Treatment Effect Identification as An Out-of-Distribution Generalization Method for Semi-Markovian Causal Model

Abstract

Evaluating treatment effect plays a vital role in individual medicine in which the interpretability of the prediction model is critical due to unobservable confounders. The challenge is to guarantee consistency when generalizing over unknown distributions. However, current researches mainly focus on treatment effect estimation on specific hypotheses. Identification from structure hypothesis, as the base of estimation, has not been emphasized and integrated into the treatment effect estimation framework. In this paper, we introduce graphical identification methods to create predictors automatically and transform the data distribution specification task into prediction task which was well handled by deep learning. It will take the unknown common cause variables and hidden mechanisms into consideration without modeling them directly. Then we propose a treatment effect estimation algorithm based on identifiable semi-Markovian causal model. The estimator created by identification outperformed the traditional estimator in our linear out-of-distribution testing. The experiment results show the potential ability of complete identification methods to generalize over unknown distribution.

1 INTRODUCTION

In the individual medicine domain, predicting an individual's response when specific treatments are assigned or not is a critical problem for precision medicine and individual therapy. Learning from observational data to predict individuals' potential outcomes has been emphasized these years due to the accessibility of many clinic data. However, the point of existing methods, such as Louizos et al. [2017], Atan et al. [2018], and Bica et al. [2020], is focusing on finding function families of more accurate estimators from specific identification result following proposed hypotheses, rather than creating estimators and combined them from identification result automatically.

Traditional regression models do regression and predict the expectation of the potential outcome of different treatments from covariate and treatment directly, such as T-learner, S-learner, and X-learner. Such a model requires an unconfounded assumption that no confounder exists between treatment and outcome. Louizos et al. [2017] assumes a hidden common cause among covariate, treatment, and response. Their method models and learns the latent common cause to compute causal effect, although the causal diagram of this proxy-variable assumption is not identifiable if the invertibility of some matrix is not satisfied Lee and Bareinboim [2021]. Atan et al. [2018] indicates that treatment bias may exists due to the hidden common casue between covariate and treatment. The individual treatment balancing by their bias removing network is similar to propensity score method Rosenbaum and Rubin [1983] which is an "efficient estimator of the adjustment estimand" Pearl [2009] to help deal with high dimension problem of covariates' representation space. The equivalence of identification between propensity score and adjustment formula was proved in Pearl [2009]. Wang and Blei [2019] and Bica et al. [2020] add mutil-cause assumption based on causal diagram of adjustment graph and learn substitute confounder to do causal inference. The mutil-cause assumption means confounder between treatments and potential outcome has at least two children that are treatments.

However, all these models can be seen as semi-Markovian causal models with three classes of variables (covariate, treatment, and outcome) with some hidden common causes, as shown in figure 2. We also demonstrate the causal diagrams and identified results of those hypotheses in figure 2. And grouping variables into three classes ignore the complex causal mechanisms inside the class and cross those class. For example, if clinic data didn't satisfy unconfoundness assumption and the true mechanisms can be represent

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

by napkin model as figure 1, the identification methods they used, such as back-door or front-door, may not be able to identify the treatment effect without the invertibility of some matrices Lee and Bareinboim [2021]. Also, identifying every structure hypothesis should not be a human job because the ADMG number increase with the rate of $\Theta(2^{N^2})$ where N is the number of observable variables (Appendix B). Also, out-of-distribution challenges from parametric intervention Wang et al. [2021a] and exogenous distribution shifting that was reflected in the data is not considered in those works.

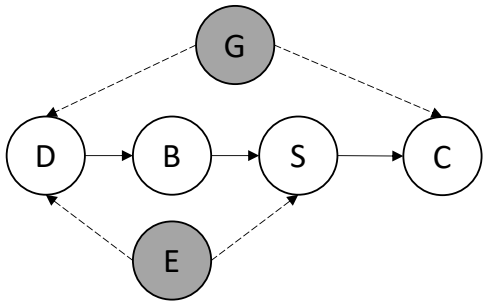


Figure 1: Example of napkin graph. D means dopamine; B means brain environment; G means unobserved gene/physique; E means social environment not easy to measure. S means smoking behaviour, and C means cancer. For example, $E \rightarrow D$ may represent some life pressures, and $E \rightarrow S$ may be unconscious mimic nature.

In this paper, what we do is to automatically create estimators that can be learned by deep learning from identification results and then estimate causal effect to attain unbiased estimation results. And we will use prediction error distribution as proxy of original conditional distribution, to estimate the treatment effect leveraging the causal factorization.

Our main contributions are as follows. First, we formulate Intervention Confidence Expression Graph (ICE Graph) to represent intervention query. The leaf nodes are perdition error distribution and prediction model, the non-leaf nodes are responsible for out-of-distribution generalization ability. Second, we propose an learn and inference algorithm using Mont-Carlo method for ICE Graph to attain likelihood of causal query and develop a novel treatment effect estimation algorithm to create automatic estimators. It enables robustness for unknown distribution and we evaluate it in both continuous and discrete setting.

2 RELATED WORKS

2.1 TREATMENT EFFECT ESTIMATION

Traditional treatment effect estimation is often based on specific hypotheses without explicit inductive bias, such as causal diagrams. Most works about treatment effect estimation divide observed variables into three groups (covariate, treatment, and outcome) to attain an estimator for causal effects, such as ATE, CATE, and ITE.

