Stable Prediction with Leveraging Seed Variable

Kun Kuang, Haotian Wang, Yue Liu, Ruoxuan Xiong, Runze Wu, Weiming Lu*, Yueting Zhuang, Fei Wu*, Peng Cui, Bo Li*

Abstract—In this paper, we focus on the problem of stable prediction across unknown test data, where the test distribution might be different from the training one and is always agnostic when model training. In such a case, previous machine learning methods might exploit subtly spurious correlations induced by non-causal variables in training data for prediction. Those spurious correlations can vary across datasets, leading to instability of prediction across unknown test data. To address this problem, we propose an algorithm based on conditional independence tests to screen out non-causal features and reduce spurious correlations by leveraging a seed variable. We show, both theoretically and with empirical experiments, that our algorithm can precisely screen out the isolated non-causal variables, which have no causal relationship with other variables, and remove the spurious correlations induced by them, increasing the stability of prediction across unknown test data. Extensive experiments on both synthetic and real-world datasets demonstrate that our algorithm outperforms state-of-the-art methods for stable prediction across unknown test data.

Index Terms—Stable Prediction, Seed Variables, Conditional Independence, d-separation

1 Introduction

Many machine learning algorithms have been shown to be very successful for prediction when test data have the same distribution as the training data. In real-world scenarios, however, test data is always agnostic in model training and we cannot guarantee the unknown test data will have the same distribution as the training data. For example, different geographies, schools, or hospitals may draw from different demographics, and the correlation structure among demographics may also vary (e.g., one ethnic group may be more or less disadvantaged in different geographies). The model may exploit the subtle, but genuine, statistical relationships among predictors present in the training data to improve prediction, resulting in the instability of prediction across test data that differs from the training distribution. Hence, how to learn a model for stable prediction across unknown test data is of paramount importance for both academic research and practical applications.

To address the stable/invariant prediction problem, recently, many algorithms have been proposed, including domain generalization [1], [2], causal transfer learning [3], [4] and invariant causal prediction [5], [6]. The motivation of these methods is to explore the invariant or stable structure between predictors and the response variable across multiple training data for stable prediction. But they cannot handle the test data whose distribution are out of all training environments. Kuang et al. [7], [8] and Shen et al. [9], [10] proposed to recover causation between predictors and response variable by global sample weighting, and separate stable variables for stable prediction. However, they either assume all predictors are binary or analyze based on linear model, which are impractical in real scenarios.

In this paper, we focus on the problem of stable prediction via separating causal and non-causal variables. In the stable prediction problem [7], [8], all predictors $\mathbf{X}$ can be separated into two categories, including causal (stable) variables $\mathbf{C}$ and non-causal variables $\mathbf{N}$, by whether it has a direct causal link (causal effect) to the response variable Y or not, that is $\mathbf{X} = \{\mathbf{C}, \mathbf{N}\}$. For example, ears, noses, and whiskers are casual variables of cats to identify whether an image contains a cat or not, while the grass or other backgrounds are non-causal variables to recognize the cat. Then, the data generating process of the response Y can be written as $Y = f(\mathbf{X}) + \epsilon = f(\mathbf{C}) + \epsilon$, where non-causal variables $\mathbf{N}$ should be independent with the response variable Y conditional on the full sets of causal variables $\mathbf{C}$. But they might be spuriously correlated with either causal variables, response variable or both because of sample selection bias in data. For example, the variable "grass" would be spuriously correlated with label "cat" and become a powerful predictor if the training data has many images with "cat on the grass". Those spurious correlations between non-causal variables and the response variable vary and are unstable across datasets with different distributions, leading to instability of prediction across unknown test data. Hence, to address the stable prediction problem, one possible solution is to screen out those non-causal variables and separate causal variables for model training and prediction. However, in practice, analysts always have no prior knowledge on which variables are casual variables and which are non-causal variables.

- *Kun Kuang, Weiming Lu and Yueting Zhuang are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang Province, China*
- *Yue Liu is with the Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China.*
- *Ruoxuan Xiong is with the Department of Quantitative Theory and Methods, Emory University, GA, 30322, USA.*
- *Runze Wu is with the Fuxi AI Lab, NetEase Games, Hangzhou, Zhejiang Province, China of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang Province, China*
- *Fei Wu is with the Institute of Artificial Intelligence, Zhejiang University; Shanghai Institute for Advanced Study, Zhejiang University; Shanghai AI Laboratory.*
- *Peng Cui is with the Department of Computer Science and Technology in Tsinghua University, Beijing, China.*
- *Bo Li is with the School of Economics and Management in Tsinghua University, Beijing, China.*

* *Corresponding authors.*

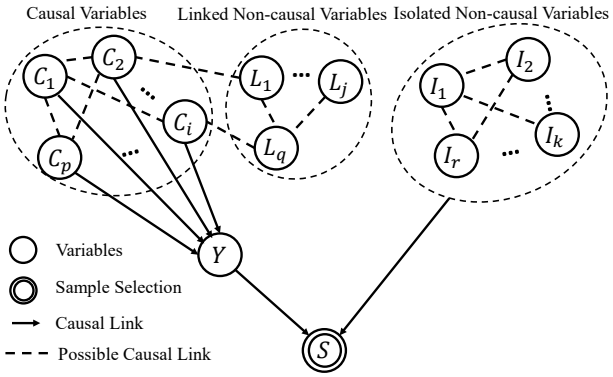


Fig. 1: Structural causal model (SCM) in our problem. All variables are categorized into three sets, including causal variables $\mathbf{C}$, which has a direct causal link[1] to response variable Y; linked non-causal variables $\mathbf{L}$, which have no direct causal link to Y but might be causal linked with causal variables $\mathbf{C}$; and isolated non-causal variables $\mathbf{I}$, which have no direct causal link to Y and are also isolated from (without causal link to) $\mathbf{C}$ and $\mathbf{L}$. Under the sample selection [11], [12] (indexed by variable S) between Y and $\mathbf{I}$, some isolated non-causal variables $\mathbf{I}$ might be highly correlated with either response variable, causal variables, or both[2], leading to spurious correlation between Y and $\mathbf{I}$, and instability of prediction.

Variable/Feature selection plays a very important role in machine learning. Traditional correlation based feature selection methods utilize either the correlation criteria [13] or mutual information criteria [14] without differentiating causation from spurious correlation, leading to unstable prediction across test data that are out of training distribution. In the literature of causality, causal discovery and causal estimation techniques can be adopted for causal variables selection. PC [15], FCI [15] and CPC [16] are three of the most prominent causal discovery methods based on conditional independence (CI) tests, but their complexity grow exponentially with the number of variables. Moreover, PC method need assume causal sufficiency, i.e., the assumption that all common causes of observed predictors are observed. Causal inference [17], [18], [19] and treatment effect estimation methods [20], [21], [22], [23] can approximately identify causal variables via estimating the causal effect of each variable, but they focused on binary predictors and required the ignoriablity assumption that all causal variables are observed.

Motivated by the practical scenarios that causal sufficiency assumption is not met and some of the causal variables are unobserved or unmeasured, in this paper, we propose a novel CI test based non-causal variables screening and causal variable separation method for stable prediction. In our paper, we assume the non-causal variables $\mathbf{N}$ can be separated into two sets: linked non-causal variables $\mathbf{L}$, which have no direct causal link to response variable Y but might be causal linked with some causal variables; and isolated non-causal variables $\mathbf{I}$, which have no direct causal link to Y and are also isolated from (without causal link to) both $\mathbf{C}$ and $\mathbf{L}$. Fig. 1 illustrates the structural causal model (SCM) in our problem. Then, we theoretically prove that one can screen out the isolated non-causal variable with a single CI test per variable. Specifically, as shown in Fig. 1, if we know a seed variable $\mathbf{C}_0$ is one of the causal variables, then each isolated non-causal variable $\mathbf{I}_k$ should satisfy that $\mathbf{I}_k \perp\!\!\!\perp \mathbf{C}_0 \mid Y$, and each causal variable $\mathbf{C}_i$ should satisfy that $\mathbf{C}_i \not\perp\!\!\!\perp \mathbf{C}_0 \mid Y$. With those theoretical analyses, we present a CI test based non-causal variables screening method for stable prediction. At a first step, we apply our non-causal variables screening method on synthetic data, which leads to high precision on separation of isolated non-causal variables. Screening out those isolated non-causal variables help to reduce the spurious correlation in training data and bring stability for model training and prediction. In real-world applications, we also demonstrate that our algorithm outperforms baseline algorithms in both causal variable separation task and stable prediction task.

Comparing with previous CI based causal discovery methods [15], [16], [24], [25], our method do not rely on the assumption of causal sufficiency and remain unaffected even some causal variables are unobserved. Moreover, our algorithm screen out the isolated non-causal variables with a single CI test per variable, scaling algorithmic complexity from exponential to linear with the number of variables. Comparing with sample based work on stable prediction [7], [8], our method can be applied for continuous settings and separate the causal variables without assumptions on regression model. Our work is similar with a recent paper [26], which also adopt CI for causal variable selection. But the tailored problems are totally different in the following ways: (i) [26] focused on detecting direct and indirect causes of a response variable under i.i.d settings, while our algorithm is designed for screening out a part of non-causal variables under the biased settings with sample selection bias; (ii) [26] is tailored for the problem in which a cause variable of each candidate causal variable is known, while our algorithm attempts to screen out those isolated non-causal variables by leveraging a seed causal variable. Moreover, we applied our method to address agnostic distribution shift issue between training and unknown test data for stable prediction.

