Journal of Machine Learning Research (2021) 1-23 Submitted 5/21; Published 8/21

Estimating individual-level optimal causal interventions combining causal models and machine learning models

Keisuke Kiritoshi K.KIRITOSHI@NTT.COM
NTT Communications.
Tokyo, Japan

Tomonori Izumitani TOMONORI.IZUMITANI@NTT.COM
NTT Communications.
Tokyo, Japan

Kazuki Koyama KAZUKI.KOYAMA@NTT.COM
NTT Communications.
Tokyo, Japan

Tomomi Okawachi T.OKAWACHI@NTT.COM
NTT Communications.
Tokyo, Japan

Keisuke Asahara KEISUKE-ASAHARA@BIWAKO.SHIGA-U.AC.JP
Shiga University
Shiga, Japan

Shohei Shimizu SHOHEI-SHIMIZU@BIWAKO.SHIGA-U.AC.JP
Shiga University
Shiga, Japan

Editor: Thuc Le, Jiuyong Li, Greg Cooper, Sofia Triantafyllou, Elias Bareinboim, Huan Liu, and Negar Kiyavash

Abstract

We introduce a new statistical causal inference method to estimate individual-level *optimal causal intervention*, that is, to which value we should set the value of a certain variable of an individual to obtain a desired value of another variable. This is defined as an optimization problem to minimize the error between a desired value and the value that would have been attained under the setting for the individual. To solve the optimization problem, we first train a machine learning model to predict the value of an objective variable and then estimate the causal structure of variables. We then combine the machine learning model and causal structure into a single causal model to estimate counterfactual value of the predicted objective variable. This is effective in achieving a more accurate estimation of individual-level optimal causal intervention. We further propose a gradient descent algorithm to compute the optimal causal intervention. Our method is generally applicable to continuous variables that are linearly and non-linearly related. In experiments, we evaluate the effectiveness of our method using artificial data generated by non-linear causal structures and real data.

Keywords: Causal Discovery, Optimal Intervention, Machine Learning, Counterfactual, Causal Inference

1. Introduction

Analyzing causal effects for making an action is an important problem in many domains. For example, a doctor decides the most effective treatments for patients from the patients background, and an operator controls various equipments in a chemical plant to improve the quality of manufactured products on the basis of the causal relationships of measured sensor data. Randomized experiments are conducted to investigate causal effects but they can be costly and unethical. We can estimate population-level causal effects from a variable to another variable on the basis of a statistical calculation with intervention on known causal structures and observed data (Pearl, 2009), but situations in which the causal structures are identified are rare. In such situations, statistical causal discovery methods (Shimizu et al., 2006; Peters et al., 2014; Mooij et al., 2016) for estimating causal structures from observed data are effective.

Individual-level causal effects (Pearl, 2009) are important to examine the cause of a phenomenon or make decisions for not a population but individuals. For example, a plant operator will find that temperature should have been raised to achieve a better quality value of a certain product sample. Pearl (Pearl, 2009) defines individual causal effects and proposes a method for estimating counterfactuals based on noise values, i.e. exogenous variable values of individuals in structural causal models.

In many proceeding researches, a control variable for an individual causal effect is a binary variable such as whether a doctor should administer treatment, but what if the variable is continuous such as the example of the plant operator? To achieve a desired quality value of a certain target product by controlling temperature, the operator needs to know what a concrete value of the controlled variable is. We call this value *individual-level optimal causal intervention* and introduce a new problem setting of estimating *individual-level optimal causal intervention* to obtain a given desired value of an objective variable from observational or non-experimental data. In this problem setting, correct causal structures may have non-linear or linear relationship between variables, and the intervened variable is continuous.

There are two difficulties to estimate individual-level optimal causal intervention. First, to estimate counterfactuals, we have to estimate structural equation models from observed data and identify noise values of individuals (Pearl, 2009). When the true causal relationship is non-linear, we can estimate structural equation model (SEM) by assuming some models which are identifiable such as additive noise model (Hoyer et al., 2008a) or Post Non-linear Model (Zhang and Hyvarinen, 2012). However, estimating these SEM is costly because in existing methods we have to train some non-linear models for each pair-wise variable to identify the causal structures. (Peters et al., 2014; Mooij et al., 2016). Moreover, we may not obtain enough estimation accuracy of SEM because of validity of model assumptions and accuracy of prediction models. Second, there are some proposed methods to estimate optimal causal intervention for total effects based on interventional data and observational data (Aglietti et al., 2020b,a). However, methods which focus on individual-level causal effects based on observational data are still not proposed.

In this paper, to estimate individual-level causal effects, we propose a method of combining a prediction model and an estimated causal structure into a single causal model based on the framework of Blöbaum et al. (Blöbaum and Shimizu, 2017). We also propose

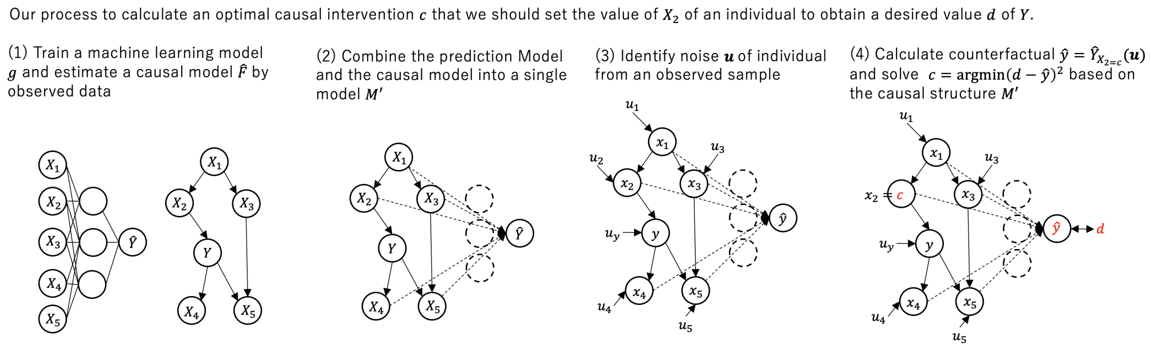

Our process to calculate an optimal causal intervention c that we should set the value of X_2 of an individual to obtain a desired value d of Y.
(1) Train a machine learning model g and estimate a causal model $\hat{F}$ by observed data
(2) Combine the prediction Model and the causal model into a single model M'
(3) Identify noise $\boldsymbol{u}$ of individual from an observed sample
(4) Calculate counterfactual $\hat{y} = \hat{Y}_{X_2=c}(\boldsymbol{u})$ and solve $c = \mathbf{argmin}(d - \hat{y})^2$ based on the causal structure M'
X_1
X_2
X_3
X_4
X_5
$\hat{Y}$
Y
x_1
x_2
x_3
x_4
x_5
y
$\hat{y}$
u_1
u_2
u_3
u_4
u_5
u_y
$x_2 = c$
d

Figure 1: The Overview of our proposed method

an optimization problem to minimize an error between an estimated individual-level causal effect and the desired value and develop a gradient descent algorithm to optimize the minimization problem for a non-linear prediction model. As experiments, we validate accuracy of counterfactuals and optimal causal interventions with our proposed method and some baselines.

The contributions of this paper is below:

- We propose a new problem setting of estimating *individual-level optimal causal intervention* and formulate it as a minimization problem.
- We propose estimate individual-level causal effects by combining a machine learning model and an estimated causal structure into a single causal model.
- Based on the combination of the prediction model and the causal structure, we propose a method to calculate individual-level optimal causal intervention by optimizing the proposed minimization problem with a gradient descent algorithm.

2. Related Works

Pearl (Pearl, 2009) defines the concepts of counterfactual and its calculation from the structural equation model (Pearl, 2009; Bollen and Hoyle, 2012). An individual is identified by the values of its exogenous variables and its counterfactual is calculated by the structural equation models with the exogenous variables. Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al., 2006) is a structural equation model based on the identifiability of linear DAG (directed acyclic graph) under non-Gaussian noises. Additive Noise Model (Hoyer et al., 2008a) is the model represented by arbitrary non-linear functions and added noise to identify non-linear causal relation. Post-Nonlinear Causal Model (Zhang and Hyvarinen, 2012) generalize Additive Noise Model and apply a non-linear function to the sum of variables and noise.

