Mining Invariance from Nonlinear Multi-Environment Data: Binary Classification

Austin Goddard, Kang Du, Yu Xiang

Department of Electrical and Computer Engineering

University of Utah

{austin.goddard, kang.du, yu.xiang}@utah.edu

Abstract—**Making predictions in an unseen environment given data from multiple training environments is a challenging task. We approach this problem from an invariance perspective, focusing on binary classification to shed light on general nonlinear data generation mechanisms. We identify a unique form of invariance that exists solely in a binary setting that allows us to train models invariant over environments. We provide sufficient conditions for such invariance and show it is robust even when environmental conditions vary greatly. Our formulation admits a causal interpretation, allowing us to compare it with various frameworks. Finally, we propose a heuristic prediction method and conduct experiments using real and synthetic datasets.**

I. Introduction

It is common practice to collect observations of a set of features $X = (X_1, \ldots, X_m)$ and response Y from different environments to train a model. The prediction of the response in an unseen environment is often referred to as multi-environment domain adaptation, with practical applications in various fields (e.g., genetics [1] and healthcare [2]). A common assumption in such problems is the principle of invariance, modularity, or autonomy [3]–[8]. This invariance assumption states that the conditional distribution of Y given X is invariant with respect to different environment.

The invariant causal prediction (ICP) framework [9], along with its various extensions [10], [11], employ the invariance principle to identify invariant predictors across environments. Following this framework, various domain adaptation approaches have been developed [12]–[14]. Specifically, the stabilized regression (SR) [14] approach relies on a weaker form of invariance dependent on expectation as opposed to probability. The common assumption for the approaches mentioned is that the assignment of Y does not change over environments. In a causal sense, from which much of the literature in this area stems, this is referred to as an *intervention* on Y [8]. When Y is intervened, the invariance principle, as well as the frameworks mentioned above, fail. In a series of recent works [15], [16], an alternative approach called the invariant matching property (IMP) has been developed to detect *linear* invariant models in a *regression* setting even when the assignment of Y is altered over environment.

In this work, we extend general principles developed in [15], [16] to the binary classification setting as an attempt to generalize to nonlinear settings. The proposed approach works even when data-generating models change over environments (e.g., Y can be generated using a probit model for one environment and a logistic model in another). Additionally, the approach is not constrained by the data type, meaning it can be useful on continuous, discrete, or categorical variables.

II. Problem Formulation

Consider the following setting. For different environmental conditions indexed by the set $\mathcal{E}$, we have a random vector $X = (X_1, \ldots, X_m)$ and a binary random variable Y whose elements form a joint distribution $\mathcal{P}_e := \mathcal{P}_e^{X,Y}$ dependent on $e \in \mathcal{E}$. Denote X and Y as X^e and Y^e for a specific $e \in \mathcal{E}$, respectively. The supports of X and Y are $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \{0, 1\}$, respectively. Let X_S be a random vector containing the elements in X indexed by the set $S \subseteq \{1, \ldots, m\}$, and let $\mathcal{X}_S$ be its support. To simplify notation, let $X_0^e := Y^e$. For each $e \in \mathcal{E}$, we keep the distribution $\mathcal{P}_e$ general, with the exception that there exists an X_i^e generated according to the form

$$X_i^e = g(X_{S_i}^e) + \epsilon^e, \text{ for some } i \in \{1, \ldots, m\}, \qquad (1)$$

where $X_{S_i}^e$, for $S_i \subseteq \{0, \ldots, m\} \backslash i$, represents the variables that directly effect X_i^e, and ϵ^e is an independent, zero mean, noise variable. We assume the output of the function g is not constant with regards to any of its inputs; g is a constant function when $S_i = \varnothing$.

Additionally, while the function g does not change over environment (i.e., does not depend on e), the distribution of ϵ^e can change arbitrarily as long as the mean of the distribution remains zero. Aside from a binary Y and the form of X_i^e in (1), we make no assumptions on the distribution or functional form of any variable. As such, this formulation applies to any set of features, be it continuous, discrete, or a mixture of the two.