In traditional Baysian estimator, we use covariates and treatments to estimate potential outcome, $P(Y(t)) = P(Y|T, X)$. The idea of this Baysian estimator is intuitive. However, there are two weakness of such estimator, it presumed that the covariate and treatment is independent and there are no hidden common cause. CEVAE Louizos et al. [2017] and Deconfounder Wang and Blei [2019] use proxy-based based estimation methods. where use back-door criterion to condition on proxy of hidden common cause. However, proxy-based methods may requires some invertible mechanism Lee and Bareinboim [2021]. Deep-treat Atan et al. [2018] is focusing on remove direct influence from covariate to treatment to debias.

Despite of identification of specific assumptions, improving estimator performance is also important.

Künzel et al. [2019] model factual and counterfactual separately. T-learner estimate treatment effect based treatment or not independently. X-learner estimate $Y(0)$ and $Y(1)$ for all individuals, and then compute treatment effect.

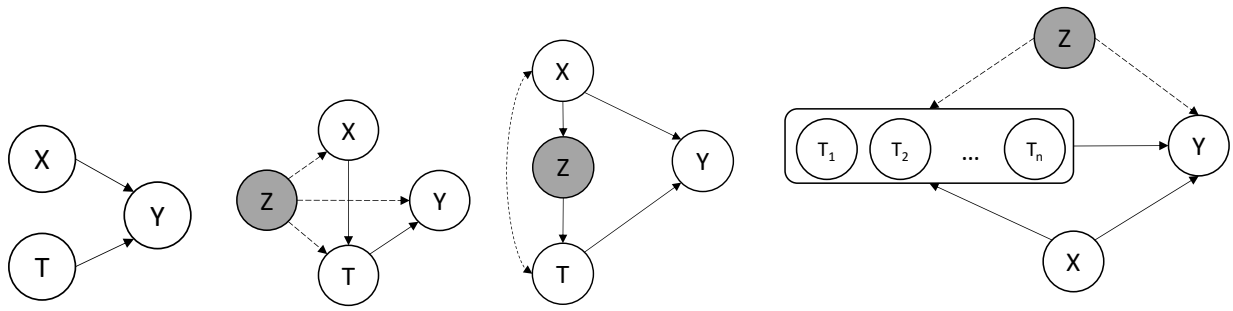
2.2 CROSS-LAYER IDENTIFICATION

Identification strategy in this paper is to transform the queries of estimand Imbens and Rubin [2015] (such as $P(Y(t))$ and $P(Y(t)|X)$) into conditional, marginal and interventional probability estimations that are computable from observational data.

Identification from observation data for treatment effect estimation is feasible if the data satisfy some assumptions. Potential outcome framework Imbens and Rubin [2015] often requires ignorability, positivity, and stable unit treatment variable assumption (SUTVA).

Treatment effect identification is also practicable if we assume causal structure constrains the data generation mechanism rather than the data itself. In 2002, Tian and Pearl [2002a] proposed a sound graphical identification algorithm based on c-factorization for unconditional causal effect estimation. Shpitser and Pearl [2006] proved the completeness and soundness of the hedge criterion and proposed ID and IDC algorithms for unconditional and conditional effect identification, respectively. The soundness and completeness of those two algorithms are also proven by Shpitser and Pearl [2006]. Huang and Valtorta [2006] proposed another complete and sound identification algorithm in the same year independently.

Recently, there are also proxy-based works for the identification of intervention queries. Xia et al. [2021] defined neural effect identification and proved the equivalence between traditional graphical identification and neural identification. Lee and Bareinboim [2021] connected c-factorization and



(a) Baysian (b) CEVAE (c) Deep-treat (d) Deconfounder

Figure 2: Causal graph of related works. In those figures, we use X to denote covariates, T to denote treatment assignment, Y to denote outcomes, and Z to denote proxy of hidden variables. The identification for $P(Y(t))$ in figures are $P(Y|T)$, $\int_z P(z)P(Y|T,z)dz$, $\int_x P(x)P(Y|T,x)dx$, and $\int_{z,x} P(z,x)P(Y|T,z,x)d(z,x)$ respectively. The identification for $P(Y(t)|X)$ in those figures are $P(Y|T,X)$, $\int_z P(Y|T,z)P(z|X)dz$, $P(Y|T,X)$, and $\int_z P(Y|T,z,X)dz$. We emphasise that Z of deconfounder means proxy of mutil-cause confounders, not including single-cause confounders.

matrix equation and proposed a sound identification algorithm that leveraged the advantages of existing approaches.

3 BACKGROUND AND NOTATION

Notation Black letters denote variable set, and calligraphic letters denote mechanism transformation and exogenous distribution shifting. Larger letters indicate variable or mechanism, and small letters represent the specific assignment of variable or mechanism. For consistency with other works, we use $\mathbf{V} = (\mathbf{X}, \mathbf{T}, \mathbf{Y})$ to denote the set of endogenous variables, which includes observable covariates $\mathbf{X}$, treatment variables $\mathbf{T}$, and outcome variables $\mathbf{Y}$. Further, we use $\mathbf{U}$ to denote exogenous variables. We use $\mathbf{F}$ to denote assignment function set of random variables and $\mathcal{W}$ to denote the set of perturbations or transformation that induce out-of-distribution problem.

3.1 SEMI-MARKOVIAN CAUSAL MODEL

The connection of the semi-Markovian causal model and causal Bayesian network with hidden variables can be seen in Tian and Pearl [2002b]. Before we introduce the semi-Markovian causal model, we will first define acyclic directed mixed graphs.