In many practical cases, we are not given any structural equation models at first. In such cases, we have to estimate the causal structures in the observed data through causal discovery methods under these assumptions. ICA-LiNGAM (Shimizu et al., 2006) is a causal discovery method with FastICA Algorithm(Hyvarinen, 1999) under LiNGAM assumption. DirectLiNGAM (Shimizu et al., 2011) extend ICA-LiNGAM algorithm to improve its convergence. For the assumption of Additive Noise Model, there are some methods to identify causal direction of pair-wise variables on the basis of measuring independency between predicted residuals from regression of variables and the explanatory variable (Mooij et al., 2016). As the measurements of independence, Hilbert-Schmidt Independence Criterion(Hoyer et al., 2008b) and differential Shannon entropy(Kpotufe et al., 2014) are used. RESIT (Peters et al., 2014) is a method to estimate causal structures of multiple variables under Additive Noise Model. It repeat procedures which regress one variable on the other variables and compare a dependence score of the predicted residual and input variables of the regression model to identify causal order.

In this paper, we use these causal discovery methods to estimate structural equation models from observed data to calculate individual-level causal effects in Pearl's sense.

To calculate individual-level causal effects when we intervene on binary variable, there are some machine learning methods without (estimating) structural equation models. Shalit et el. (Shalit et al., 2017) utilize neural network models that directly predict controlled and intervened values of objective variable, respectively. This model train to make pre- and post-intervention distributions of intermediate representations of explanatory variables close to obtain balanced representations of explanatory variables on randomized experiments. Similarly, Guo et al. use Graph Neural Networks for causal inference of social networks (Guo et al., 2020) and Alaa et al (Alaa and van der Schaar, 2017) propose a method based on multi-task Gaussian processes to estimate individualized treatment effects. These methods use Rubin-Neyman potential outcomes framework (Rubin, 2005) with binary treatment values. On contrast, our method focus on continuous variables as targets of intervention and not only estimate individual-level causal effects but also estimating optimal causal intervention. Moreover, we use estimated causal structures to calculate individual causal effects based on Pearl's definition.

Some proposed methods calculate optimal causal intervention for total effects. Aglietti et al. proposes Causal Bayesian Optimization framework, combining Bayesian optimization and causal inference to explore optimal intervention for total effects (Aglietti et al., 2020b). They also construct a function with a method for representing a function with the value of the intervention as the input and the total effect of the intervention as the output through multi-task Gaussian process model (Aglietti et al., 2020a), through which they calculated optimal causal intervention as well. On contrast, our method calculate an optimal intervention for individual-level causal effect and to explore the optimal intervention, we use a gradient descent algorithm.

3. Background

We assume that data-generating processes can be graphically represented as a direct acyclic graph (DAG), where the causal influence of random variable X_i on another variable X_j is indicated with an arrow between two variables.

3.1 Additive Noise Model

Structural equation models (Pearl, 2009; Bollen and Hoyle, 2012) are used in a typical representation of linear causal structures in causal discovery when the observed data are continuous-valued. To generalize these models for non-linear causal relationships, Hoyer et al. proposed an additive noise model (Hoyer et al., 2008a) defined by the observed value x_i associated with X_i and independent additive noise u_i as follows:

$$x_i = f_i(\boldsymbol{x}_{pa(i)}) + u_i \tag{1}$$

where f_i is an arbitrary function of each variable and $\boldsymbol{x}_{pa(i)}$ means the set of observed values associated with the parent variables of X_i in the DAG. Noise u_i has arbitrary probability densities. We define the observed value as $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and each element associate with $X_1, X_2, \ldots, X_n$. Each variable is affected by independent exogenous variables $U_1, U_2, \ldots, U_n$. Their observed values are also defined as $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$. As a special case, this model includes a linear non-Gaussian acyclic model (LiNGAM) (Shimizu et al., 2006) if all the f_i are linear and the distributions of all u_i are non-Gaussian. When there is a linear causal influence, its causal strength between X_i and X_j is described by b_{ij} and its causal relationship of all variables is described by matrix $\mathbf{B}$.

3.2 Interventions and Counterfactuals

DAGs can show when an X_i has a fixed constant value c and how other variables in the DAG change. To analyze this behavior, we introduce *intervention* as a *do* operator $do(X_i = c)$ (Pearl, 2009). This means fixing X_i to a constant c. Pearl defines the probability of the variable Y after intervention of X_i as $P(Y|do(X_i = c)$ and its expected intervention effects, called *total effects*, as

$$\mathbb{E}[Y|do(X_i = c)]. \tag{2}$$

The total effects show the effects on the population level not on the individual level. Pearl introduced the concepts of counterfactuals (Pearl, 2009, 2018) to analyze such individual causal effects. To calculate a counterfactual, the structural equation model M need to be known and the value of observed variable $\boldsymbol{x}$ and exogenous variables U are necessary to be obtained. The value of the exogenous variables $U = u$ corresponds to a certain individual or individual "situation". Therefore, counterfactual "Y would be y had X been c in situation $U = u$", that is $y = Y_{X=c}(\boldsymbol{u})$ on M, is calculated with the following procedure:

1. Obtain $\boldsymbol{u}$ by applying observed value $\boldsymbol{x}$ to model M.

2. Replace the equations corresponding to variables in set X with the equations $X = c$ in M, and obtain modified model M_x.

3. Use M_x to compute counterfactual $Y = y$ from $\boldsymbol{u}$.

4. Problem Setting

In this paper, our challenge is to meet the following situation:

- Observations of $X_1, X_2, \ldots, X_n$ and Y are given but true causal structures are unknown.
- Each of X_i and Y is a continuous variable and is given by an user.
- The optimal intervention c that results in $Y_{X_i=c}(\boldsymbol{u}) = d$ should be identified.

This situation is common in many system control cases. For example, in a manufacturing system, a certain Y, which is the product quality (e.g. product hardness), is observed as y, and other variables $X_1, X_2, \ldots, X_n$ are also observed (e.g. temperature and pressure) as $x_1, x_2, \ldots, x_n$. They may have causal influences on Y and themselves, but the causal structure and causal strength are unknown in general. We suppose that $Y = d$ is the most desirable value. In this situation, the question is how to control X_i given by an user to change Y into d for individual products. Controlling x_i to the optimal value means calculating such c that achieves $Y_{X_i=c}(u) = d$ because X_i has a causal relationship with other variables including Y. Moreover, because *optimal interventions* differ among individual products, we need to estimate not the total intervention effects based on all products ($\mathbb{E}[Y|do(X_i = c)]$) but individual-level causal effects based on individual products ($Y_{X_i=c}(u)$).

Individual-level causal effects from observed data on the basis of intervening individuals have been studied (Shalit et al., 2017). However, intervened variables that they focused on have only binary values such as "whether the treatment should have been administered?", rather than continuous variables on our problem setting. Moreover, many machine learning based methods without causal structures estimate $\mathbb{E}[Y_1 - Y_0|\boldsymbol{u}]$[1] instead of $Y_{X_i=c}(u)$. In our problem setting, we have to estimate causal structure and values of exogenous variables to calculate counterfactual $Y_{X_i=c}(u)$.

Existing studies attempted to answer the question, "If we control a certain X_i, how many effects on the value of Y of an individual $\boldsymbol{u}$ are there?". In contrast, we attempted to answer the question, "When we need to change a certain Y of an individual $\boldsymbol{u}$ to d, how should we control a certain X_i?" We call this problem estimating *optimal causal intervention.* To the best of our knowledge, this problem has not been discussed in the causal inference field.