Our contributions are summarized as follows:

- We investigate the problem of non-causal variables screening to improve the stability of prediction across unknown test environments.
- We propose an elaborative but effective non-causal variables screening algorithm based on conditional independence test for separating causal and non-causal variables.
- We give theoretical analysis on our proposed algorithm and proved that our algorithm can precisely screen out the set of isolated non-causal features and select causal features for stable prediction.
- Extensive experiments on both synthetic and real world datasets demonstrate the superior performance of our proposed algorithms on causal features selection and stable prediction.

The rest of this paper is organized as follows. Section

1. Variable "A" has a causal link to variable "B" refers to that "A" has direct causal effect on "B".

2. The distribution under sample selection is always conditioned on sample selection variable S.

2 reviews the related work. Section 3 gives the notations and formulates our problem. The details of our proposed algorithm for non-causal features screening and stable prediction are introduced in Section 4. Experimental results and analyses are reported in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

There is a significant body on research related to our problem, which we categorize into three groups: correlation based feature selection, causation based feature selection and stable/invariant prediction.

Correlation based features selection: The focus of feature selection is to select a subset of features from the input which can efficiently describe the input data while reducing effects from irrelevant and noise features and still provide good prediction results [27]. Traditional feature selection method can be clustered into three categories: filter methods [14], [28], [29], wrapper methods [30], [31] and embedded methods [32], [33]. Here, we focus on filter methods, which use features ranking techniques as criteria for features selection by ordering, where correlation criteria [28], [34] and mutual information criteria [14], [29], [35] are often used to describe the dependency, relevancy and redundancy of a feature to the data or the outcome variable for features selection. Traditional correlation based feature selection methods are widely used and achieved good performance in many applications. But their performance cannot be guaranteed in non-stationary environments, where the correlations in test data might be very different with the one on training.

Causation based features selection: Both causal discover methods and treatment effect estimation methods in causal inference literature can be employed for causal features selection. By causal discover methods [15], [25], [36], one can identify whether a predictor is an cause of outcome variable or not. But the complexity of these algorithms grows exponentially as the dimension of features. With methods for treatment effect estimation [20], [21], [22], [23], one can estimate the causal effect of each predictor on outcome variable for causal features selection, but these methods required that all the causal features are observed.

Our work is very similar with a recent paper [26], which is also adopts conditional independence test for causal feature selection. But the settings and main assumption are totally different: (i) the main assumption in [26] is that they can observe a cause for each possible causal feature of outcome variable, while our method suppose that the causal features are independent with non-causal features, and (ii) the conditional variable in [26] is a predictor, while our method is conditional on outcome variable. Comparing with [26], our method has the following advantages: (i) our method can be applied for stable prediction under non-i.i.d settings; (ii) the causal sufficiency assumption is not necessary for our algorithm; and (iii) our algorithm is not affected by the unobserved causal variables, but missing some causal variables would decrease the performance of predictive model on prediction.

Stable/Invariant prediction: Many methods have been proposed from different aspects for enhancing the stability and robustness of AI, such as artificial general intelligent [37], adversarial learning [38], [39], [40], and distributional robustness optimization [41]. In this paper, we focus on stable/invariant prediction across unknown data. Recently, some works have been proposed to address the stable prediction problem by either invariant component learning or causation recovery. Kuang et al. [7], [8] defined the problem of stable prediction, and proposed novel sample reweighting methods for isolating the effect of each predictor, which help to recover the causation between predictors and outcome variable, and finally identify the causal features for stable prediction. To achieve uniformly error on any data point, Shen et al. [42] proposed a sample reweighted decorrelation operator to decorrelate the predictors for stable prediction. Peters et al. [5] proposed invariant causal prediction algorithm to identify causal features by exploring the invariance of the conditional distribution of outcome variable across multiple training data. Domain generalization [1] methods estimate an invariant representation of data by minimizing the dissimilarity among multiple training data. These methods are facing the challenges from either non-convex optimization, high dimensional predictors, unaccepted complexity, or strong prior knowledge.

Out-of-distribution (OOD) Generalization: Improving the generalization capability of a learning model is a long-standing problem ranging from statistical analysis to machine learning area (e.g., Distributionally robust optimization (DRO), techniques such as Dropout and Mixup, and settings with explicit distributional shift such as Domain adaptation and Domain Generalization, and heterogeneous risk minimization) [43], [44], [45], [46]. Original attempt to OOD generalization can be attributed to the robust machine learning, which aims to improve the model performance towards the outliers or hard samples. Then researchers are not satisfied to deal with hard samples or perturbed samples: they turn to force the learning model generalize well on "hard" or shifted distribution, which results in the propose of transfer learning, domain generalization, and OOD generalization.

Retraining model might be a possible solution of OOD generalization in a few cases, while it is not practical for the most of real-world scenes. Retaining model requires the knowledge of new testing data to fine-tune the pretrained model for address the problem of OOD generalization. In practice, however, the bottleneck for retraining the model is the cost to collect and annotate the new test dataset (sometimes data for retrain is even not accessible) and the computational resource ("high performance computing and accelerators (GPU, FPGA etc)") for quickly fine-tune.

3 Problem and Notations

Let $\mathcal{X}$ denote the space of observed features and $\mathcal{Y}$ denote the space of response variable. We define an **environment** to be a joint distribution $P_{\mathbf{X}Y}$ on $\mathcal{X} \times \mathcal{Y}$, and let $\mathcal{E}$ denote the set of all environments. In each environment $e \in \mathcal{E}$, the dataset $\mathbf{D}^e = (\mathbf{X}^e, Y^e)$ is sampled from the corresponding distribution P^e_{XY}, where $\mathbf{X}^e \in \mathcal{X}$ are predictor variables and $Y^e \in \mathcal{Y}$ is a response variable. Let P^e_{XY} denote the joint distribution of features and outcomes on $(\mathbf{X}, Y)$ in environment e. The joint distribution of features and outcomes on

$(\mathbf{X}, Y)$ can vary across environments: i.e., $P^e_{XY} \neq P^{e'}_{XY}$ for $e, e' \in \mathcal{E}$.

In this paper, we consider a setting where a researcher has a single data set (data from one environment), and wishes to train a model that can then be applied to other environments. This type of problem might arise when a firm creates an algorithm that is then provided to other organizations to apply, for example, medical researchers might train a model and incorporate it in a software product that is used by a range of hospitals; academics might build a prediction model that is applied by governments in different locations. The researcher may not have access to the end user's data for confidentiality reasons. The problem can be formalized as a stable prediction problem [7] as follows:

Problem 1. *(Stable Prediction). **Given** one training environment $e \in \mathcal{E}$ with dataset $\mathbf{D}^e = \{\mathbf{X}^e, Y^e\}$, the task is to **learn** a predictive model that can **stably** predict across unknown test environments $\mathcal{E}$.*

In this problem, let $\mathbf{X} = \{\mathbf{C}, \mathbf{N}\}$, we define $\mathbf{C}$ as causal variables, and $\mathbf{N}$ as non-causal variables with the following assumption [7]:

Assumption 1. *There exists a stable probability function $P(y|c)$ such that for all environment $e \in \mathcal{E}$, $P(Y^e = y|\mathbf{X}^e = x) = P(Y^e = y|\mathbf{C}^e = c, \mathbf{N}^e = n) = P(Y^e = y|\mathbf{C}^e = c) = P(y|c)$.*

Assumption 1 illuminates that the non-causal variables $\mathbf{N}$ do not affect the response variable during the data generation processing (i.e., $Y = f(\mathbf{X}) + \epsilon = f(\mathbf{C}) + \epsilon$), but it might be spuriously correlated with either response variable, causal variables, or both since sample selection bias problem as shown in Fig. 1. These spurious correlations might vary across environments. Hence, to make a stable prediction, one should guarantee the prediction only depending on the causal variables. Thus, one can address the stable prediction problem by separating causal variables $\mathbf{C}$ and learning the stable function $P(y|c)$. But, in practice, we have no prior knowledge on which variables are causal and which are non-causal.

In this work, we further divide the non-causal variables, which have no direct causal link to Y, into two categories as shown in Figure 1: linked non-causal variables $\mathbf{L}$, which might be causal linked with causal variables $\mathbf{C}$; and isolated non-causal variables $\mathbf{I}$, which are isolated from (without causal link to) both $\mathbf{C}$ and $\mathbf{L}$. Here, we focus on screening out those isolated non-causal variables, hence reduce part of spurious correlations in training data and improving the stability of prediction.