Definition 1. ***Acyclic Directed Mixed Graphs (ADMG)****: an ADMG is a tuple* $(\mathbf{V}, \mathbf{D}, \mathbf{B})$*, where* $V \in \mathbf{V}$ *is a vertex in the graph,* $D \in \mathbf{D}$ *is a directed edge, and* $B \in \mathbf{B}$ *is a bi-directed edge. The constrain of ADMG is that there is no directed circle in ADMG, which means the subgraph* $(\mathbf{V}, \mathbf{D})$ *of ADMG* $(\mathbf{V}, \mathbf{D}, \mathbf{B})$ *is a directed acyclic graph.*

The ADMG describes the assignment procedure of endogenous variables with hidden confounders. Directed edge means the head variable of an arrow is an input of the tail. Bi-directed edge means the tails of bi-directed edge have a common noise as the input of their assignment procedure. Each variable without bi-directed edges has a noise variable as its parent, but we usually do not plot them explicitly for simplification. The nonparametric ADMG didn't indicate the specific function or distribution among those variables. The instantiation of ADMG is called semi-Markovian causal model or semi-Markovian model.

Definition 2. ***Semi-Markovian causal model****: a semi-markovian causal model is defined by a tuple* $(\mathbf{G}, \mathbf{F}, \mathbf{U})$*, in which* $\mathbf{G}$ *is the causal diagram of the model which is an ADMG,* $\mathbf{F}$ *is the assignment program of each variable and* $\mathbf{U}$ *is set of distributions of hidden confounders and independent noise.*

Figure 2 give examples of the ADMG and some possible mechanisms for corresponding semi-Markovian models. We emphasize that the assignment functions are distinct because if variable A and variable B have the same assignment function, they should be represented as one variable in the graph.

3.2 IDENTIFICATION

Then identifiability of an estimand in semi-Markovian model can be defined as follows,

Definition 3. ***Identifiability****: let* $Q(Y, T, C)$ *be any estimand, such as* $P(Y(T))$*,* $P(Y(T)|X)$*. We say that* Q *is identifiable in semi-Markovian causal model set* $\mathbf{M}$ *with same causal diagram* G *if for any pair of models* m_1 *and* m_2 *from* $\mathbf{M}$*,* $q_1 = Q_{m_1} = Q_{m_2} = q_2$ *whenever* $P_{m_1}(\mathbf{v}) = P_{m_2}(\mathbf{v}) > 0$*. For simplicty, we called it* Q^G *is identifiable.*

A direct corollary of definition 3 is as following,

Corollary 1. *If $\mathbf{Q}^G$ is the value set of identifiable queries, and $\mathbf{P}^G(V)$ is the value set of $P^G(V)$. Then there exists a projection $I : \mathbf{P}^G(V) \to \mathbf{Q}^G$ that $Q^G = I(P^G(V))$ when $P^G(V) > 0$.*

Proof. Due to definition 3, for any identifiable $Q_G \in \mathbf{Q}_G$, there exist unique Q_G corresponding with $P_G(V)$. So there exist projection I from $\mathbf{P}_G(V)$ to $\mathbf{Q}_G$. □

The process of finding such projection is identification. We should notice that the projection $I \in \mathbf{I}$ might have different forms. However, for specific estimand $Q_G(\mathbf{Y}, \mathbf{T}, \mathbf{C})$, $I_1(P_G(V)) = I_2(P_G(V))$ if it is identifiable.

What we should notice is that for same observation $P(V)$ and different causal diagrams G_1 and G_2, the identification result may be same. That means, for some cases, $Q_{G_1}(Y, T, C) = Q_{G_2}(Y, T, C)$. For example, there are only five different identification results for query $P(Y(T))$ in three variables cases with hidden confounders of all 200 ADMG (Appendix A): not identifiable (41 ADMG), $P(Y)$ (128 ADMG), $P(Y|T)$ (24 ADMG), $\sum_x P(x)P(Y|T, x)$ (6 ADMG), $\sum_x P(x|T) \sum_t P(Y|t, x)P(t)$ (1 ADMG).

4 INTERVENTION CONFIDENCE EXPRESSION GRAPH

4.1 INTERVENTION EXPRESSION GRAPH

Now, we will introduce the tool we will use to perform causal inference among observation data. Theorem 1 give the definition and expressiveness of the intervention expression graph.

Theorem 1. ***Intervention Expression Graph****: In any semi-Markovian model $M(G(\mathbf{V}, \mathbf{D}, \mathbf{B}), \mathbf{F}, \mathbf{U})$, let a topological order π over G is $V_0 < V_1 < ... < V_n$ where $V_0 = \emptyset$ and denote variables set $\{V_0, V_1, ..., V_i\}$ to $\mathbf{V}_\pi^{(\mathbf{i})}$. For any identifiable intervention query $Pr^M(\mathbf{Y}(\mathbf{T})|\mathbf{X})$, there exist an directed acyclic expression graph with only one root, in which any leaf nodes $\alpha \in \{Pr^M(V_i|\mathbf{V}_\pi^{(\mathbf{i-1})})\}$ and any non-leaf nodes $\beta \in \{\int_{\mathbf{V_c}}(*)d\mathbf{V_c}, (*)^{-1}, \prod(*)\}$ where $\mathbf{V_c} \subseteq \mathbf{V}$. Pr can be probability or probability density.*

The basic idea to prove theorem 1 is to represent the condition probability following a topological order of G and build the expression graph from bottom to top by running the IDC algorithm, which is complete to identify any intervention query with condition Shpitser and Pearl [2006], including continuous variables Tikka and Karvanen [2018]. And the conditional probability calculation in non-leaf nodes can be replaced by $\int_{\mathbf{V_c}}(*)d\mathbf{V_c}$ and $(*)^{-1}$. Thus, an intervention expression graph for the intervention query will be created.

An expression graph can not only be built by graphical identification methods; it can also be determined by other identification strategies, such as unconfounded assignment mechanism (also ignorable)Imbens and Rubin [2015] if the query is identifiable based on such identification strategies. Although possible ADMG number is enormous, corresponding intervention expression graph number may be far smaller (Appendix A). So given estimand, learning expression graph can be easier than learning causal structure.