The optimal intervention framework of Blöbaum et al. (Blöbaum and Shimizu, 2017) estimates linear intervention effects on prediction models on the basis of the combination of prediction models and structural equation models. In their framework, c is calculated to explain how the output changes when we intervene on a certain X_i when the prediction model predicts d. This framework is defined as the following minimization problem $c = \underset{c}{\mathrm{argmin}}(\mathbb{E}[\hat{Y}|do(X_i = c)] - d)^2$, where $\hat{Y}$ denotes the output value of the prediction model. To calculate c, they introduced a procedure to estimate the total effects of all variables through the linear causal structure by connecting the causal structure and prediction model. They also introduce an algorithm to numerically solve c on the basis of the optimization problem.

However, their problem setting differs from ours. They estimate c for the output $\hat{Y}$ of the prediction model, but we need to estimate the optimal intervention for Y of the causal structure on the basis of the generation process. In addition, they focus on total effects $\mathbb{E}[\hat{Y}|do(X_i = c)]$, but we target individual-level causal effect $Y_{X_i=c}(\boldsymbol{u})$. While they

[1] Y_0 and Y_1 mean potential outcomes corresponding to intervention $t = \{0, 1\}$ defined by Rubin (2005).

proposed an algorithm for linear structural equation models and linear prediction models, our problem setting includes a non-linear causal structure.

This problem setting may need to be extended to time-series data for real-world applications. In this study, we deal with non-time-series data for simplicity.

5. Proposed Method

Fig 1 shows the overview of our method. In our problem "*how much intervention on a certain variable X_i of an individual $\boldsymbol{x}$ needs to be determined to obtain the desired value d of another variable Y ?*", we define the following formula:

$$c = \underset{c}{\operatorname{argmin}} (Y_{X_i=c}(\boldsymbol{u}) - d)^2. \tag{3}$$

Algorithm 1 and 2 shows the overall of our proposed method to solve this problem when estimated causal structures are represented by linear functions.

5.1 Estimate Individual-level Causal Effects

To solve the optimization problem of Formula (3), calculating $Y_{X_i=c}(\boldsymbol{u})$ is necessary. When the structural equation model is known, it is calculated with the procedure of counterfactual calculations by Pearl given in Section 3.2. If the causal structure is represented using an additive noise model (Formula (1)), counterfactual $Y_{X_i=c}(\boldsymbol{u})$ is calculated sequentially by

$$Y_{X_i=c}(\boldsymbol{u}) = \begin{cases} c, \text{ if } j = i, \\ x_j, \text{ if } X_j \text{ is a root variable,} \\ f_j(pa(X_j)_{X_i=c}(\boldsymbol{u})) + u_j, \text{ otherwise.} \end{cases} \tag{4}$$

where X_j is a random variable in a set of ancestors of Y and $pa(X_j)$ denotes the parent variables of X_j, f_j is an arbitrary function to generate X_j, and u_j is individual noise and calculated by $x_j - f_j(pa(x_j))$. LiNGAM is a special linear additive noise model, and we obtain the counterfactual by replacing $f_j(pa(X_j)_{X_i=c}(\boldsymbol{u}))$ with $\sum_{X_k \in pa(X_j)} b_{kj} X_{k X_i=c}(\boldsymbol{u})$. Intuitively, Formula (4) means that individual causal effects by intervention on X_i propagate all descendants of X_i along with the causal relationship.

If we can obtain a correct additive noise model, counterfactual $Y_{X_i=c}(\boldsymbol{u})$ can be calculated accurately but it is generally difficult because the data-generation process can follow another model such as the post non-linear model (Zhang and Hyvarinen, 2012). A function f_j is generally not known, so we have to approximate it by using non-linear regression models such as a machine learning regression model and a highly accurate regression model is necessary to calculate an accurate counterfactual.

To calculate $Y_{X_i=c}(\boldsymbol{u})$, we combine the estimated causal structures and machine learning regression model into a single causal model. This idea of using prediction models is based on supporting the accuracy of estimation of causal discovery methods by the predictive performance of machine learning models.

When the objective variable Y has descendants and the explanatory variables include the descendants, the prediction model has high regression accuracy. This is because the

Algorithm 1 Calculate Individual-level Optimal Intervention for Linear Causal Structure

```
Input: $\boldsymbol{x}, \mathbf{B}, f, \boldsymbol{u}, i, d$
$\boldsymbol{x}$: a vector of target sample, $f$: a prediction model,
$d$: desired value, $\mathbf{B}$: causal coefficient matrix,
$i$: the index of the intervened variable,
$\boldsymbol{u}$: noise vector of target sample.
$S \leftarrow \{1, 2, 3, \dots, n\} \backslash \{i\}$
$S \leftarrow$ RemoveIndicesOfRootVariables($\mathbf{B}$)
$\boldsymbol{\alpha} \leftarrow$ VectorOfZerosForEachVariable($\mathbf{B}$)
$\alpha_i \leftarrow 1$
$x_i, u_i \leftarrow 0$
while $S$ is not empty do
    $k \leftarrow$ GetNextIndex($S$)
    if $X_k$ has no parents in $S$ then
        $x_{tmp}, \alpha_{tmp} \leftarrow 0$
        for all parents of $X_k$ do
            $q \leftarrow$ GetIndexOfNextParent($X_k$)
            $\alpha_{tmp} \leftarrow \alpha_{tmp} + b_{kq}\alpha_q$
            if $q$ is not $i$ then
                $x \leftarrow x + b_{kq}x_q$
            end if
        end for
        $x_k \leftarrow x + u_k$
        $\alpha_k \leftarrow \alpha_{tmp}$
        $S \leftarrow$ RemoveIndex($S, k$)
    end if
    $k \leftarrow$ GetNextIndex($S$)
end while
$c \leftarrow$ BackwardFunction($\boldsymbol{x}, \boldsymbol{\alpha}, f, d$)      :Algorithm 2
Output: $c$
```

descendant of Y has an effect on Y including noise u_y. When calculating the counterfactual of Y on a certain sample $\boldsymbol{x}$, this information of u_y is essential to identifying individuals in accordance with Pearl 's calculation process of a conterfactual (Section 3.2). If estimated causal structures and their coefficients of causal effects are not accurate, this u_y is also not accurate. On the other hand, prediction models can use accurate u_y directly through the descendants of Y to regress y. Moreover, while considering generalization performance of causal discovery method is difficult, prediction models trained by machine learning techniques such as cross validation are expected to improve generalization performance. This is effective on real-world application like our problem setting. In Appendix A, we discuss the predictive performance of prediction models with and without descendants of objective variables.

We first train a regression model g, such as a linear regression or neural network, by explanatory variables $X_1, X_2, \dots, X_n$ and objective variable Y. Next, we estimate the causal

Algorithm 2 Calculate Optimal Intervention for Neural Network Model

```
Input: $\boldsymbol{x}, \boldsymbol{\alpha}, f, d$
$\boldsymbol{x}$: target sample, $f$: prediction model,
$d$: objective value, $\epsilon$: hyper-parameter
$c \leftarrow init()$, $L \leftarrow \infty$
while $L > \epsilon$ do
    $\boldsymbol{x}' \leftarrow \boldsymbol{x} + c\boldsymbol{\alpha}$
    $\boldsymbol{x}' \leftarrow$ RemoveObjectiveIndex($\boldsymbol{x}'$)
    $\hat{y} \leftarrow f(\boldsymbol{x}')$
    $L \leftarrow (d - \hat{y})^2$
    $c \leftarrow c - \eta \nabla L$
end while
return $c$
Output: $c$
```

structure $\hat{F}$ of all variables by using a causal discovery models such as an additive noise model or LiNGAM. We then define a causal structure $M'(g, \hat{F})$ by combining the causal structure of the prediction mechanism of g and estimated $\hat{F}$. As shown in Fig 1, in $M'(g, \hat{F})$, $\hat{Y}$ is introduced and all explanatory variables are parents of $\hat{Y}$.