We assume only a subset of all environments are observed and denote this set by $\mathcal{E}_{\text{obs}} \subseteq \mathcal{E}$. Where $\mathcal{E}_{\text{obs}} = \mathcal{E}_{\text{train}} \cup \{e^{\text{test}}\}$, and $Y^{\text{test}} := Y^{e^{\text{test}}}$, our goal is to make predictions on Y^{test}, given a set of training environments $\mathcal{E}_{\text{train}}$. As such, we aim to find a function $\phi_e : \mathcal{X} \to \mathcal{W}$ such that, the probability of Y given $\phi_e(X)$ does not vary over any environment. Specifically, for all $w \in \mathcal{W}$ and $e, h \in \mathcal{E}_{\text{obs}}$,

$$\mathcal{P}_e(Y|\phi_e(X) = w) = \mathcal{P}_h(Y|\phi_h(X) = w). \qquad (2)$$

As Y is binary, it is equivalent to write (2) in the form: $\mathsf{E}_{\mathcal{P}_e}[Y|\phi_e(X) = w] = \mathsf{E}_{\mathcal{P}_h}[Y|\phi_h(X) = w]$, for all $w \in \mathcal{W}$ and $e, h \in \mathcal{E}_{\text{obs}}$. It is well-known that (2) is satisfied if $\phi_e(X) = X_{S_Y}$ and for $S_Y \subseteq \{1, \ldots, m\}$,

$$Y^e = f(X_{S_Y}^e) + \epsilon_Y, \qquad (3)$$

where ϵ_Y is an independent noise that does not vary over environment [9]. However, we are interested in a more general setting where *the function f and distribution of the noise can vary over environment*. From a causal perspective, this would indicate that Y^e had been *intervened* (see Section IV-A). In such a setting, $\phi_e(X) = X_{S_Y}$ is no longer useful and other approaches must be considered. We now consider one such alternative, starting with a motivating example.

III. MOTIVATING EXAMPLE

Consider the following setting with $X^e = (X_1^e, X_2^e, X_3^e)$. Let X_1^e and X_2^e be independent and follow $X_1^e \sim \mathcal{N}(\mu_1^e, \sigma_1^2)$ and $X_2^e \sim \mathcal{N}(\mu_2^e, \sigma_2^2)$. The variable Y^e is generated such that $Y^e | X_1^e, X_2^e$ forms a probit model. Specifically,

$$Y^e = \begin{cases} 1, & \text{if } \beta_1^e X_1^e + \beta_2 X_2^e + \epsilon_Y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Following a similar form as (1), X_3^e is linear given Y^e so that

$$X_3^e = \begin{cases} \gamma_1 X_1^e + \epsilon_3, & \text{if } Y^e = 1, \\ \gamma_0 X_1^e + \epsilon_3, & \text{if } Y^e = 0. \end{cases}$$

The noise variables ϵ_Y and ϵ_3 are i.i.d. $\mathcal{N}(0, \sigma^2)$. Suppose we wish to predict Y^e given only X_1^e. Predicting Y^e for a particular $e \in \mathcal{E}$ becomes difficult as β_1^e and μ_2^e vary with environment. Specifically,

$$\mathsf{E}_{\mathcal{P}_e}[Y|X_1 = x_1] = \Phi\left(\frac{\beta_1^e x_1 + \beta_2 \mu_2^e}{\sqrt{(\beta_2 \sigma_2)^2 + \sigma^2}}\right), \qquad (4)$$

where Φ is the cumulative distribution function of a standard normal random variable. As (4) varies over environment, it is not practical to use $\mathsf{E}_{\mathcal{P}_e}[Y|X_1]$ to estimate Y^e on different environments. Even while conditioning on both X_1^e and X_2^e (the variables that directly affect Y^e), the variance (w.r.t. environment) still remains through β_1^e.

We can, however, decompose (4) into various variant and invariant components such that $\mathsf{E}_{\mathcal{P}_e}[Y|X_1 = x_1]$ becomes the following (see the proof of Proposition 1 for a general case),

$$\frac{\mathsf{E}_{\mathcal{P}_e}[X_3|X_1 = x_1] - \mathsf{E}_{\mathcal{P}_e}[X_3|X_1 = x_1, Y = 0]}{\mathsf{E}_{\mathcal{P}_e}[X_3|X_1 = x_1, Y = 1] - \mathsf{E}_{\mathcal{P}_e}[X_3|X_1 = x_1, Y = 0]}, \quad (5)$$

where $\mathsf{E}_{\mathcal{P}_e}[X_3|X_1 = x_1]$ is

$$\Phi\left(\frac{\beta_1^e x_1 + \beta_2 \mu_2^e}{\sqrt{(\beta_2 \sigma_2)^2 + \sigma^2}}\right)(\gamma_1 - \gamma_0)x_1 + \gamma_0 x_1, \quad (6)$$