Notations. In our paper, n refers to the sample size, and p is the dimensions of variables. For any vector $\mathbf{v} \in \mathbb{R}^{p\times 1}$, let $\|\mathbf{v}\|_2^2 = \sum_{i=1}^p v_i^2$, and $\|\mathbf{v}\|_1 = \sum_{i=1}^p |v_i|$. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times p}$ (e.g., $\mathbf{X}$, $\mathbf{C}$, $\mathbf{N}$, $\mathbf{I}$, $\mathbf{L}$ in our problem), we let $\mathbf{A}_j$ represent the j^{th} variable in $\mathbf{A}$.

4 METHODS

In this section, we first give the background of causal graph, then introduce the details of isolated non-causal variable screening algorithm with single conditional independence (CI) test per feature for stable prediction.

4.1 Background on Causal Graph

Firstly, we revisit key concepts and theorems related to d-separation and CI in causal graph.

Let $G = \{\mathbf{V}, \mathbf{E}\}$ represents a causal directed acyclic graph (DAG) with nodes $\mathbf{V}$ and edges $\mathbf{E}$, where a node denotes a variable and an edge represents the direct dependence or causal direction between two variables. In a DAG, $\mathbf{V}_i \to \mathbf{V}_j$ refers to that $\mathbf{V}_i$ is a cause of $\mathbf{V}_j$ and $\mathbf{V}_j$ is an effect of $\mathbf{V}_i$.

Definition 1 (d-separation [18])**.** *In a DAG G, a path π is said to be d-separated by a set of nodes $\mathbf{Z}$ if and only if (i) π contains a chain $\mathbf{V}_i \to \mathbf{V}_k \to \mathbf{V}_j$ or a fork $\mathbf{V}_i \leftarrow \mathbf{V}_k \to \mathbf{V}_j$ such that the middle node $\mathbf{V}_k$ is in $\mathbf{Z}$, or (ii) π contains a collider $\mathbf{V}_i \to \mathbf{V}_k \leftarrow \mathbf{V}_j$ such that the middle node $\mathbf{V}_k$ is not in $\mathbf{Z}$ and such that no descendant of $\mathbf{V}_k$ is in $\mathbf{Z}$.*

Definition 2 (Conditional Independence)**.** *Given two distinct variables $\mathbf{V}_i, \mathbf{V}_j \in \mathbf{V}$ are said to be conditionally independent given a subset of variables $\mathbf{Z} \subseteq \mathbf{V} \setminus \{\mathbf{V}_i, \mathbf{V}_j\}$ (i.e. $\mathbf{V}_i \perp\!\!\!\perp \mathbf{V}_j|\mathbf{Z}$), if and only if $P(\mathbf{V}_i, \mathbf{V}_j|\mathbf{Z}) = P(\mathbf{V}_i|\mathbf{Z})P(\mathbf{V}_j|\mathbf{Z})$. Otherwise, $\mathbf{V}_i$ and $\mathbf{V}_j$ are conditionally dependent given $\mathbf{Z}$ (i.e. $\mathbf{V}_i \not\perp\!\!\!\perp \mathbf{V}_j|\mathbf{Z}$).*

The connection between d-separation and CI is established through the following lemma:

Lemma 1 (Probabilistic Implications of d-Separation [18], [47])**.** *If variables $\mathbf{V}_i$ and $\mathbf{V}_j$ are d-separated by $\mathbf{Z}$ in a DAG G, then $\mathbf{V}_i$ is independent of $\mathbf{V}_j$ conditional on $\mathbf{Z}$ in every distribution compatible with the DAG G. Conversely, if $\mathbf{V}_i$ and $\mathbf{V}_j$ are not d-separated by $\mathbf{Z}$ in a DAG G, then $\mathbf{V}_i$ and $\mathbf{V}_j$ are dependent conditional on $\mathbf{Z}$ in at least one distribution compatible with G.*

4.2 Isolated Non-Causal Variables Screening

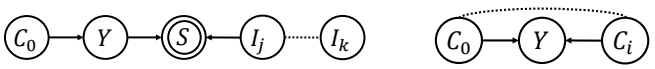
Based on lemma 1, in this paper, we propose an elaborative but effective algorithm to screen out the isolated non-causal variables by combining the mechanisms of d-separation and causality with the following assumption.

Assumption 2. *We have prior knowledge on one causal variable as seed variable. Formally, we know $\mathbf{C}_0 \in \mathbf{C}$.*

Under assumption 2, we have the following theorem to support for precisely screening out those isolated non-causal variables and reduce the spurious correlation in training data, hence improving the stability of model training and prediction across unknown test data.

Theorem 1. *Given a causal variable $\mathbf{C}_0$, observed variables $\mathbf{X}$ and response variable Y, and assuming 1 and 2, then, for each causal variable $\mathbf{C}_i \in \mathbf{C}$, we have $\mathbf{C}_i \not\perp\!\!\!\perp \mathbf{C}_0 \mid Y$; and for each isolated non-causal variable $\mathbf{I}_k \in \mathbf{I}$, we have $\mathbf{I}_k \perp\!\!\!\perp \mathbf{C}_0 \mid Y$.*

Proof. Assumption 1 implies that non-causal variables $\mathbf{N} = \{\mathbf{L}, \mathbf{I}\}$ are not direct causes of response Y, but causal variables $\mathbf{C}$ are the direct causes. Hence, in our causal DAG, there exists a direct edge from each causal variable $\mathbf{C}_i$ to response Y, but $\mathbf{N}$ have no any edges that directly point to Y. With the definition of linked and isolated non-causal variables, we know the linked non-causal variables $\mathbf{L}$ might be causally linked with causal variables, whiling the isolated non-causal variables $\mathbf{I}$ have no causal link with both $\mathbf{C}$



(a) Path between $\mathbf{C}_0$ and $\mathbf{I}_k$ (b) Path between $\mathbf{C}_0$ and $\mathbf{C}_i$

Fig. 2: Causal paths between a known causal variable $\mathbf{C}_0$ and other variables, including isolated non-causal variable $\mathbf{I}_k$ and other causal variable $\mathbf{C}_i$. The dash line between two variables refers to the causal path between them is unknown.

and $\mathbf{L}$. The inner causal structure in each variables set (i.e., $\mathbf{C}$, $\mathbf{L}$, and $\mathbf{I}$) might be very complex and unknown. With considering the sample selection bias is generated based on the response Y and part of isolated non-causal variables $\mathbf{I}$, the structural causal model (SCM) in our problem is shown in Figure 1.

From Figure 1, the path between the seed causal variable $\mathbf{C}_0$ and any isolated non-causal variable $\mathbf{I}_k$ can be represented as Figure 2a, where the causal links between $\mathbf{I}_j$ and $\mathbf{I}_k$ are unknown, could be very complex or could be that $\mathbf{I}_j$ is exactly $\mathbf{I}_k$ if sample selection is based on $\mathbf{I}_k$ and Y. With the definition of d-separation, we have that $\mathbf{C}_0$ and $\mathbf{I}_k$ are d-separated by variable Y. Hence, $\mathbf{I}_k \perp\!\!\!\perp \mathbf{C}_0 \mid Y$ for any $\mathbf{I}_k \in \mathbf{I}$ guaranteed by the lemma 1.

On the other hand, the path between the seed causal variable $\mathbf{C}_0$ and any other causal variable $\mathbf{C}_i$ can be represented as Fig. 2b, where the causal links between $\mathbf{C}_0$ and $\mathbf{C}_i$ are unknown. Similarity, with the definition of d-separation, we know that the response variable Y is a collider and cannot d-separate $\mathbf{C}_0$ and $\mathbf{C}_i$. Therefore, with the lemma 1, we have $\mathbf{C}_i \not\perp\!\!\!\perp \mathbf{C}_0 \mid Y$ for any $\mathbf{C}_i \in \mathbf{C}$.

Overall, we can screen out the isolated non-causal variables from the causal variables by a single CI test per variable. □

Based on theorem 1, we know that a very simple conditional test can help to screen out the isolated non-causal variables since the p-value of CI test between I and $\mathbf{C}_0$ conditional on Y would be significantly higher than the one of causal variables $\mathbf{C}$. Thus, we propose a causal variables selection algorithm via one single CI test per variable to screen out the isolated non-causal variables. The details of our algorithm are summarized in Algorithm 1. With screening out the isolated non-causal variables and selecting top-k causal variables, we can learn a more stable predictive model for prediction across unknown test data.

Remark 1. *From the proof of theorem 1, we know that our algorithm only need a single CI test of that variable and a known causal variable conditional on the response variable, with no need to know the other causal variables or common causes of observed variables. Then, we conclude that (i) the causal sufficiency assumption is not necessary for our algorithm; and (ii) our algorithm is not affected by the unobserved causal variables, but missing some causal variables would decrease the performance of predictive model on prediction.*

4.3 Discussion and Analysis

Complexity Analysis. Note that our algorithm requires only a single CI test per variable. Therefore, it speeds up the causal variables separation as it scales almost linearly with the number of variables. Specifically, the time complexity of the proposed algorithm (see Algorithm 1) consists of two main components: the for loop (statements 1-3 in Algorithm 1) for calculating p-value of CI test on each variable, and ranking the p-value (statement 4 in Algorithm 1) of all variables. From the calculation of p-value of CI test, we know the complexity of each single CI test should be related with the number of sample size. Different CI test methods can be applied in our algorithm for causal variable selection, and different CI test methods might be slightly different on the time complexity. Therefore, we use $t(n)$ to denote the complexity of a single CI test in the statement 2 in our algorithm 1, where n refers to the number of sample size. Then, the for loop (statements 1-3) in algorithm 1 requires the complexity of $\mathcal{O}(t(n)p)$, where p refers to the dimension/number of variables. The ranking/sorting procedure (statement 4) in algorithm 1 requires $\mathcal{O}(p \log p)$. Overall, the time complexity of the proposed algorithm 1 is $\mathcal{O}(t(n)p) + \mathcal{O}(p \log p)$. More details about the complexity and analysis on a single CI test can be found in [48], [49].