As the approximation of $Y_{X_i=c}(\boldsymbol{u})$, we use $\hat{Y}_{X_i=c}(\boldsymbol{u})_{M'}$, that is, the value of $\hat{Y}$ had X_i been c for individual $\boldsymbol{x}$ on the causal structure $M'(g, \hat{F})$. The sequential calculation of counterfactuals on $\hat{F}$ of $M'(g, \hat{F})$ follows Formula (4) when $\hat{F}$ represents additive noise models or linear causal models. The calculation procedure on the causal relationship $X_j \rightarrow \hat{Y}$ depends on the g. If on g, the relationship of the explanatory variables Xs and variable $\hat{Y}$ is given by linear or non-linear regression functions following the function rule of an additive noise model (e.g. linear regression and Gaussian Process (Williams and Rasmussen, 2006)), we can calculate the counterfactual along with Formula (4). Feedfoward neural networks can follow Formula (4) because they can be reduced to causal structures in which input nodes and output nodes are directly connected (Chattopadhyay et al., 2019).

In the experiment section, we validate accuracy of individual-level causal effects estimated by our proposed method and some baselines.

5.2 Estimate Optimal Causal Intervention

To calculate individual-level *optimal causal intervention*, solving the optimization problem of Formula (3) is necessary. We consider the following three cases.

In the first case that g is a linear regression model and $\hat{F}$ is a causal structure estimated using a linear causal model (such as LiNGAM), we propose an extended algorithm of Blobaum et al. for individuals. This algorithm is described in Algorithm 1. The concept of the algorithm is to separate the effects of intervention and propagation of noise on the basis of Formula (4). The vector $\boldsymbol{\alpha}$ represents the individual causal effect on each variable by intervening on a certain variable $X_i = 1$, and $\boldsymbol{\alpha}$ is calculated by the propagation through $\hat{F}$. The vector $\boldsymbol{u}$ is noise calculated by $\boldsymbol{x} - \mathbf{B}\boldsymbol{x}$ and is propagated in $\hat{F}$ along with Formula (4). On the basis of $\boldsymbol{x}$ and $\boldsymbol{\alpha}$, we formulate the value of each variable affected by individual causal effects as $\boldsymbol{x} + c\boldsymbol{\alpha}$. When g is a linear regression model, we can easily calculate individual-level optimal intervention $\hat{c}$ by the following formula:

$$\hat{c} = \frac{d - \boldsymbol{w}^T\boldsymbol{x} - w_0}{\boldsymbol{w}^T\boldsymbol{\alpha}} \tag{5}$$

where $\boldsymbol{w}$ is the coefficient vector of g when w_y is 0.

On the other hand, in the second case that g is a feedforward neural network and $\hat{F}$ is a causal structure estimated using a linear causal model, calculating optimal intervention analytically is difficult. Therefore, in Algorithm 2, we propose a gradient descent algorithm to obtain an approximate solution c from g on the basis of the squared error (SE) between d and predicted value $\hat{y} = g(x')$, where x' is removed index of the objective variable from $\boldsymbol{x} + c\boldsymbol{\alpha}$.

Note that the accuracy of c involves the smoothness of the regression surface of the model, and when the surface is non-smooth, the solution reaches local minima. With a neural network, the activation functions of the model influence it. In particular, the rectified linear unit (ReLU) function has local linearity, and its regression surface is roughly piecewise linear with many transitions (Ghorbani et al., 2019). The Softplus function (Dugas et al., 2000) reduces the maximal curvature of the regression surface (Dombrowski et al., 2019). We compared the optimization loss and accuracy of c for activation functions.

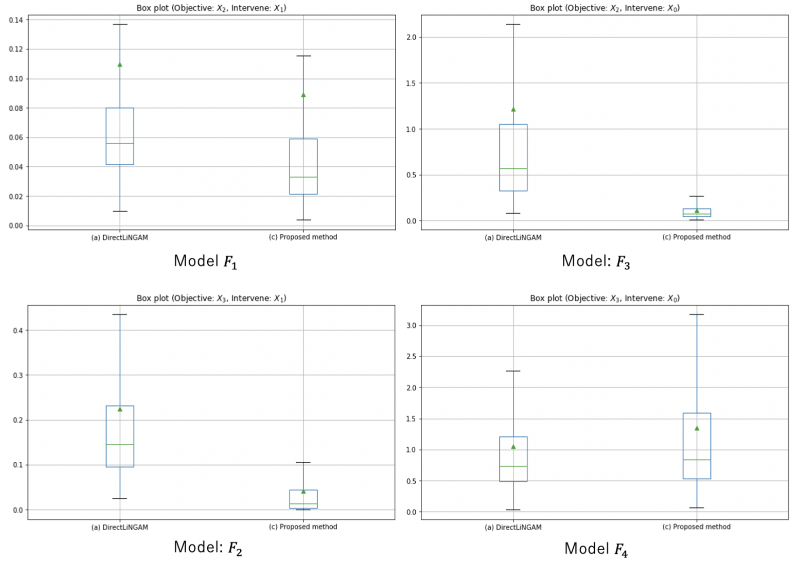
In the case that g is a feedforward neural network and $\hat{F}$ is a non-linear causal structure estimated using a non-linear causal model (such as the additive noise model), if all functions f_j of an additive noise model are differentiable such as neural networks, Formula (3) can be optimized with our gradient descent algorithm by considering the combination model M' as a single formula. In this study, we applied our method to cases (i) and (ii).

6. Experiments

We evaluate our method from two perspectives: estimated accuracy of counterfactual and individual-level optimal causal intervention. We applied our proposed method and several baselines to artificial data generated by various processes. We used the common preparations of the following steps. (1) Define structural equation model $F(X_1, \ldots, X_n, Y)$ and generate sample data D. Randomly separate $\boldsymbol{x}$ into training data D_{tr} and test data D_{ts}. The data D_{tr} and D_{ts} are normalized. (2) Estimate a causal structure by using DirectLiNGAM with the likelihood ratio (Shimizu et al., 2011; Hyvärinen and Smith, 2013) with D_{tr} and obtain estimated causal model $\hat{F}(X_1, \ldots, X_n, Y)$. (3) Train a prediction model $\hat{Y} = g(X_1, \ldots, X_n)$ with explanatory variables $X_1, \ldots, X_n$ and an objective variable Y by using D_{tr}.

6.1 Evaluate Individual-level Causal Effects Estimation

To evaluate accuracy of estimating individual-level causal effects, we introduce four causal models as the generative models F. In each model, noise $u_i\,(i = 1, 2, 3, 4)$ is generated by Laplace$(0, 0.05)$, and we generate data with sample size $10,000$ (training data: $8,000$ and



Model F_1

Model: F_3

Model: F_2

Model F_4

Figure 2: Evaluating accuracy of estimating individual-level causal effects. Boxplot of averaged squared error (SE) for each sample on structural equation models (lower score is better). (a) and (c) are estimated structural causal model and proposed method, respectively. Proposed method obtained lower scores in the case of generative models F_1, F_2 and F_3. The result of (b) baseline of only prediction model is omitted because (b) obtain higher SE score than (a) and (c) in all generative models (see Appendix B)

test data: $2,000$). We define four additive noise models (F_1, F_2, F_3, F_4) with various patterns of non-linearity and causal order of the objective variable Y as the following formulas.

$$F_1 = \begin{cases} X_0 = u_0 \\ X_1 = 0.6X_0 + u_1 \\ Y = f_y(0.8X_0 + 1.5X_1) + u_2 \\ X_3 = 0.5X_1 - 2.0Y + u_3 \end{cases} \tag{6}$$

$$F_2 = \begin{cases} X_0 = u_0 \\ X_1 = 0.6X_0 + u_1 \\ X_2 = 0.8X_0 + 1.5X_1 + u_2 \\ Y = f_y(0.5X_1 - 2.0X_2) + u_3 \end{cases} \tag{7}$$

$$F_3 = \begin{cases} X_0 = u_0 \\ X_1 = f_1(0.6X_0) + u_1 \\ Y = f_2(0.8X_0 + 1.5X_1) + u_2 \\ X_3 = f_3(0.5X_1 - 2.0Y) + u_3 \end{cases} \tag{8}$$

