2020.

[63] R. Bhattacharya, R. Nabi, and I. Shpitser, "Semiparametric inference for causal effects in graphical models with hidden variables," *arXiv preprint arXiv:2003.12659*, 2020.

[64] M. H. G. L. P. O. M. O. V. S. Keith Battocchi, Eleanor Dillon, "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation," https://github.com/microsoft/EconML, 2019, version 0.x.

[65] A. Sharma, E. Kiciman *et al.*, "DoWhy: A Python package for causal inference," https://github.com/microsoft/dowhy, 2019.

[66] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao, "Causalml: Python package for causal machine learning," *arXiv preprint arXiv:2002.11631*, 2020.

[67] R. W. Kobrosly, "causal-curve: A python causal inference package to estimate causal dose-response curves," *Journal of Open Source Software*, vol. 5, no. 52, p. 2523, 2020.

[68] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *The Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019.

[69] S. Tikka, A. Hyttinen, and J. Karvanen, "Causal effect identification from multiple incomplete data sources: A general search-based approach," *arXiv preprint arXiv:1902.01073*, 2019.

[70] S. Tikka and J. Karvanen, "Identifying causal effects with the r package causaleffect," *arXiv preprint arXiv:1806.07161*, 2018.

[71] J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison, "Robust causal inference using directed acyclic graphs: the r package 'dagitty'," *International journal of epidemiology*, vol. 45, no. 6, pp. 1887–1894, 2016.

[72] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[73] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[74] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[75] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[76] D. A. Freedman, "On regression adjustments to experimental data," *Advances in Applied Mathematics*, vol. 40, no. 2, pp. 180–193, 2008.

[77] W. Lin, "Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 295–318, 2013.