Discussions on assumptions. Assumption 1 refers to that the underlying predictive mechanism is invariant across environments, which is the basic assumption for causal variables identification and stable/invariant prediction [5], [7]. As for assumption 2, we think it is reasonable and acceptable in real applications. For example, if we want to predict the crime rate, we could know the income is one causal variable. Moreover, one can identify a causal variable as the seed variable by estimating its causal effect [20], [21], [22], [23]. In this paper, we employ causal effect estimator [21] to identify one causal variable as seed variable without assumption 2.

Discussion on non-causal variables. Throughout our paper, we indeed achieves a trade-off between the computational feasibility and the complexity of causal structure, by making a mild assumption that no causal relationship exists between linked non-causal variables $\mathbf{L}$ and the sample selection $\mathbf{S}$. With such an assumption, the feasibility of computation can be promoted towards some specific real-data with appropriate prior knowledge. Considering the general cases without any constrains on causal relationship, we require exponential skeleton search methods (e.g., PC, FCI) to first recover the total causal structure and make further inference. However, such methods designed in general cases requires strict assumptions (e.g., causal sufficiency) and are sensible towards high-dimensional data due to the large number of statistical CI tests. By the assumption on independence between $\mathbf{L}$ and $\mathbf{S}$, our method could get rid of the untestable causal sufficiency assumption and only require linear number of CI tests with respect to the number of features. On the other hand, there indeed exists realistic cases which mirror our assumption. An classical example is presented in the field of non-invasive brain stimulation [26], where the independence between biased non-causal variables and causal variables can be provided by prior. On the contrary, once this assumption is violated, we could not conclude that $\mathbf{I}_k \perp\!\!\!\perp \mathbf{C}_0 \mid Y$, as some path through $\mathbf{C}$ between $\mathbf{I}_k$ and $\mathbf{C}_0$ cannot be blocked by only conditioning on $\mathbf{Y}$. In the empirical results, we demonstrate the effectiveness of the proposed algorithm even this assumption is violated.

Algorithm 1 top-k Causal Variables Selection

Require: $\mathbf{X} \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^n$, $\mathbf{C}_0$ and parameter k
Ensure: top-k casual variables
1: **for** each variable $\mathbf{X}_i \in \mathbf{X}$ **do**
2: Calculate p-value of CI test: $pv_i = \text{CI-test}(\mathbf{X}_i \perp\!\!\!\perp \mathbf{C}_0 \mid Y)$
3: **end for**
4: $\mathbf{X}_{ranking} = Ranking(\mathbf{X}, \mathbf{pv})$ ▷ Ranking $\mathbf{X}_i \in \mathbf{X}$ by their p-value pv_i in ascending order
5: **return** top-k ranked variables in $\mathbf{X}_{ranking}$

Discussions on the spurious correlation. Under the SCM shown in Fig. 1, there are many kinds of spurious correlations between non-causal variables and response variable such as spurious correlation between $\mathbf{L}$ and Y from confounding bias, and spurious correlation from sample selection between $\mathbf{I}$ and Y (also can be from sample selection between $\mathbf{L}$ and Y). In this paper, we focus on addressing the spurious correlation from sample selection between $\mathbf{I}$ and Y with theoretical guarantee. Although it is only a part of spurious correlation in data, removing it can still improve the stability of prediction as we will demonstrate in experiments. In the future, we will try to address other spurious correlations.

5 Experiments

In this section, we evaluate the performance of our algorithm on both synthetic and real world datasets.

5.1 Baselines

We adopt the following variable selection methods as baselines. (i) **Correlation based methods**, including minimal Redundancy Maximal Relevance (mRMR) [14], Random Forest (RF) [34] and LASSO [28], they would be affected by the spurious correlation between non-causal variable and the response variable, and select non-causal variables for prediction. (ii) **Causation based methods**, including PC-simple[1] [24] and causal effect (CE) estimator [20], [21], [22], [23], they need to assume all causal variables are observed, moreover, PC-simple requires causal sufficiency and with curse of dimensionality. (iii) **Stable/Invariant learning based methods**, including invariant causal prediction (ICP)[2] [5] and global balancing algorithm (GBA) [7], [8], [9], ICP need multiple training environments for reveal causation and GBA requires tremendous training data for global sample weighting.

We do not compare with a recent causal variable selection method [26], since it requires the knowledge of a cause variable of each candidate causal variable, which is not applicable in our problem.

1. Previous CI based methods either need observe all causal variables, or assume causal sufficiency, moreover, with curse of dimensionality. So, we only compare with PC-simple, a prominent CI based method.

2. ICP method cannot be applied for variables ranking, but selecting a subset of variables for prediction, where the size of that subset variables is determined by its algorithm. Hence, the experimental results of ICP reported in this paper is based on its unique subset of selected variables.

In our algorithm, we employ causal effect estimator [21] to identify one causal variable as seed variable without assumption 2. Then, we execute CI test with the bnlearn method [50], denoted as *Our+BNCI*, and the RCIT [51] method, denoted as *Our+RCIT*. More specifically, the BNCI-CI test algorithm we used in this paper is derived from the information-theoretical approach with mutual information, which is proportional to the log-likelihood ratio test G^2 [50]. Meanwhile, the RCIT method for fast CI testing is based on random Fourier Operator Approximation, which aims to accelerate the computation of large-size kernel matrix. Based on the selected variables from each algorithm, we apply a linear model[3] for prediction to check their stability across unknown test data.

5.2 Evaluation Metrics

In this paper, we have two main tasks, including causal feature separation/selection and stable prediction with the selected causal variables.

To evaluate the performance of causal variable separation/selection, we use precision@k and ranking index of unstable non-causal variable as evaluation metrics. Precision@k refers to the proportion of top-k selected variables that are hitting the true causal variables set as follows:

$$Precision@k = \frac{|\{x_i | x_i \in \hat{\mathbf{C}}, index(x_i) < k, x_i \in \mathbf{C}\}|}{k},$$

where $\hat{\mathbf{C}}$ and $\mathbf{C}$ refer to the set of selected causal variables and true causal variables, respectively. $index(x_i)$ is the ranking index of variable x_i in the selected variables $\hat{\mathbf{C}}$.

To evaluate the stable prediction with the selected causal variables, similar to [7], we also adopt Average_Error and Stability_Error to measure the performance of stable prediction with the following definition:

$$\text{Average\_Error} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \text{RMSE}(\mathbf{D}^e),$$
$$\text{Stability\_Error} = \sqrt{\tfrac{1}{|\mathcal{E}|-1}\sum_{e\in\mathcal{E}} \left(\text{RMSE}(\mathbf{D}^e) - \text{Average\_Error}\right)^2},$$

where $\text{RMSE}(\mathbf{D}^e)$ represents the Root Mean Square Error on dataset $\mathbf{D}^e$ with following definition:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}, \quad (1)$$

where $\hat{Y}_i$ and Y_i refer to the predicted and true outcome of sample i, and n is the sample size.

5.3 Experiments on Synthetic Data

In this section, we check the performance of our algorithm with extensive simulations.

3. For simplification, we use a linear model to evaluate the selected variables, other models can also be applied.

5.3.1 Dataset

To generate the synthetic datasets, we consider the sample size $n = 2000$ and dimension of observed variables $p = \{10, 20, 40, 80\}$. We first generate the observed variables $\mathbf{X} = \{\mathbf{C}, \mathbf{L}, \mathbf{I}\}$ based on our SCM as shown in Fig. 1. Specifically, we generate $\mathbf{X} = \{\mathbf{C}_1, \cdots, \mathbf{C}_{p_C}, \mathbf{L}_1, \cdots, \mathbf{L}_{p_L}, \mathbf{I}_1, \cdots, \mathbf{I}_{p_I}\}$ with the help of auxiliary variables $\mathbf{Z_C}$ and $\mathbf{Z_I}$ with independent Gaussian distributions as:

$$\begin{aligned}
&\mathbf{Z}_{\mathbf{C}1}, \cdots, \mathbf{Z}_{\mathbf{C}p} \overset{iid}{\sim} \mathcal{N}(0,1);\\
&\mathbf{C}_i = 0.8 * \mathbf{Z}_{\mathbf{C}i} + 0.2 * \mathbf{Z}_{\mathbf{C}i+1},\ i = 1, 2, \cdots, p_C\\
&\mathbf{L}_j = 0.1 * \mathbf{C}_j + 0.3 * \mathbf{C}_{j+1} + \mathcal{N}(0,1),\ j = 1, 2, \cdots, p_L\\
&\mathbf{Z}_{\mathbf{I}1}, \cdots, \mathbf{Z}_{\mathbf{I}p} \overset{iid}{\sim} \mathcal{N}(0,1);\\
&\mathbf{I}_k = 0.8 * \mathbf{Z}_{\mathbf{I}k} + 0.2 * \mathbf{Z}_{\mathbf{I}k+1},\ k = 1, 2, \cdots, p_I,
\end{aligned}$$

where the number of causal variables $p_C = 0.3 * p$, number of linked non-causal variables $p_L = 0.3 * p$ and the number of isolated non-causal variables $p_I = 0.4 * p$. $\mathbf{C}_i$, $\mathbf{L}_j$ and $\mathbf{I}_k$ represent the i^{th}, j^{th} and k^{th} variable in $\mathbf{C}$, $\mathbf{L}$ and $\mathbf{I}$, and i, j, k are given by $i = mod(i, p_C)$, $j = mod(j, p_L)$, $k = mod(k, p_I)$, respectively. The function $mod(x, y)$ returns the modulus after division of x by y.