We use $\sum_{\mathbf{V_c}}(*)$ to denote $\int_{\mathbf{V_c}}(*)d\mathbf{V_c}$ and $\sum_{\mathbf{V_c}}(*)$ in the following part for simplicity. To make it clear, two examples of expression graphs which are determined by IDC algorithm for napkin case and determined based on unconfoundenss were given in Figure 3 and Figure 4. And figure 3 can be further simplified as figure 5.

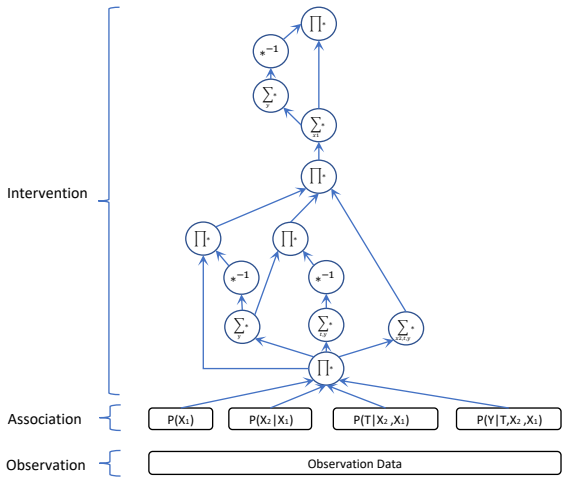


Figure 3: Expression graph of napkin

4.2 INTERVENTION CONFIDENCE GRAPH

Based on theorem 1, we can rewrite any identifiable intervention query into an intervention expression graph. However, there are still challenge we should deal with: the conditional distribution forms of leaf nodes are difficult to specify and learn, especially for sparse data.

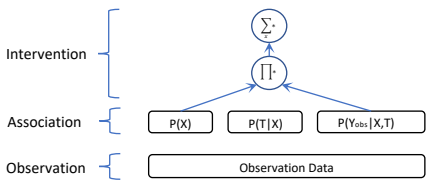


Figure 4: Expression graph of unconfoundness

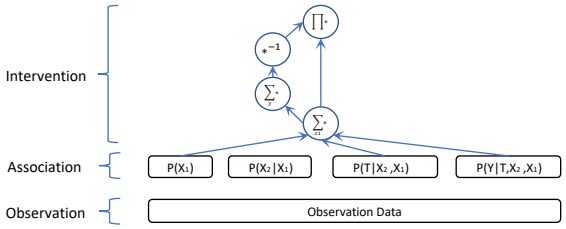


Figure 5: Simplified expression graph of napkin

4.2.1 Prediction model as proxy of data

Inspired by Friston [2009], we use the prediction and error distribution as a proxy of original conditional probability to address the first challenge. The hypothesis is that prediction and prediction error of V_i are conditionally independent given condition variables $V_\pi^{(i-1)}$. So that distribution of V_i given $V_\pi^{(i-1)}$ can be separated into prediction function and error distribution for continuous variables and can be modeled as categorical distribution for discrete variables as following,

$$V_i = \begin{cases} h_i(V_\pi^{(i-1)}) + \epsilon_i & \text{if } V_i \text{ is continuous} \\ Cat_i(K_i, \overrightarrow{P_i}(V_\pi^{(i-1)})) & \text{if } V_i \text{ is discrete} \end{cases} \quad (1)$$

where h_i is the prediction function which input is $V_\pi^{(i-1)}$ and output is V_i, ϵ_i is the distribution of prediction error that is also determined by $V_\pi^{(i-1)}$, and Cat_i is the proxy categorical variable which is determined by $V_\pi^{(i-1)}$, K_i is the value number of V_i. Here, we emphasise that the prediction function is not an assignment mechanism. It only represents a bi-directional equation relationship rather than a causal relationship. So, our hypothesis is not inconsistent with the causal diagram and causal assumptions.

For specific value v_i and $v_\pi^{(i-1)}$, we can calculate probability or density,

$$Pr(v_i|v_\pi^{(i-1)}) = \begin{cases} f_{\epsilon_i}(v_i - h_i(v_\pi^{(i-1)})) & \text{if } V_i \text{ is continuous} \\ p_i(v_i|v_\pi^{(i-1)}) & \text{if } V_i \text{ is discrete} \end{cases} \quad (2)$$

where f_{ϵ_i} is the probability density function of ϵ_i and p_i is probability of $V_i = v_i$ given $V_\pi^{(i-1)} = v_\pi^{(i-1)}$. For example, figure 6 is the ICE Graph of query $P(Y(T))$ in ADMG napkin.

From equation 2, we can transform the data distribution specification and parameterization into prediction (such as regression and classification) problems. The prediction performance in high dimension has been well handled in many fields under IID assumption based methods like deep learning Pouyanfar et al. [2018]. As for prediction error distribution, it is often to be Gaussian-like distribution, which is more easier to parameterize than data distribution. The number of possible ADMG is enormous, but the number of corresponding intervention expression graphs is far smaller for certain estimand. So learning ICE Graph can be helpful to ease the misspecification problem.

4.2.2 Learning

Two parts need to be determined in our ICE Graph: the intervention and the association parts. The learning goal of association parts is to improve predictors' performance under IID assumption and traditional matrices (such as F1-score, etc.). In comparison, the intervention part is to get an unbiased estimation of estimand with acceptable variance from predictors. Usually, the intervention part can be made directly by symbolic algorithms like IDC, but it can also be built by assumptions in potential outcome framework and even pure neural parametric methods Xia et al. [2021]. Association learning is to learn the prediction model and the confidence distribution of leaf nodes, which will influence the variance of the final estimation result.