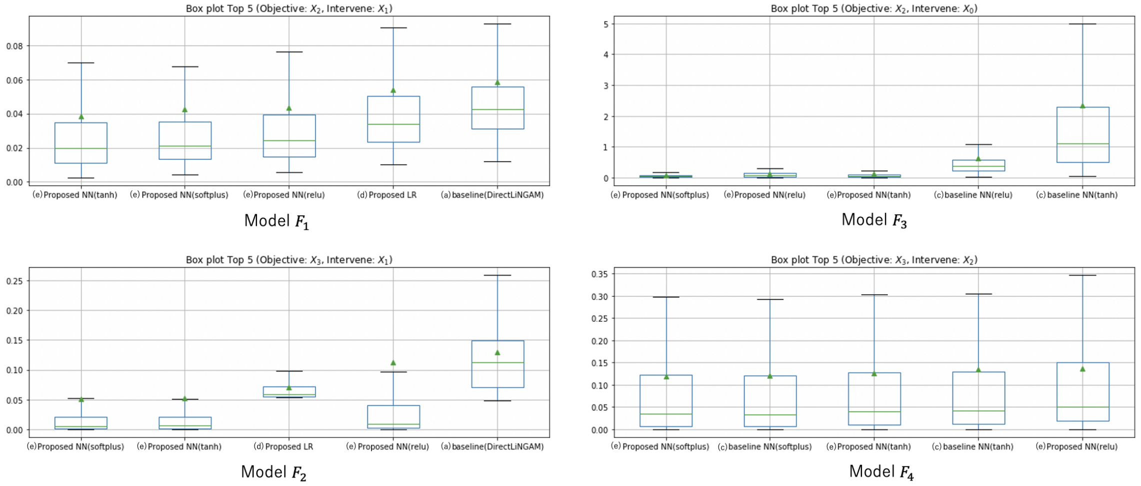


Figure 3: Evaluating accuracy of optimal causal intervention. Boxplot of averaged squared error (SE) for each sample on simple structural equation models and additive noise models (lower score is better). (a)-(e) are associated with experimental setting and we show only top five methods. Green delta points are MSE. Proposed NN obtained lower scores.

$$F_4 = \begin{cases} X_0 = u_0 \\ X_1 = f_1(0.6X_0) + u_1 \\ X_2 = f_2(0.8X_0 + 1.5X_1) + u_2 \\ Y = f_3(0.5X_1 - 2.0X_2) + u_3 \end{cases} \tag{9}$$

where f_y in F_1 and F_2 is a sigmoid function described as $f_y(x) = \frac{1}{1+e^{-\tau x}}$, and τ is the hyper-parameter to adjust non-linearity when x is near 0; we set $\tau = 10$ in the experiments. In F_3 and F_4, each non-linear function is $f_1 = \frac{1}{1+e^{-\tau x}}$, $f_2 = x + \sin(\pi x)$, and $f_3 = (x - 0.25)^3$. In F_1 and F_3, the explanatory variables include the descendant of Y, and in F_2 and F_4, Y is the last of the causal order. These two type structural models aim to evaluate effects of descendants of the objective variable and non-linearity of causal relationship on estimation accuracy of individual-level causal effects.

In the experiment, we compare values of counterfactuals calculated by generative models ($F_1 \ldots F_4$), baselines, and our proposed method. First, as ground truth, we do intervention on X_i with $c \in C$ ($C = \{-1.0, -0.5, 0, 0.5, 1.0\}$) and calculate counterfactual $Y_{X_i=c}(\boldsymbol{u})$ in D_{ts} for each generative model on the basis of the calculation procedure of individual-level causal effects in section 3.2. Second, we also calculate $\hat{Y}_{X_i=c}(\boldsymbol{u})$ by baselines and our proposed methods. The baselines are (a) calculating $\hat{Y}_{X_i=c}(\boldsymbol{u})$ from an estimated causal structure $\hat{F}$ directly; (b) estimating counterfactual using a prediction model g of MLP without $\hat{F}$ on the basis of Algorithm 2. The proposed method is (c) MLPs g combined with estimated causal structure $\hat{F}$. The detailed experimental settings and results are given in Appendix B.

The boxplot [2] of error score of individual-level causal effects of each sample is shown in Fig 2 . As evaluation scores, we calculate the average of squared error between ground truth and each method defined by $\frac{1}{|C|}\sum_{c\in C}(Y_{X_i=c}(\boldsymbol{u}) - \hat{Y}_{X_i=c}(\boldsymbol{u}))^2$. The result shows that out proposed method obtain lower average and median of error in the models F_1, F_2 and F_3. The baseline method demonstrates lower error score in F_4. We consider this phenomenon in discussion section.

6.2 Accuracy of Optimal Causal Intervention

To evaluate accuracy of estimating the optimal causal intervention of the proposed method, instead of giving c, we calculate individual-level optimal causl intervention $\hat{c}_i$ of a desired value d for sample $\boldsymbol{x}_i \in D_{ts}$ on the basis of baselines and proposed method. Second, we apply $\hat{c}_i$ to original generation process $F_1, \ldots, F_4$ on the basis of the counterfactual procedure and obtain predictive value $\hat{d}_i$ for each sample. We vary $d \in D$ ($D = \{-1.0, -0.5, 0, 0.5, 1.0\}$) and compare $\hat{d}_i$ and d. The detailed experimental settings and results are given in Appendix B.

The baselines are (a) estimating $\hat{c}$ from $\hat{F}$ directly on the basis of the counterfactual procedure; (b) Blöbaum et al.s' method; (c) calculating $\hat{c}$ by using a prediction model g without $\hat{F}$ on the basis of Formula (5) or Algorithm 2; (d) linear regression; and (e) MLPs (ReLU, tanh, Softplus) combined with estimated causal structures.

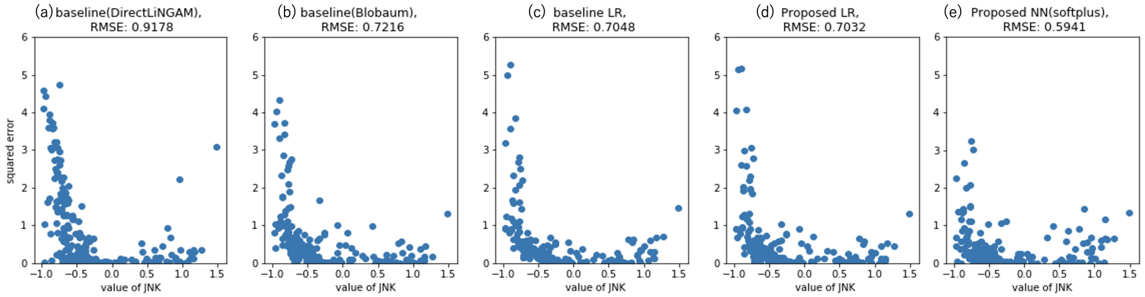
Fig 3 illustrates the average of squared error between d and $\hat{d}$ by boxplots of top five methods. The figure shows that our method with neural network models obtain the lowest MSE score for individuals.

6.3 Real Data

We also carried out qualitative experiments with a real-world dataset. Protein Signaling Data (Anti-CD3/CD28) (Sachs et al., 2005) is a cellular network dataset with its causal relationships between the features already identified by perturbation experiments. The original dataset has 853 samples with 11 features each. In this experiments, we selected 8 features of them following the setting of Balu and Borle(Balu and Borle, 2019) (see Appendix C). The given causal relationships includes only causal structure and not their strengths, which makes it difficult to evaluate the result of our proposed method.

As the generative model F, we therefore trained MLP regressors to predict the value of each variable from its parent variables along the given causal directions. We set JNK and PKC as the objective and intervened variables, respectively. We investigated errors of individual $\hat{d}_i$ predicted by each method when $d = 0$ and select samples which the value of JNK is from -1.5 to 1.5 to remove effects of outliers. The detailed experimental settings and results are in Appendix C.