Then, we generate the response variable Y from a nonlinear function as:

$$Y = \sum_{i=1}^{p_C} \alpha_i \cdot \mathbf{C}_i + \sum_{j=1}^{p_C} \beta_j \cdot e^{\mathbf{C}_j \mathbf{C}_{j+1} \mathbf{C}_{j+2}} + \varepsilon,$$

where $\alpha_i = (-1)^i \cdot p_C/i$, $\beta_j = I(mod(j, 3) \equiv 1)$ and $\varepsilon = \mathcal{N}(0, 0.3)$. It is also noteworthy that the index of $\mathbf{C}_j$ for generating Y is given by $j = mod(j, p_C)$. The $I(\cdot)$ is the indicator function and function $mod(x, y)$ returns the modulus after division of x by y.

5.3.2 Generating Environments via Biased Sample Selection

From the generation of Y, we know that Y is only affected by the causal variables $\mathbf{C}$, and independent with the non-causal variables $\mathbf{N} = \{\mathbf{L}, \mathbf{I}\}$. In real applications, however, some non-causal variables might be spuriously correlated with Y since sample selection bias as shown in Fig. 1, and their correlation might vary across datasets. To check the stability of algorithms under that practical setting, we generate a set of environments, each with a stable probability $P(Y|\mathbf{C})$, but a distinct spuriously correlation $P(Y|\mathbf{N})$. For simplification, we only set one isolated non-causal variable $\mathbf{I}_{pI}$ as the *unstable non-causal variable*, and change its spuriously correlation $P(Y|\mathbf{I}_{pI})$ across environments.

Specifically, we vary $P(Y|\mathbf{I}_{pI})$ via biased sample selection with a bias rate $r \in [-3, -1) \cup (1, 3]$ based on $\mathbf{I}_{pI}$ and Y as shown in Fig. 1. For each sample, we select it with probability $Pr = |r|^{-5*D_i}$, where $D_i = |Y - sign(r) * \mathbf{I}_{pI}|$. If $r > 0$, $sign(r) = 1$; otherwise, $sign(r) = -1$.

Note that $r > 1$ corresponds to positive spurious correlation between Y and $\mathbf{I}_{pI}$, while $r < -1$ refers to the negative spurious correlation between Y and $\mathbf{I}_{pI}$. The higher value of $|r|$, the stronger correlation between $\mathbf{I}_{pI}$ and Y. Different value of r refers to different environments. All methods are trained with $r_{train} = 2.0$, but tested across environments with different $r_{test} \in [-3, -1) \cup (1, 3]$. To screen off the randomness, we generate 50 datasets for each number of dimension p with different random seeds and report their average results.

TABLE 1: Results of precision@k, where k equals the number of stable variables and linked non-causal features, namely $k = p_C + p_L = 0.6 * p$. ICP method cannot be applied for selecting variable with specific size. Results are averaged on 50 generated synthetic datasets for each p with different random seeds. We denote the best and the second best method with **bold** and underscore, respectively.

Dimension	p=10	p=20	p=40	p=80
mRMR	0.500	0.583	0.586	0.598
RF	0.640	0.668	0.594	0.592
LASSO	0.693	0.706	0.622	0.601
PC-simple	0.663	0.636	0.598	0.616
CE	0.643	0.688	0.689	0.665
ICP	-	-	-	-
GBA	0.640	0.660	0.655	0.669
Our+BNCI	**0.910**	**0.827**	**0.774**	**0.713**
Our+RCIT	0.900	0.820	0.767	0.703

TABLE 2: Ranking index of the unstable non-causal variable $\mathbf{I}_{pI}$, where "Y" denotes that the unstable non-causal variable is in the selected subset in ICP method. Results are averaged on 50 generated synthetic datasets for each p with different random seeds. we denote the best and the second best method with **bold** and underscore, respectively.

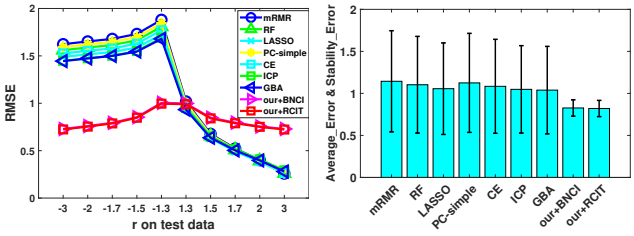
Dimension	p=10	p=20	p=40	p=80
mRMR	1	1	1	1
RF	1	1	1	1
LASSO	2.42	1	1	1
PC-simple	1	1	1	1
CE	3.48	2.08	2.32	2.28
ICP	Y	Y	Y	Y
GBA	3.54	3.5	2.06	1.24
Our+BNCI	**8.34**	14.64	27.92	**53.48**
Our+RCIT	8.2	**14.7**	**28.04**	49.24

5.3.3 Results on Causal Variables Separation/Selection

We report the results on causal variable selection from two aspects, including the ranking of causal variable with precision@k in Tab.1 and ranking of unstable non-causal variable in Tab. 2. The ranking of stable variables determines the average error of prediction across environments, the closer to 1 of precision@k, the better; while the ranking of unstable non-causal variable determines the stability error of prediction across environments, the lower ranking, the better. From Tab. 1 and 2, we conclude that: (i) Traditional correlation based variables selection methods, including mRMR, Random Forest and LASSO cannot precisely select the stable variables (with lower precision@k) and rank the unstable non-causal variable with a higher ranking. The main reason is that the spurious correlation from non-causal variable $\mathbf{I}_{pI}$ is more significant than causation from causal variables under the sample selection bias. (ii) The performance of PC-simple is similar to correlation based method, since it's hard to search the optimal solution for PC-simple via naively random search, moreover, it relies on the causal sufficiency assumption and needs to observed all causal variables. (iii) The performance of causation based methods, including CE and GBA, is better than those correlation

TABLE 3: Results of Average_Error and Stability_Error with different dimension p. For each performance under different p, we denote the best and the second best method with **bold** and underscore, respectively. Results are averaged on 50 generated synthetic datasets for each p with different random seeds.

Dimension	p=10		p=20		p=40		p=80	
Metrics	Average_Error	Stability_Error	Average_Error	Stability_Error	Average_Error	Stability_Error	Average_Error	Stability_Error
mRMR	1.055	0.540	1.144	0.602	1.175	0.621	1.183	0.625
RF	0.988	0.501	1.103	0.575	1.170	0.618	1.184	0.626
LASSO	0.988	0.501	1.056	0.545	1.165	0.614	1.184	0.626
PC-simple	1.020	0.522	1.125	0.590	1.173	0.620	1.183	0.625
CE	0.662	0.230	1.084	0.559	1.150	0.604	1.176	0.621
ICP	0.697	0.319	1.062	0.547	1.185	0.630	1.167	0.573
GBA	0.611	0.191	1.034	0.520	1.135	0.595	1.174	0.620
Our+BNCI	**0.410**	**0.026**	0.827	0.097	**0.942**	**0.136**	**1.030**	**0.200**
Our+RCIT	0.512	0.042	**0.820**	**0.096**	0.949	0.138	1.040	0.247



(a) RMSE across different test data (b) Average Error (green bar) & Stability Error (black line)

Fig. 3: Prediction results across unknown test data with $n = 2000, p = 20$. All methods are trained with $r_{train} = 2.0$, but tested across test environments with different $r_{test} \in [-3, -1) \cup (1, 3]$. Results are averaged on 50 generated synthetic datasets for each p with different random seeds.

based methods with higher precision@k and lower ranking of unstable non-causal variable. Since by revealing part of causations among variables, they can reduce spurious correlations in training data. But their performances are still worse than our methods in high dimensional settings, since they need enough training data for a better sample reweighting, moreover, they need to observe all causal variables. (iv) Our methods achieve the best performance for the selection of stable variables (with highest precision@k) and screening of unstable non-causal variables (with lowest ranking of unstable non-causal variable).