Algorithm 1 is our learning algorithm for ICE Graph. There is two-step learning: intervention expression learning and association learning. The intervention expression learning transforms an estimand into estimations and creates predictors for the association layers. Then we will model the predictors and learn them from the data. Finally, we give those predictors in the association layer and the structure of intervention part.

Algorithm 1: Intervention Confidence Expression Graph (ICE Graph) Learning

Input : Data set $\mathcal{D}$, ADMG $\mathcal{G}$, Query $Q(Y, T, C)$
Output : ICE Graph G with Predictor set $\mathcal{P} = \{(f_i, \epsilon_i)\}_{i=1}^n$

$G := \textbf{IDC}(\mathcal{G}, Q)$;
$G := \textbf{Simplify}(G)$;
$\mathcal{P} := \textbf{LearnPredictor}(G, \mathcal{D})$;
return $G, \mathcal{P}$

4.2.3 Inference

There is still one technical challenge to inference likelihood: the integration or summation calculation. Here, in our inference algorithm 2, the Monte-Carlo method is used to put the marginal operator on distribution to attain the likelihood of such query.

Similarly to the completeness of 2, algorithm 1, 2 is complete to calculate any identifiable intervention query's likelihood.

Theorem 2. *Algorithm 1,2 is sound and complete to compute likelihood of any identifiable estimand $Q(Y, T, C)$ in any semi-Markovian model with ADMG G.*

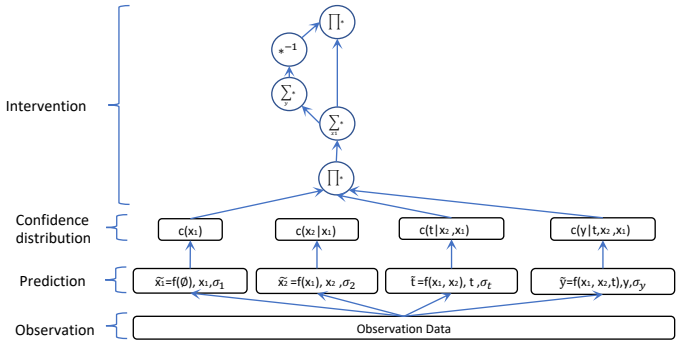


Figure 6: ICE Graph of napkin

Algorithm 2: Intervention Confidence Expression Graph (ICE Graph) Inference

Input : Sample $x = \{y, t, c\}$, ICE Graph G with Predictor set $\mathcal{P} = \{(f_i, \epsilon_i)\}_{i=1}^{n}$

Output : likelihood p for sample x

$a := Root(G)$;
if $a = \prod(*)$ **then**
  return $\prod_i$ **Inference**$(x, sub_i(G, a))$;
else if $a = (*)^{-1}$ **then**
  return **Inference**$^{-1}(x, sub(G, a))$;
else if $a = \sum_{V_C}(*)$ **then**
  $X = repeat(x, len(V_C))$;
  $X.V_c = V_C$;
  return $\sum_i$ **Inference**$(X_i, sub(G, a))$;
else if $a = \int_{V_c}(*)dV_c$ **then**
  $V_c =$ **UniformSampling**(V_C);
  $X = repeat(x, len(V_c))$;
  $X.V_c = V_c$;
  return $\sum_i$ **Inference**$(X_i, sub(G, a))$;
else if *a is leaf* **then**
  return $\epsilon_a(x)$;

Proof. For any identifiable estimand in any semi-Markovian model, the intervention part can always be built due to theorem 1 and completeness of IDC. From the law of large numbers, we know that the Monte-Carlo method in algorithm 2 is unbiased for any query. So, algorithms 1,2 can learn from observational data and calculate the likelihood of any identifiable estimand. □

For example, $P(Y(T))$ is often the interest estimand in reality. The expected outcome of $Y(T)$ can be estimated with slight structure bias and controlled variance from both predictors and structure of ICE Graph as algorithm 3.

The intervention part of the confidence graph can deal with any structure bias that was introduced by causal structure among variables in real world, such as confounding bias, selection bias, M-bias, mediation bias, once we define the

Algorithm 3: Treatment Effect Estimation Algorithm based on ICE Graph

Input : individual covariate x; ICE Graph G with Predictor set $\mathcal{P} = \{(f_i, \epsilon_i)\}_{i=1}^{n}$ for Query $Q(Y, T, \emptyset)$

Output : maximum likelihood outcomes of $Y(0)$ and $Y(1)$

$P_0 = \emptyset, P_1 = \emptyset$;
$S = \{(t, y)\} \leftarrow$ **UniformSampling**$((T, Y))$;
for $(t, y) \in S$ **do**
  $p \leftarrow$ **Inference**$(G, (y, t, x))$;
  if $t = 0$ **then** $P_0 = P_0 \cup \{(p, y)\}$;
  else $P_1 = P_1 \cup \{(p, y)\}$;
end
$E_x(Y(0)) = \arg_y \max p \in P_0$;
$E_x(Y(1)) = \arg_y \max p \in P_1$;

estimand that we want to estimate in the situation of interventionism causality and give ADMG candidates. The generalization ability of out-of-distribution is attained only from intervention part which is built by identification strategies. Also, intervention part may affect the variance of estimation result.

The association layer is one source of estimation variance. Predicting more accurately often introduce more accurate intervention estimation. The requirement of predictors $(P(V_i|V_i^{(i-1)}))$ is to get smaller variance, which means our prediction model should be as accurate as we can and keep generalization in the IID data set by train/valid/test splitting. In some cases, the counterfactual problem can be seen as missing data problem in observation layer if we add all potential outcome (such as $Y(1), Y(0)$) as endogenous variable of the semi-Markovian model.