[2] Boxplot in our experiments shows that green delta points are sample average of SE and the bottom, middle, and top line of boxes shows quantiles $0.25, 0.50, 0.75$ respectively. The whiskers extend from the box to show min and max of data. We remove outliers that have values greater than $Q1 - 1.5(Q3 - Q1)$ ($Q1$ means the the value of quantile 0.25).


(a) baseline(DirectLiNGAM), RMSE: 0.9178

(b) baseline(Blobaum), RMSE: 0.7216

(c) baseline LR, RMSE: 0.7048

(d) Proposed LR, RMSE: 0.7032

(e) Proposed NN(softplus), RMSE: 0.5941

Figure 4: The result in experiments of real-world data comparing with linear baselines and our proposed methods (regarding kinds of methods, see Section 6). The blue points means each sample. The vertical axis is the squared error of d and predicted individual $\hat{d}_i$ (the lower is better). The horizontal axis is the original value of JNK and limited from -1.5 to 1.5 because of removing outliers. RMSE is calculated by d and $\hat{d}_i$ in the range of horizontal axis.

Fig 4 shows sample plots of squared error score in some linear baseline methods and our proposed methods. While the baseline methods held higher errors when the target value of the objective variable is around -1, our method with softplus activation achieved lower errors in that range. Fig 5 shows another comparison between neural network baselines and our proposed methods. All the baseline methods failed extremely on some data points to get large SE scores, while our methods do not.

7. Discussion

In artificial experiments, we evaluate our methods and baseline by data generated by two simple models (F_1: Formula (6) and F_2: Formula (7)) and two complex models (F_3: Formula (8) and F_4: Formula (9)) and our proposed methods achieve the lowest error scores in most experiments. In real-world data, our proposed methods improve the results of other baseline methods. We discuss the result from three perspectives.

Effects of descendants of objective variables: In the two artificial experiments, our methods show better scores than other baselines except for F_4 model. These results indicate that including descendants variables of the objective variable in a prediction model, as explanatory variables, can improve prediction accuracy. This is consistent with the concept of Markov blanket where one's children nodes and their parent nodes are not conditional independent of the node focused on in a graphical model. F_2 also does not include the descendants of the objective variable as explanatory variable but our method achieves lower error score. This should be related to non-linearity and we state next paragraph.

Effects of Non-linearity: In two experiments, our proposed method obtain higher performance in the model F_2 and F_3 and this results coincide with regression accuracy on non-linear and linear regression in Fig 6. Intuitively, error scores of baseline methods in

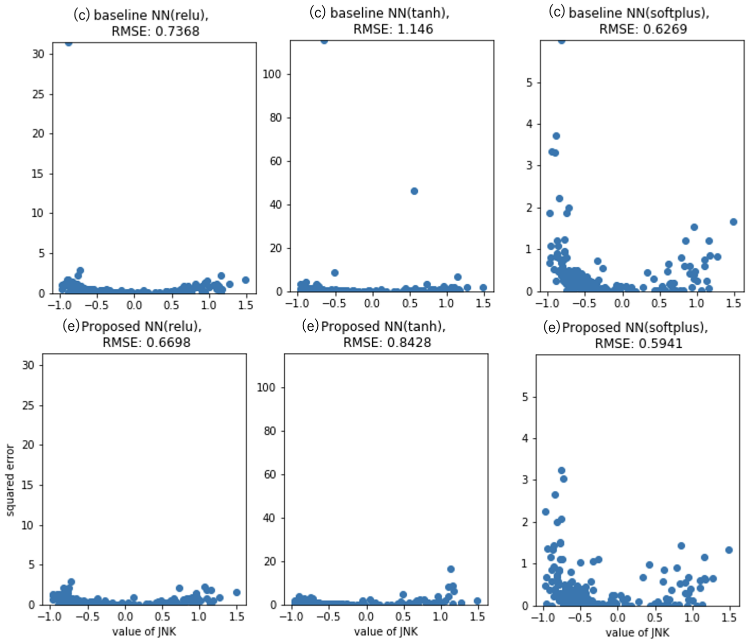


Figure 5: The result in experiments of real-world data comparing with neural network baseline (the upper) and our proposed methods (the lower). The ranges of the vertical axes are differ from each activation function. The baseline methods have some outliers with extreme large squared errors.

non-linear generative models can be lower because we use linear causal discovery method (DirectLiNGAM). Therefore, the factor to achieve lower error score in our proposed method is considered to combine non-linear prediction models with linear causal models. We can estimate non-linear causal structures by existing method based on additive noise models and achieve enough accuracy. However, it is costly because when the number of variables is p, ANM estimation method such as RESIT (Peters et al., 2014) have to train $p!$ regression models. On the other hands, our method have the advantage of calculation cost because it trains only one regression model.

Convergence of loss and Activation Function: Fig 3 shows that the neural network models with softplus and tanh achieve lower error score. Fig 6 shows that such activation functions can more reduce optimization loss in Algorithm 2 than ReLU in most cases. We can say that our framework is related to regression surfaces of neural networks and therefore smoother activation is suitable. In addition, in most cases, our method can decrease loss more than the baselines with the prediction model alone. This is because that our proposed method can vary descendant values of X_i indirectly through coefficient $\boldsymbol{\alpha}$ in the process of updating c. On the other hand, baselines only update a direct relation $X_i \rightarrow \hat{Y}$.

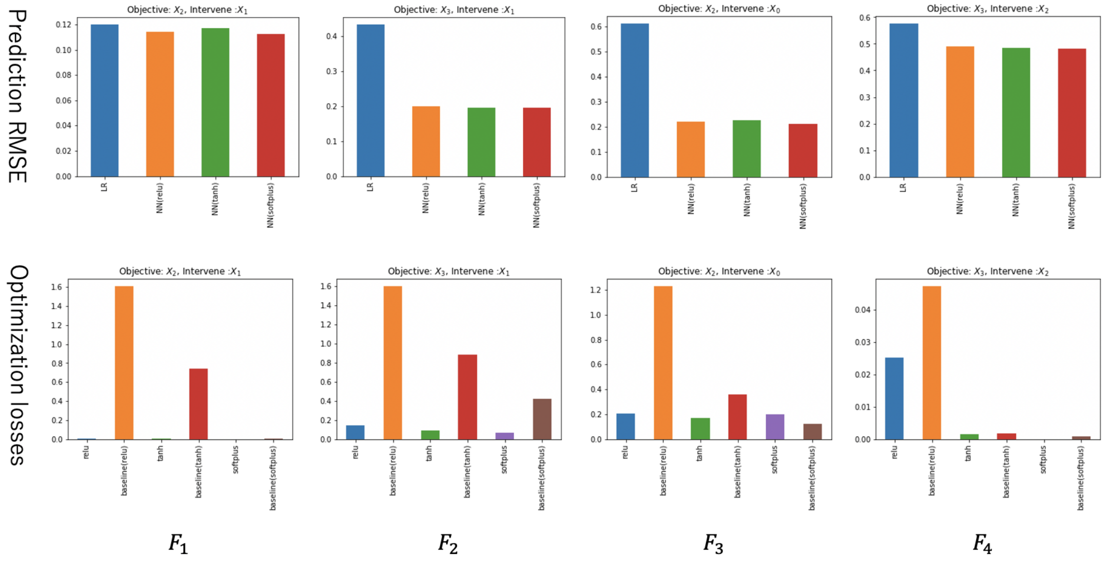


Figure 6: Artificial data experiments: prediction RMSE score (the upper) and the loss L in Algorithm 2 (the lower) for models F_1, F_2, F_3, F_4.