5.3.4 Results on Stable Prediction.

With the variable ranking list from each algorithm, we select top-k ranked variables to evaluate their performances on stable prediction across unknown test environments, where k is set as the number of causal variables (i.e., $k = p_C = 0.3 * p$. Different k is with similar results, here, we only report the results when $k = p_C$ for saving space). Fig. 3 and Tab. 3 demonstrate the experimental results on stable prediction. From Fig. 3, we find that (i) the performance of our methods are worse than baselines when $r_{test} > 1.5$. This is because the spurious correlation between unstable non-causal variable and the response variable are highly similar between training data ($r_{train} = 2.0$) and test data when $r_{test} > 1.5$, and that spurious correlation can be exploited for improving predictive performance; (ii) the performance of our methods are much better than baseline when $r_{test} < -1.3$, where that spurious correlation are totaly different between training ($r_{train} = 2.0$) and test data $r_{test} < -1.3$, leading to under-performance of baselines on prediction; (iii) our methods achieve the most stable prediction (with smallest Average_Error and Stability_Error) across all test data, since our algorithm can precisely select the causal variables and achieve the lowest ranking of unstable non-causal variable as reported in Tab.1 and Tab. 2.

To clearly demonstrate the advantages of our algorithm on stable prediction, we report the detail results under different synthetic settings in Tab. 3. From the results, we can conclude that our algorithm can make stable prediction across unknown environments via non-causal variables screening and causal variables selection.

5.4 Experiments on Real-World Data

To evaluate the performance of our algorithm in real-world datasets, we apply it to a Parkinson's telemonitoring dataset[4] and a House Pricing prediction dataset [5].

5.4.1 Parkinson's Dataset.

Parkinson's dataset was wildly used for the problem of domain generalization [1], [52] and other regression tasks [53]. This dataset consists of biomedical voice measurements from 42 patients with early-stage Parkinson's disease recruited for a six-month trial of a telemonitoring device for remote symptom progression monitoring. For each patient, there are about 200 recordings, which were automatically recorded in the patients' home. The task is to predict the clinician's motor UPDRS scoring of Parkinson's disease symptoms from patients' features, including their age, gender, test time and many other measures.

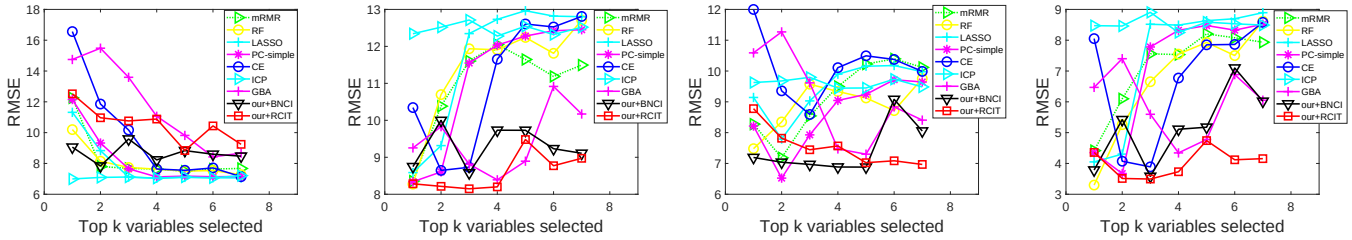
5.4.2 Experimental Settings on Parkinson's dataset.

In our experiments, we set the motor UPDRS scoring as the response variables Y. To test the stability of all methods, we generate different environments by biased data separation based on different patients. Specifically, we separate the whole 42 patients into 4 patients' groups by their order in data, including group 1 (G1) with recordings from 21 patients[6], and other three groups (G2, G3 and G4) are all with recordings from different 7 patients, where the different

4. https://archive.ics.uci.edu/ml/datasets/parkinsons+telemonitoring

5. https://www.kaggle.com/c/house-prices-advanced-regression-techniques

6. Data in group 1 would be continuous separate into training data and test data by patients id, hence, we set 21 patients in group 1.



(a) RMSE on env. G1 (b) RMSE on env. G2 (c) RMSE on env. G3 (d) RMSE on env. G4

Fig. 4: Results of RMSE with top-k selected variables on different environments. All algorithms are trained with data from environment G1, but tested on the data from each environment. When the test environment is different from the training one (e.g., G2, G3, and G4), our algorithm achieves better performance than baselines.

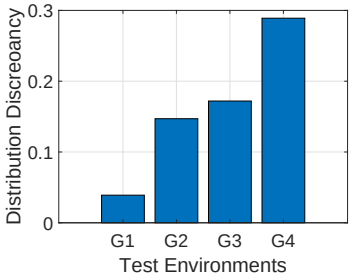


Fig. 5: Distribution discrepancy between the training and test environments by comparing their difference over the mean value of variables.

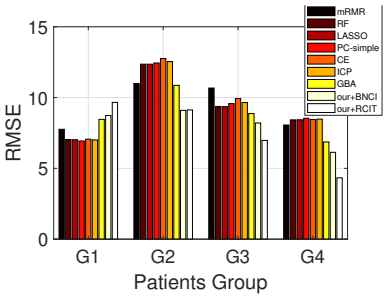


Fig. 6: Prediction across patients' group with the selected top-7 features for each model. Models are trained on datasets where patients are from "G1", but tested on datasets across patients' groups.

groups correspond to different environments[7]. Considering a practical setting where a researcher has a single data set and wishes to train a model that can then be applied to other environments, in our experiments, we trained all models with data from environment G1, but tested them on all 4 groups.

5.4.3 *Experimental Results on Parkinson's dataset.*

Before reporting the results of causal feature selection and prediction, we compute the distribution discrepancy[8] between the training and test environments as shown in Fig. 5. From this result, we can find that the G1 test environment is with very small discrepancy (0.039) with the training environments, but G2, G3 and G4 are with heavy discrepancy with the training environment. Figure 4 demonstrates the experimental results of RMSE with top-k ranked variables with 10 times independent experiments by independent re-selections on test data[9]. Figure 4a shows that correlation based methods (LASSO, mRMR and RF) outperform causation based methods (GBA and our method), this is because the training and test have the similar distribution (small discrepancy) on environment G1, hence the spurious correlation between non-causal variables and response variable can bring positive power for prediction. The main reason might be that the test environment has small discrepancy with training one, thus the spurious correlations might be similar on the training and test data. Hence, the correlation based methods can exploit those spurious correlations to improve their prediction on test data, while the causation based methods reduced those spurious correlation thus achieved poor performance. Moreover, we find ICP method achieves good performance in environment G1 since it cannot differentiate the spurious correlation from only one training environment. Fig. 4b, 4c and 4d demonstrate that causation based methods are better than correlation based methods when the test distributions are out of the training one, and our method, especially the method "our+RCIT", can almost achieve the best performance. The main reason is that spurious correlation on training could be different on testing, while causation based methods could discover causal variables for more stable prediction across environments, and our method performs the best on causal variables ranking and separation. In addition, we observed that in non-i.i.d settings[10], the prediction performance might seriously decrease as inputting more selected variables, since

7. We separate the data by patients id, different patients would have different features distribution and 7 patients in each group cannot guarantee fully randomness, leading to different environments. The best way is to consider each patient as a environment, but it make the sample size of each environment is too small. Hence, we set each 7 patients as an environment.

8. Here, we compute the distribution discrepancy between two environments by directly comparing their difference over the mean value of variables.

9. The variance of RMSE over 10 times independent experiments is about 0.1-0.2 for all algorithms, and we do not plot the variance in Figure 4 for easy reading.

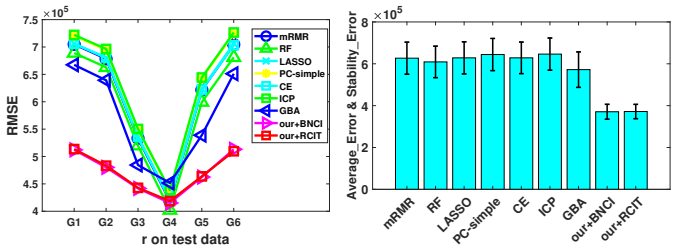
10. The test distribution is different from the training one.

some selected variables could be spuriously correlated with the response and unstable across environments.

Fig. 6 shows the prediction results across patients' groups with the top-7 features from each method. From the results, we can have the similar observations that (i) correlation based methods achieve better performance than causation based method on G1 where training and test data have similar or even the same distribution; (ii) causation based methods (i.e., GBA) obtain better performance on G2, G3 and G4 where the test distribution might be different from the on on training; (iii) our methods achieve the best performance when the test data is from different patients' group of the training.

Moreover, by combining Fig. 4 and Fig. 5, we can find that as the increasing of the discrepancy between test and training environments, the relative improvement of our method is more significant (i.e., the relative improvements of our method on G4 is more significant than G3 and G2 as shown in Fig. 6).

5.4.4 House Pricing Dataset.



(a) RMSE on House Pricing (b) Stability Error on House Pricing

Fig. 7: RMSE and stability error of House price prediction across groups with the selected top-6 features for each model. Models are trained on dataset where the built year of samples falls in [1960,1979] ("G4"), but tested on datasets across other year intervals.