Actually, neural methods Xia et al. [2021] and mixed methods Lee and Bareinboim [2021] can also be used for identification to create such an ICE Graph.

4.2.4 Confidence of Predictor

The σ of error distribution can be learned by variational and reparameterization methods Kingma and Welling [2013] where prior of σ will be set to sample variance of prediction error. Based on confidence graph, we can calculate any likelihood of $P(y(t)|x)$ given x, y, t.

There are mainly three advantages of our algorithms. First, our estimator is unbiased, without any structure bias (confounding, selection, M-bias, mediation), and efficiently analyses the variance's source. The unbiased estimation is from identification and sampling. The Second is that the estimator in our algorithm has better interpretability and reconfigurability. Every predictor has a specific meaning which indicate a conditional probability. They can be reused for other estimand due to the information that include in the predictor. Third, we transform the problem of learning data distribution of real world into adjusted prediction task with confidence distribution where input is $V_\pi^{(i-1)}$, and output is V_i.

5 ROBUSTNESS METRICS

5.1 PARAMETRIC INTERVENTIONS ROBUSTNESS OF ESTIMATORS

Now, we consider the matrix of parametric interventions robustness Wang et al. [2021a] in the semi-Markovian model for causal effect prediction. To measure different aspects of out-of-distribution generalization ability, we define exogenous robustness, endogenous robustness for the semi-Markovian model respectively, based on independent mechanism assumption. This assumption allows us to rewrite the assignment function $v := f_v(U_v, Pa_v)$ into $v := f_v(g_v^1(u_v^1), ..., g_v^k(u_v^k), g_v^1(pa_v^1), ..., g_v^h(pa_v^h))$. We call $g_v^i(u_v^i)$ outer mechanism (OM), $g_v^j(pa_v^j)$ inner mechanism (IM),f_v composition mechanism (CM), and u_v outer distribution (OD).

Definition 4. *Exogenous robustness: denote S to be parameters space of $g_V(U_V)$, S_1 to be parameters of semi-Markovian model M_1 in learning domain, and unknown transformation C is defined in space S. Denote E to be an estimator of treatment effect. The performance of E in the testing domain M_2 which parameters of $g_V(U_V)$ are $C_U(S_1)$ and other parameters are fixed, reflects exogenous robustness of E.*

Similarly, endogenous robustness can be defined by a transformation C_V in parameters space of $g_V(Pa_V)$.

In our ICE Graph, mechanisms that generate the data and statistics are disentangled into independent components. The intervention domain keeps invariant for parametric interventions, although the prediction domain is shifted due to mechanism changes. Such invariance helps us perform better in an unknown situation that was induced by transformation C and modularly reconfigure the model in a new environment.

6 COMPARISON AND CONNECTION WITH OTHES

The traditional potential outcome framework to deal with causal questions is to propose estimand from interest problem and manually factorize the estimand into estimations by assumptions or independence. Then build a model and learn from data to attain estimations for answering the estimand. However, such causal inferences are very tricky, and it is challenging for a human to factorize general estimand from hypotheses when there are more than three class endogenous variables except for cases with specific patterns. Specific examples of such methodology has been introduced in the related works.

Another direction to answer the causal query is pure neural methods Papantonis and Belle [2021]Xia et al. [2021]. The basic idea is to use truncated distribution or causal mechanism assumptions as regularization terms or constrain terms. Xia et al. [2021] is trying to answer the causal questions by using pure neural methods end-to-end, and they prove the existence of such a pure neural method which is complete. However, they didn't give a methodology for finding such function forms as data prior for different data distribution. Also, the method that uses symbolic identification judgement first and then estimation is indicated in this paper, but the core function of symbolic identification is to factorize the intervention query into smaller observation queries for building intervention expression graph if the query is identifiable, and return hedge which wit the non-identification of such query, rather to judge a query with causal diagram is answerable or not. The advantages of symbolic identification are not leveraged very much. Another weakness of their algorithm is that whenever we change the value of variable in query, we have to learn the model again, which inducing low re-usability of existed models and make it very hard to be applied to continuous cases.

Causal representation learning is to find causal variables from high dimension space. Structural decoder Leeb et al. [2020] is to create hierarchical coders, which is similar to our idea is of learning topological models in the association layer firstly. The difference is that our variables and topological order is presumed, and their topological order and variables are learned from reconstruction process. However, identification of query in diagram is not considered in the structural decoder which is very important to remove structure bias inside those coders for answering arbitrary intervention query. There are also other disentangled methods to create disentangled representation, such as CausalVAE Yang et al. [2020], stable learning Zhang et al. [2021],