8. Conclusion

In this paper, we introduce a new problem setting to estimate individual-level *optimal causal intervention* for a continuous variable and formulate it as a optimization problem to minimize mean squared error between the counterfactual of objective variable and a user given desired value. We propose a method that combines an estimated causal structure and regression model into a single causal structure. That can improve individual-level causal effects estimations. We also propose the method to estimate individual-level optimal causal intervention using a gradient descent algorithm. In the experiments of evaluating estimation of individual-level causal effects and the optimal causal intervention, we compare our proposed methods with some baseline methods by artificial and real-world data and confirm that our method can estimate optimal causal intervention better than other baselines.

As a future work, more theoretical analysis is necessary for non-linearity. In the experiments, we use the combinations of linear causal structures and the non-linear prediction models for data generated by the non-linear causal structures and our method show higher performance than the baselines. To analyze this phenomenon, we can investigate how linear causal structures represent non-linear causal relationships. Moreover, the relationship between $\hat{Y}$ and Y, and, between the optimal intervention values derived from $\hat{Y}$ and from Y, should be theoretically analyzed. In the experiments, we achieve high accuracy but it is not clarified that its theoretical validity and what situation is suitable for. In this study, we use a simple causal model with some assumptions such as causal sufficiency for simplicity, as a first attempt for this problem. Development of methods including more general or more complex causal models remains as a future challenge.

Appendix A. Toy Experiment

We compare the accuracy on the two cases, (a) when the explanatory variables are only the parents of the objective variable and (b) when the explanatory variables including the children of the objective variable. Formula F_1 and F_2 generate 1500 and 500 training and test samples, respectively. We train MLP models (the number of nodes is 100, two intermediate layers, and activation is ReLU) on two cases (a) and (b). (e.g. in the case (a) of F_1 data, the explanatory variables are X_0, X_1, X_3 and in the case (b), the explanatory variables are X_0, X_1). Regression accuracy is shown in Table 1. This result shows that case (b) of the model F_1 achieves better regression accuracy because the prediction model of case (a) is not given noise of the objective variable X_2 by the explanatory variables. Intuitively, In the calculation of counterfactual, it is necessary to predict the objective value added noise so we can say that case (b) is suitable for predicting counterfactual by noise included by descendants of the objective variable.

Table 1: The Average of RMSE

	Model F_1	Model F_2
case (a)	0.144	0.160
case (b)	**0.0729**	0.161

Appendix B. Detailed Settings and Results for Artificial Data

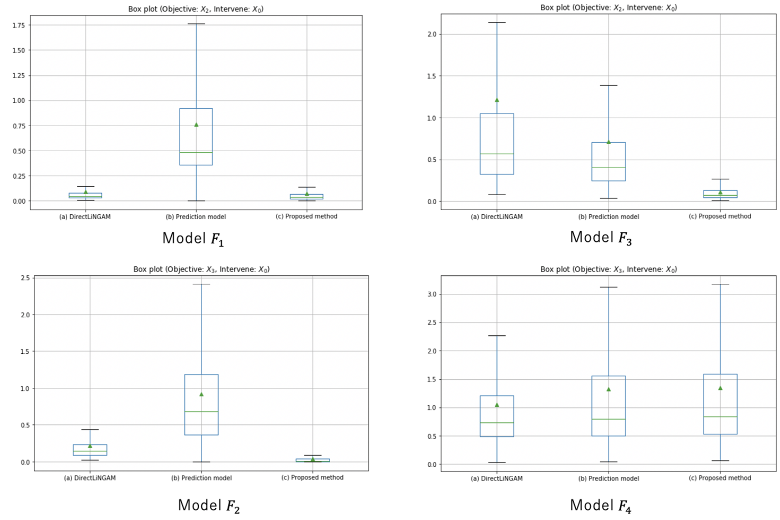
In the artificial experiments, MLPs has 256 nodes of two intermediate layers. The learning rate and optimizer are 0.01 and SGD, respectively. The number of epoch and batch size are 1000 and 64, then the hyper-parameters ϵ and η is 10^{-10} and 0.1, respectively. Fig **??** show that the estimated causal matrix.

Evaluation for estimating individual-level causal effects: Fig 6 show that experimental results regarding the RMSE of the prediction model gs in the experiment step 3 (see Section 6). In this experiment, activation functions of MLP is ReLU. Fig 7 show boxplots of that the averaged squared error between a predicted value $\hat{Y}_{X_i=c}(\boldsymbol{u})$ and a desired value d for all methods.

Evaluation for estimating optimal causal intervention: Fig 6 show that experimental results regarding the RMSE of the prediction model gs in the experiment step 3 (see Section 6) and the optimization loss L in Algorithm 2 for each prediction models, respectively. Fig 9 and 10 show boxplots of that the averaged squared error between a predicted value $\hat{d_i}$ and a desired value d for all methods.

Appendix C. Detailed Settings and Results for Real Data

For generation model F, we trained MLP regressors to predict the value of each variable from its parent variables along the causal directions given by (Balu and Borle, 2019) (Fig 11). The MLPs has 64 nodes of two intermediate layers and softplus activation. The learning rate and optimizer are 0.01 and SGD, respectively. The number of epoch and batch size are 1000 and 64.



Model F_1

Model F_3

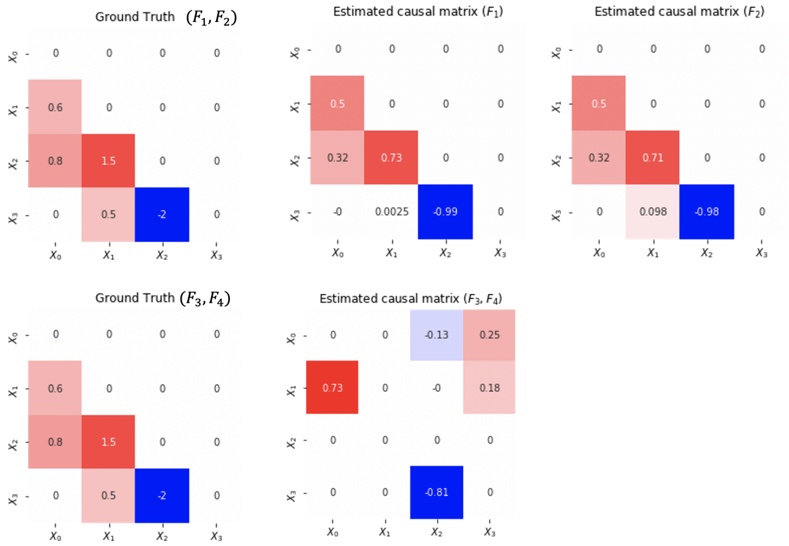
Model F_2

Model F_4

Figure 7: Evaluation for estimating individual-level causal effects: averaged SE score on structural equation models (lower score is better)

For prediction model g, We train MLP having 256 nodes of two intermediate layers. The learning rate and optimizer are 0.01 and SGD, respectively. The number of epoch and batch size are 1000 and 64, then the hyper-parameters ϵ and η is 10^{-10} and 0.1, respectively. Fig 11 shows that the estimated causal matrix.

Fig 12 show that experimental results regarding the RMSE of the prediction model gs in the experiment and the optimization loss L in Algorithm 2 for each prediction models, respectively.



Ground Truth (F_1, F_2)

Estimated causal matrix (F_1)

Estimated causal matrix (F_2)

Ground Truth (F_3, F_4)

Estimated causal matrix (F_3, F_4)

Figure 8: Artificial data experiments: estimated and ground truth causal matrix for F_1, F_2 (upper) and $F_3, F4$ (lower). The vertical and horizontal axes are parents and descendants, respectively. For example, in the ground truth of F_1 and F_2, there are a causal relation $X_2 \to X_3$ that its strength is -2.0. The values in ground truth shows coefficients before applying non-linear function. We estimate causal matrix by normalized data.