House pricing dataste is a real world regression dataset (Kaggle) of house sales prices from King County, USA, which includes the houses sold between May 2014 and May 2015. We aim to predict the outcome variable as the transaction price of the house, with each sample containing 16 predictive variables (e.g., the built year of the house, number of bedrooms, number of bathrooms, and square footage of home etc).

5.4.5 Experimental Settings on House Pricing dataset.

In our experiments, we set the transaction price of the house as the response variables Y. Similar to aforementioned approach in Parkinson's dataset, we create heterogeneous environments via biased data separation. More specifically, we split the total dataset into 6 groups/periods with each group approximately covering over a time span of two decades on the built year variable. Consequently, we refer the 6 generated environments as G1 to G6, and we trained all models on samples from environment G4 with built year falling in [1960,1979] ("G4"), but tested them on all 6 groups.

5.4.6 Experimental Results on House Pricing dataset.

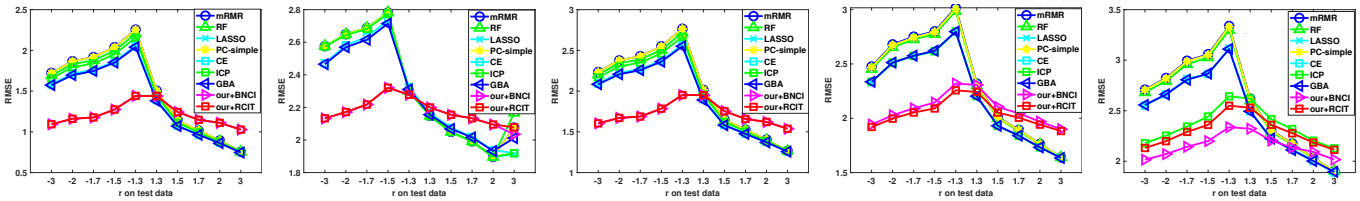
We first report the prediction results on RMSE and stability error with top-6 features for each method in Fig. 7a and 7b (Prediction results with other number of top features are omitted since the prediction fluctuation comparing to their magnitude is small). Specifically, Fig. 7a explicitly shows that under Out-of-distribution setting, our method dominates others, especially when the testing distribution shift far away from training environment (G4). The comparison on the stability performance also coincides with our motivation: selection causal variables promotes stable prediction towards varying testing environments.

5.5 Discussion on the experimental results

Importantly, our experiments indicates an interesting phenomenon that prediction performance by causal predictors is not as good as some naive regression methods (e.g., LASSO, CE) when the distribution discrepancy between training and testing environments is small, which is shown in Fig. 3a, Fig 7a and Fig 6. The underlying reason reflects that causal prediction achieves trade-off between i.i.d performance and Out-of-distribution (OOD) performance (which is suitable for any OOD prediction methods). In other words, causal prediction is "stable but conservative", which indeed sacrifices some i.i.d performance as the compensation for promotion on OOD performance. More specifically, if the training distribution P_{te} is i.i.d or very close to testing distribution P_{tr}, then straightforward prediction will outperform causal prediction, as the bias itself is informative for i.i.d prediction. In such cases, biased estimator (naive prediction) overfits on biased data and of course outperforms other methods. In contrast, most realistic cases indicate that P_{te} shifts away from P_{tr} and the causal prediction methods could dominate non-causal estimators. This is achieved by using stable causal features for prediction, where the unstable bias is eliminated during causal feature selection. For the theoretical guarantee behind this phenomenon, we refer readers to one recent paper [54], which characterizes this phenomenon using worst-case optimization.

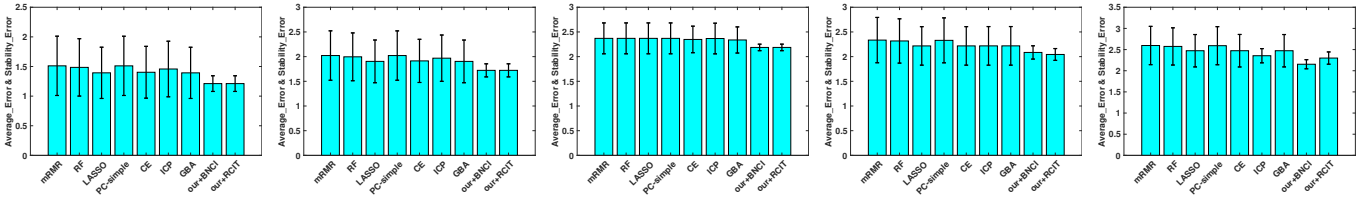
5.6 Robustness Verification

It is noteworthy that our method has to rely on the structural assumption that the causal and non-causal features should be marginally independent. However, we also note that fact that every robust/OOD method have their own prior assumption: there is no method can generalize on arbitrary latent data for testing. For instance, the well-known Invariant Risk Minimization (IRM) [55] assumes a linear classifier on top of the representations with enough number of training domains. Meanwhile, the long-standing problem named Distributionally Robust Optimization (DRO) [45] also assumes a pre-defined uncertainty set of distributions and generalization only happens when testing distribution is in/close to such uncertainty set. Concerning causal model for OOD generalization, the famous invariant causal prediction (ICP) [56] also assumes the linear structural equation and enough interventional training domains.



(a) $\gamma = 0.1$ (b) $\gamma = 0.2$ (c) $\gamma = 0.3$ (d) $\gamma = 0.4$ (e) $\gamma = 0.5$

Fig. 8: Robust test on RMSE across unknown test data with n = 2000, p = 20 and $p_L = 10$. All methods are trained with r_{train} = 2.0, but tested across test environments with different r_{test}.



(a) $\gamma = 0.1$ (b) $\gamma = 0.2$ (c) $\gamma = 0.3$ (d) $\gamma = 0.4$ (e) $\gamma = 0.5$

Fig. 9: Robust prediction on average and stability error across unknown test data with n = 2000, p = 20 and $p_L = 10$. All methods are trained with r_{train} = 2.0, but tested across test environments with different r_{test}.

To further analyze the robustness of our method when confronting with moderate violation of our prior assumption, we perform robustness analysis on synthetic data. Specifically, we generate more challenge data by inducing correlation between a subset of linked non-causal variable $\mathbf{L}$ and selection bias S, where the size of subset is controlled by varying its proportion γ to the total set $\mathbf{L}$ (Here we set $p_L = 10$). We report the results averaged for 50 times when $p = 20$ and $n = 2000$ in Fig. 8 and Fig 9. These two figures clearly reflects two facts: (a) The prediction error of each method increases as the proportion γ of biased $\mathbf{L}$ enlarges. (b) When the number of biased $\mathbf{L}$ is smaller than 0.5, our method still outperforms other methods, especially when distributional shift is large ($r_{test} \leq -1.3$). (c) When γ increases, the performance margin between our method and other decreases (from over 1 to less than 0.5).

6 Conclusion

In this paper, we focus on the problem of stable prediction with leveraging a seed variable for non-causal variables screening and causal variable selection. We argue that most of traditional prediction methods and variable selection methods are correlation based, resulting in instability problem on prediction across unknown environments. Based on conditional independence (CI) test techniques, we proposed a causal variable selection algorithm via screening out isolated non-causal variables with a single CI test per variable, and provide a series of theorems and empirical experiments to prove that our algorithm can precisely screen out the isolated non-causal variables to increase the stability of model on prediction across unknown test data. The experimental results on both synthetic and real-world datasets show that our algorithm outperforms the baselines for stable prediction across unknown test data.

Acknowledgment

This work was supported in part by National Key Research and Development Program of China (No.2018AAA0101900), the Young Elite Scientists Sponsorship Program by CAST (No. 2021QNRC001), Key R&D Projects of the Ministry of Science and Technology (No.2020YFC0832500), National Natural Science Foundation of China (No.62006207, No. 62037001, No.72171131), the Tsinghua University Initiative Scientific Research Grant (No.2019THZWJC11), Project by Shanghai AI Laboratory (P22KS00111), Technology and Innovation Major Project of the Ministry of Science and Technology of China (No.2020AAA0108400, No.2020AAA0108403).

References

[1] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, 2013, pp. 10–18.

[2] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5542–5550.

[3] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, "Invariant models for causal transfer learning," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1309–1342, 2018.

[4] J. Zhang and E. Bareinboim, "Transfer learning in multi-armed bandit: a causal approach," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017, pp. 1778–1780.

[5] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction: identification and confidence intervals," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 5, pp. 947–1012, 2016.

[6] N. Pfister, P. Bühlmann, and J. Peters, "Invariant causal prediction for sequential data," *Journal of the American Statistical Association*, vol. 114, no. 527, pp. 1264–1276, 2019.

[7] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1617–1626.

[8] K. Kuang, R. Xiong, P. Cui, S. Athey, and B. Li, "Stable prediction with model misspecification and agnostic distribution shift," in *Thirty-Fouth AAAI Conference on Artificial Intelligence*, 2020.

[9] Z. Shen, P. Cui, J. Liu, T. Zhang, B. Li, and Z. Chen, "Stable learning via differentiated variable decorrelation," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2185–2193.

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

[11] E. Bareinboim and J. Pearl, "Controlling selection bias in causal inference," in *Artificial Intelligence and Statistics*, 2012, pp. 100–108.