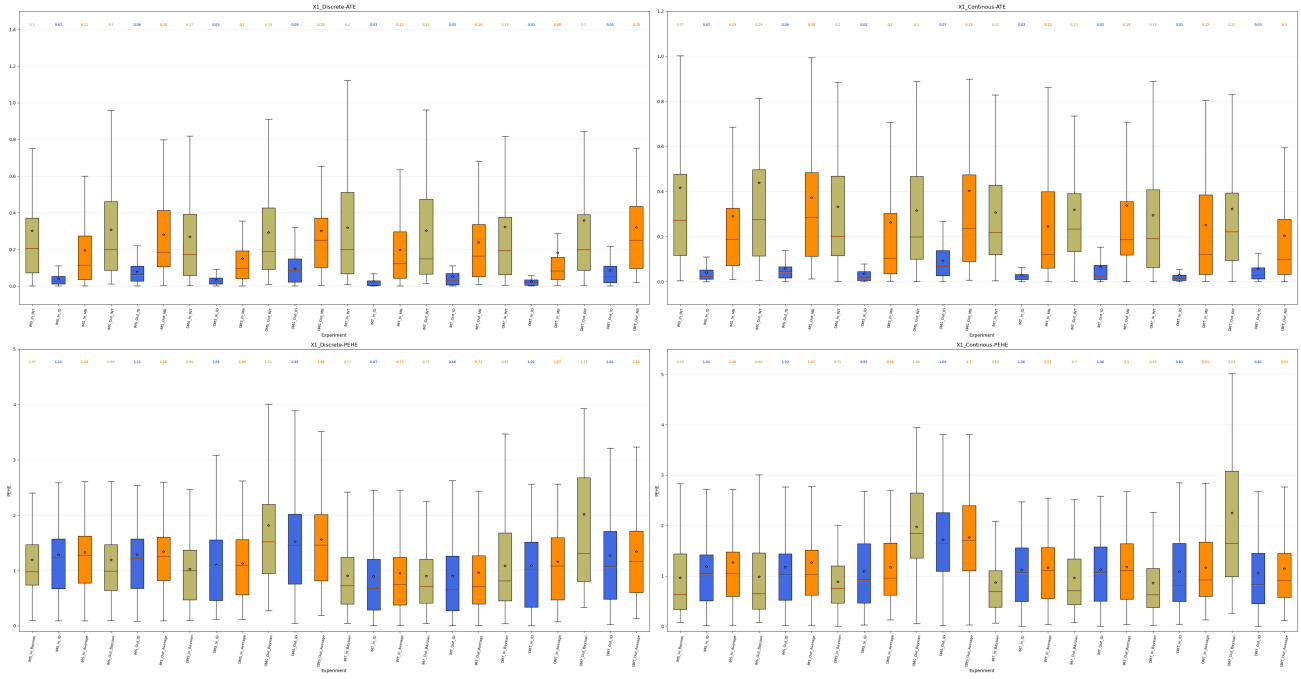


Figure 7: OOD experiment error. Star is median value. Red line is average value.'S' means mechanism shifting, and 'T' is random transformation of mechanism.

and IP-IRM Wang et al. [2021b]. Our learning algorithm 1 requires causal structure to build ICE Graph, so causal discovery Xie et al. [2020] will also be helpful.

7 EXPERIMENT

The experimental properties we are interested in about our model and algorithm 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.

In our experiment, we use the linear napkin 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 framework. For every setting, we run 50 independent experiments to evaluate the result where there are 1000 samples totally in each experiment. Other experiment details can be found in Appendix C. Although nonlinear model is not used in our experiments, it can still work if there are nonlinear predictors and environments.

Figure 7 shows the experiment results. We should notice that in-sample testing is not IID testing due to the missing counterfactual data, and our out-sample testing is under those parametric interventions. In unbiasness testing, our estimations are more unbiased than MR Freedman [2008] and INT Lin [2013] from ATE estimation result in both discrete and continuous cases. Considering estimation variance, it got better performance when outer mechanisms are changed.

8 CONCLUSION

In this paper, we propose an auto estimator framework for arbitrary identifiable estimand. It can combine the IID generalization ability of deep learning and robustness from identification strategies which is the core of causal inference. We test our idea in both discrete and continuous cases. It has little bias in all our settings and has little variance when outer mechanisms are transformed to unknown situation.

There are still many things that are needed to be improved. For example, summation and integration calculation will cost many resources and mechanism changing data may be difficult to attain in reality.

Finally, to solve the challenges in reality (such as individual treatment evaluation, and auto science), causal inference should combine with deep learning to better utilise the computing resources and big data, so that we can narrow the causal bound and reduce variance further. At the same time, we have to keep the critical idea of causal inference, such as cross-layer identification over Pearl's Causal Hierarchy.

References

Onur Atan, James Jordon, and Mihaela van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pages 884–895. PMLR, 2020.

David A Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2): 180–193, 2008.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, page 217–224, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Sanghack Lee and Elias Bareinboim. Causal identification with matrix equations. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Felix Leeb, Yashas Annadani, Stefan Bauer, and Bernhard Schölkopf. Structural autoencoders improve representations for generation and transfer. *CoRR*, abs/2006.07796, 2020. URL https://arxiv.org/abs/2006.07796.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

Ioannis Papantonis and Vaishak Belle. Closed-form results for prior constraints in sum-product networks. *Frontiers in Artificial Intelligence*, 4, 2021.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5), sep 2018. ISSN 0360-0300. doi: 10.1145/3234150. URL https://doi.org/10.1145/3234150.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002a.

Jin Tian and Judea Pearl. *Studies in Causal Reasoning and Learning*. PhD thesis, 2002b. AAI3070088.

Santtu Tikka and Juha Karvanen. Identifying causal effects with the r package causaleffect. *arXiv preprint arXiv:1806.07161*, 2018.

Benjie Wang, Clare Lyle, and Marta Kwiatkowska. Provable guarantees on the robustness of decision rules to causal interventions. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4258–4265. ijcai.org, 2021a. doi: 10.24963/ijcai.2021/585. URL https://doi.org/10.24963/ijcai.2021/585.

Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34, 2021b.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. 2021.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in Neural Information Processing Systems*, 33:14891–14902, 2020.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *CoRR*, abs/2004.08697, 2020. URL `https://arxiv.org/abs/2004.08697`.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, 2021.

A ADMG IDENTIFICATION OF THREE VARIABLES

We divide causal diagrams with same directed edge into one sets, then we have 25 sets (DAG number) in which there are 8 ADMG in each set (totally 200 ADMG).

First, there are **128** ADMG (16 sets) where there are no directed path from T to Y. From rule of deletion of action, we got $P(Y(T)) = P(Y)$.