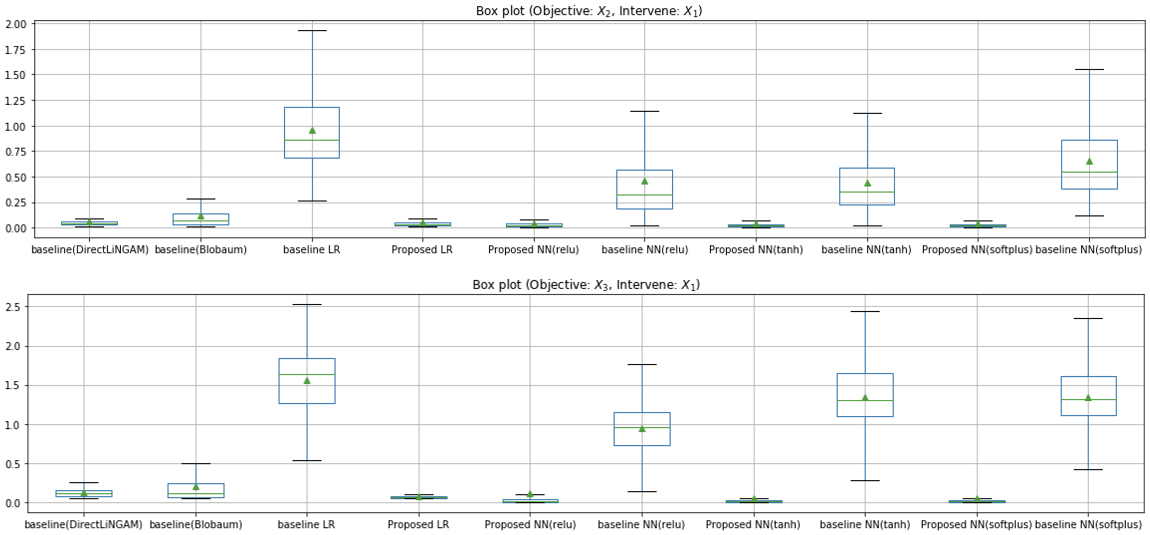


Figure 9: Evaluation for estimating optimal causal intervention: averaged SE score on simple structural equation models (lower score is better) The upper and the lower are F_1 and F_2, respectively.

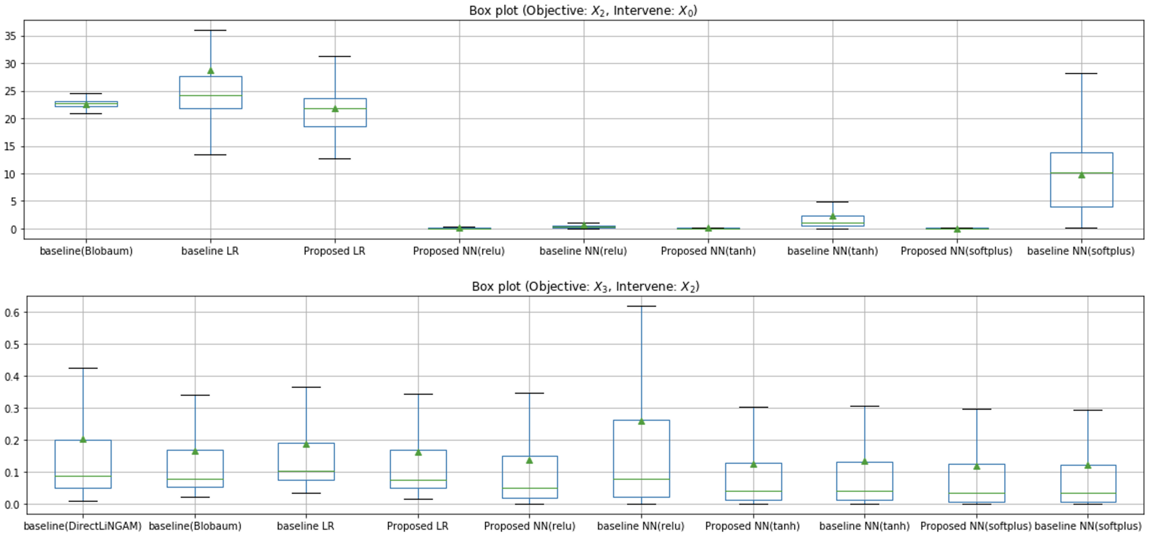


Figure 10: Evaluation for estimating optimal causal intervention: averaged SE score on simple structural equation models (lower score is better) The upper and the lower are F_3 and F_4, respectively..

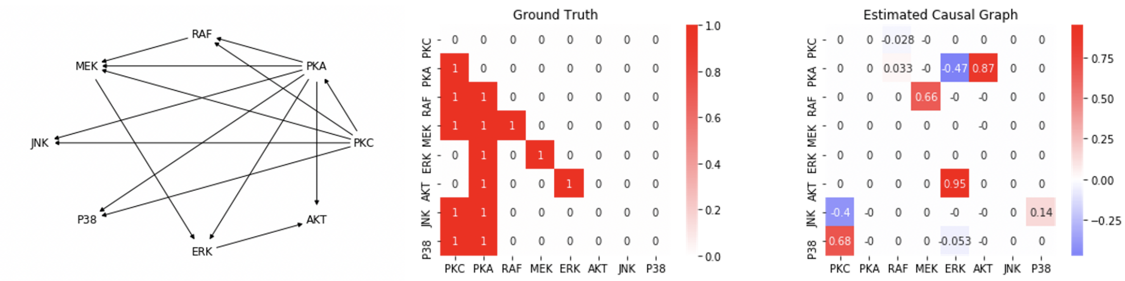
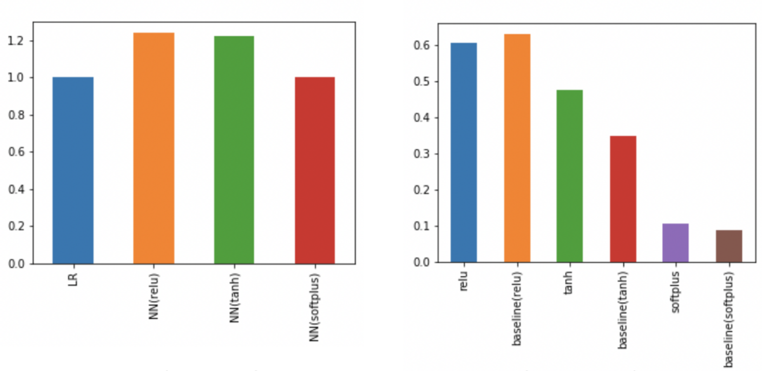


Figure 11: Causal graph of protein signaling data (Anti-CD3/CD28) (left) and ground truth and estimated causal matrix (right). The nodes of causal graph are selected by Balu's experiment setting. In the ground truth causal matrix, the connection is shown by 1.0 because dataset show only ground truth of causal connections without a structural equation model.



Prediction RMSE

Optimization losses

Figure 12: Real data experiments: prediction RMSE score (the left) and the loss L in Algorithm 2 (the right)

References

Virginia Aglietti, Theodoros Damoulas, Mauricio Álvarez, and Javier González. Multi-task causal learning with gaussian processes. *arXiv preprint arXiv:2009.12821*, 2020a.

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020b.

Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

Radhakrishnan Balu and Ajinkya Borle. Bayesian networks based hybrid quantum-classical machine learning approach to elucidate gene regulatory pathways. *arXiv preprint arXiv:1901.10557*, 2019.

Patrick Blöbaum and Shohei Shimizu. Estimation of interventional effects of features on prediction. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

Kenneth A Bollen and Rick H Hoyle. Latent variables in structural equation modeling. 2012.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. *arXiv preprint arXiv:1902.02302*, 2019.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, pages 13589–13600, 2019.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13:472–478, 2000.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

Ruocheng Guo, Jundong Li, and Huan Liu. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 232–240, 2020.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21:689–696, 2008a.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008b.

Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan): 111–152, 2013.

Samory Kpotufe, Eleni Sgouritsa, Dominik Janzing, and Bernhard Schölkopf. Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning*, pages 478–486. PMLR, 2014.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

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

Judea Pearl. Causal and counterfactual inference. *The Handbook of Rationality*, pages 1–41, 2018.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 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.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.