[12] J. D. Correa, J. Tian, and E. Bareinboim, "Identification of causal effects in the presence of selection bias," in *AAAI*, vol. 33, no. 01, 2019, pp. 2744–2751.

[13] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 8, pp. 1226–1238, 2005.

[15] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[16] J. Ramsey, P. Spirtes, and J. Zhang, "Adjacency-faithfulness and conservative causal inference," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 2006, pp. 401–408.

[17] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

[19] K. Kuang, L. Li, Z. Geng, L. Xu, K. Zhang, B. Liao, H. Huang, P. Ding, W. Miao, and Z. Jiang, "Causal inference," *Engineering*, vol. 6, no. 3, pp. 253–263, 2020.

[20] S. Athey, G. W. Imbens, and S. Wager, "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 4, pp. 597–623, 2018.

[21] K. Kuang, P. Cui, B. Li, M. Jiang, and S. Yang, "Estimating treatment effect in the wild via differentiated confounder balancing," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 265–274.

[22] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[23] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[24] P. Bühlmann, M. Kalisch, and M. H. Maathuis, "Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm," *Biometrika*, vol. 97, no. 2, pp. 261–278, 2010.

[25] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, "Causality-based feature selection: Methods and evaluations," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–36, 2020.

[26] A. Mastakouri, B. Schölkopf, and D. Janzing, "Selecting causal brain features with a single conditional independence test per feature," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 532–12 543.

[27] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[29] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1415–1438, 2003.

[30] S. Nakariyakul and D. P. Casasent, "An improvement on floating search algorithms for feature subset selection," *Pattern Recognition*, vol. 42, no. 9, pp. 1932–1940, 2009.

[31] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, no. 06, pp. 903–929, 2003.

[32] E. Romero and J. M. Sopena, "Performing feature selection with multilayer perceptrons," *IEEE Transactions on Neural Networks*, vol. 19, no. 3, pp. 431–441, 2008.

[33] Y. Peng, Z. Xuefeng, Z. Jianyong, and X. Yumhong, "Lazy learner text categorization algorithm based on embedded feature selection," *Journal of Systems Engineering and Electronics*, vol. 20, no. 3, pp. 651–659, 2009.

[34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine learning research*, vol. 5, no. Nov, pp. 1531–1555, 2004.

[36] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[37] C. Lu and S. Wang, "The general-purpose intelligent agent," *Engineering*, vol. 6, no. 3, pp. 221–226, 2020.

[38] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020.

[39] Q. Tian, K. Kuang, K. Jiang, F. Wu, and Y. Wang, "Analysis and applications of class-wise robustness in adversarial training," *KDD*, 2021.

[40] J. Liu, Z. Shen, P. Cui, L. Zhou, K. Kuang, B. Li, and Y. Lin, "Stable adversarial learning under distributional shifts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8662–8670.

[41] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, vol. 49, no. 3, pp. 1378–1406, 2021.

[42] Z. Shen, P. Cui, T. Zhang, and K. Kuang, "Stable learning via sample reweighting," in *Thirty-Fouth AAAI Conference on Artificial Intelligence*, 2020.

[43] P. Cui and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 110–115, 2022.

[44] J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen, "Heterogeneous risk minimization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6804–6814.

[45] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *International Conference on Learning Representations*, 2019.

[46] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[47] D. Geiger, T. Verma, and J. Pearl, "Identifying independence in bayesian networks," *Networks*, vol. 20, no. 5, pp. 507–534, 1990.

[48] R. Sen, A. T. Suresh, K. Shanmugam, A. G. Dimakis, and S. Shakkettai, "Model-powered conditional independence test," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2955–2965.

[49] A. Marx and J. Vreeken, "Testing conditional independence on discrete data using stochastic complexity," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 496–505.

[50] M. Scutari *et al.*, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, no. i03, 2010.

[51] E. V. Strobl, K. Zhang, and S. Visweswaran, "Approximate kernel-based conditional independence tests for fast non-parametric causal discovery," *Journal of Causal Inference*, vol. 7, no. 1, 2019.

[52] G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," *arXiv preprint arXiv:1711.07910*, 2017.

[53] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2009.

[54] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters, "Anchor regression: Heterogeneous data meet causality," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 83, no. 2, pp. 215–246, 2021.

[55] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

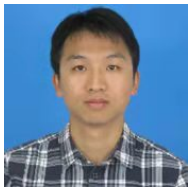
[56] C. Heinze-Deml, J. Peters, and N. Meinshausen, "Invariant causal prediction for nonlinear models," *Journal of Causal Inference*, vol. 6, no. 2, 2018.

Kun Kuang received his Ph.D. degree from Tsinghua University in 2019. He is now an Associate Professor in the College of Computer Science and Technology, Zhejiang University. He was a visiting scholar with Prof. Susan Athey's Group at Stanford University. His main research interests include Causal Inference, Artificial Intelligence, and Causally Regularized Machine Learning. He has published over 40 papers in major international journals and conferences, including SIGKDD, ICML, ACM MM, AAAI, IJCAI, TKDE, TKDD, Engineering, and ICDM, etc.

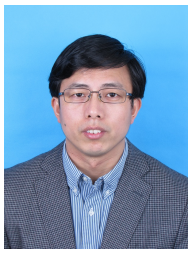
Haotian Wang received his B.E. degree in computer science and technology from National University of Defense Technology, Changsha, China, in 2017. He is currently pursuing the Doctor degree in computer science and technology with the National University of Defense Technology, Changsha, China. His research interests include machine learning and causal inference.

Yue Liu received his Ph.D. degree from Peking University in 2019. He is now an assistant professor in the Center for Applied Statistics and School of Statistics, Renmin University of China. His main research interests include Causal Inference and Trustworthy Machine Learning.

Ruoxuan Xiong completed her Ph.D. in Management Science and Engineering from Stanford University in 2020. Dr. Xiong was a postdoctoral fellow at the Stanford Graduate School of Business. Dr. Xiong is now an assistant professor in the Department of Quantitative Theory and Methods at Emory University. Dr. Xiong's research is at the intersection of econometrics and operations research, focusing on factor modeling, causal inference, and experimental design, and with applications in finance and healthcare. Her articles have been featured in the Journal of Econometrics, Journal of Business and Economic Statistics, eLife, KDD, and AAAI.



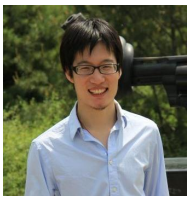
Runze Wu received the Ph.D. degree and B.E. degree from the University of Science and Technology of China, Hefei, China. He is currently a senior AI expert and supervise User Profiling Research Group in NetEase Fuxi AI Lab, Hangzhou, China. His research interests include user profiling, anomaly detection, causal inference, combinational optimization, deep learning, and various applications across online games. He has published more than 20+ papers in refereed journals and conference proceedings, such as TOIS, TKDD, TIST, KDD, IJCAI, AAAI, WWW, MM, and ICDE, etc.



Weiming Lu received his Ph.D. degree from Zhejiang University in 2009. He is now an Associate Professor in the College of Computer Science and Technology, Zhejiang University. He was a visiting scholar at Imperial College London. His main research interests include Natural Language Processing, Cross-media analysis, and Artificial Intelligence. He has published over 50 papers in major international conferences and journals including ACL, WWW, EMNLP, AAAI, IJCAI, ACM MM, TKDE and TMM, etc.

Yueting Zhuang received the B.S., M.S., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1986, 1989, and 1998, respectively. From 1997 to 1998, he was a Visitor with the Department of Computer Science and the Beckman Institute, University of Illinois at Urbana Champaign, Champaign, IL, USA. He is currently a Full Professor with the College of Computer Science, Zhejiang University, Hangzhou, China. His current research interests include multimedia databases, artificial intelligence, and video-based animation.

Fei Wu received his PhD degree from the College of Computer Science at Zhejiang University. Now he is the professor and dean of the College of Computer Science, Zhejiang University. His main research interests include multimedia information analysis and retrieval, digital library.



Peng Cui received his Ph.D. degree in computer science in 2010 from Tsinghua University and he is an Associate Professor at Tsinghua. He has vast research interests in data mining, multimedia processing, and social network analysis. Until now, he has published more than 60 papers in conferences such as SIGIR, AAAI, ICDM, etc. and journals such as IEEE TMM, IEEE TIP, DMKD, etc. He won 5 best paper awards in recent 4 years, including ICDM2015 Best Student Paper Award, ICME 2014 Best Paper Award, etc. In 2015, he was awarded as ACM China Rising Star. Now his research is sponsored by National Science Foundation of China, Samsung, Tencent, etc. He also serves as Guest Editor, Co-Chair, PC member, and Reviewer of several high-level international conferences, workshops, and journals.

Bo Li received a Ph.D degree in Statistics from the University of California, Berkeley, and a bachelor's degree in Mathematics from Peking University. He is an Associate Professor at the School of Economics and Management, Tsinghua University. His research interests are statistical methods for high-dimensional data, statistical causal inference and data-driven decision making. He has published widely in academic journals across a range of fields including statistics, management science and economics.