Second, for **32** ADMG which include both directed edge $T \to Y$ and birected edge $T \leftrightarrow Y$, they are not identifiable because they include such 'bow arc'.

Third, consider Y-rooted C-tree. There are directed edge $T \to Y$, and birected path $T \leftrightarrow X \leftrightarrow Y$, but not include birected edge $T \leftrightarrow Y$. Because T is in this tree, so those **4** ADMG include such a Y-rooted C-tree is not identifiable.

Forth, consider DAG $T \to X \to Y$, there are **5** ADMG with such DAG is not identifiable.

From rule of action/observation exchange, there are **24** ADMG is identifiable, and we got $P(Y(T)) = P(Y|T)$.

From back-door directly, there are **6** ADMG is identifiable and we got $P(Y(T)) = \sum_x P(Y|T,x)P(x)$.

From front-door directly, there are **1** ADMG is identifiable and we got $P(Y(T)) = \sum_x P(x|T) \sum_t P(Y|t,x)P(t)$.

So, finally, there are **128** ADMG whose identification results are $P(Y(T)) = P(Y)$, **41** ADMG that is not identifiable, **24** ADMG whose identification results are $P(Y(T)) = P(Y|T)$, **6** ADMG whose identification results are $P(Y(T)) = \sum_x P(Y|T,x)P(x)$, **1** ADMG whose identification results are $P(Y(T)) = \sum_x P(x|T) \sum_t P(Y|t,x)P(t)$.

B ADMG NUMBER

ADMG is determined by DAG with additional birected edges. The increasing ratio of DAG number is $O(3^N)$, and each DAG will have $2^{\frac{N(N-1)}{2}}$ possible birected edges combinations. So the ADMG number increases with the ratio $\Theta(2^{N^2})$.

C EXPERIMENT SEETING

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 X2 and Y are continuous variables. X1 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 8 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.

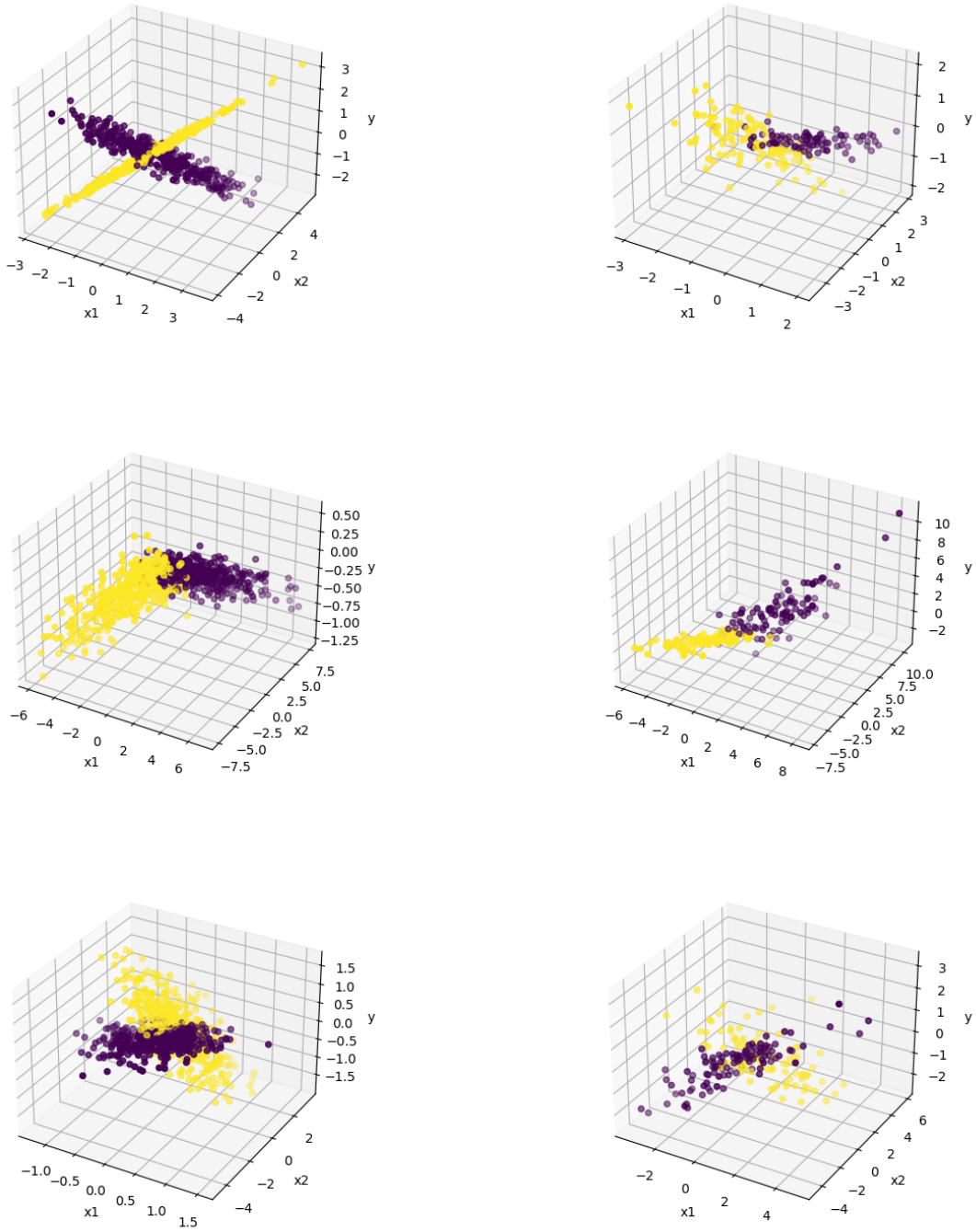


Figure 8: Some samples of generated data