and $\mathsf{E}_{\mathcal{P}_e}[X_3|X_1 = x_1, Y = y]$ is $\gamma_1 x_1$ if $y = 1$ and $\gamma_0 x_1$ if $y = 0$. We note that the variance (w.r.t environment) contributed by β_1^e and μ_2^e is completely accounted for in the term $\mathsf{E}_{\mathcal{P}_e}[X_3|X_1]$ and that $\mathsf{E}_{\mathcal{P}_e}[X_3|X_1, Y]$ is invariant over environment. Thus, (2) holds for the function $\phi_e(X) = (X_1, \mathsf{E}_{\mathcal{P}_e}[X_3|X_1])$. In addition to this, we also note that conditioning on both X_1 and X_2 leads to a similar invariance; we only condition on X_1 in this example for simplicity.

This invariance does not hold if we replace X_3^e with any other variable. For example, suppose we were to estimate Y^e, replacing X_3^e with X_2^e. We can still decompose (4) similarly to (5) by replacing X_3^e with X_2^e. As $\mathsf{E}_{\mathcal{P}_e}[X_2|X_1] = \mu_2^e$ does not contain β_1^e, the portion of $\mathsf{E}_{\mathcal{P}_e}[Y|X_1]$ that contains β_1^e must reside in $\mathsf{E}_{\mathcal{P}_e}[X_2|X_1, Y]$. i.e., $\mathsf{E}_{\mathcal{P}_e}[X_2|X_1, Y]$ is not invariant over environments as is $\mathsf{E}_{\mathcal{P}_e}[X_3|X_1, Y]$. Thus, the function $\phi_e(X) = (X_1, \mathsf{E}_{\mathcal{P}_e}[X_2|X_1])$ will no longer satisfy (2).

To further illustrate the difference in selecting X_3^e over X_2^e, suppose we wish to estimate on a new environment e^{test}. While we have access to X^{test}, we can easily construct $\mathsf{E}_{\mathcal{P}_{\text{test}}}[X_i|X_1]$ for either $i \in \{2, 3\}$. We cannot, however, use Y^{test} to construct our estimate, and $\mathsf{E}_{\mathcal{P}_{\text{test}}}[X_i|X_1, Y]$ must be obtained by leveraging invariances over environment. Thus, for either $i \in \{2, 3\}$, we construct the estimate

$$\hat{Y}_i^{\text{test}} =: \frac{\mathsf{E}_{\mathcal{P}_{\text{test}}}[X_i|X_1] - \mathsf{E}_{\mathcal{P}_e}[X_i|X_1, Y = 0]}{\mathsf{E}_{\mathcal{P}_e}[X_i|X_1, Y = 1] - \mathsf{E}_{\mathcal{P}_e}[X_i|X_1, Y = 0]}, \quad (7)$$

where $e \in \mathcal{E}_{\text{train}}$. As $\mathsf{E}_{\mathcal{P}_e}[X_3|X_1, Y]$ is invariant and $\mathsf{E}_{\mathcal{P}_e}[X_2|X_1, Y]$ is not invariant as discussed above, $\hat{Y}_3^{\text{test}}$ will provide a good estimate of Y^{test}, while $\hat{Y}_2^{\text{test}}$ will not.

In Fig. 1 we compare $\hat{Y}_3^{\text{test}}$ and $\hat{Y}_2^{\text{test}}$ by simulating $(x^{\text{test}}, y^{\text{test}})$ pairs for a set of specific parameters. The estimate $\hat{Y}_2^{\text{test}}$ does not fit the data as many x_1^{test} corresponding to $y^{\text{test}} = 0$ will be incorrectly classified to one. However, this is not the case when $\hat{Y}_3^{\text{test}}$ is used, and the fit is greatly improved (Fig. 1). The poor fit on $\hat{Y}_2^{\text{test}}$ is a result of $\mathsf{E}_{\mathcal{P}_e}[X_2|X_1, Y]$ varying across environments.
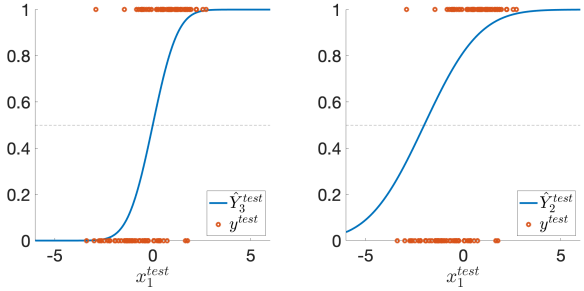

Fig. 1: Comparisons of $\hat{Y}_3^{\text{test}}$ (left) and $\hat{Y}_2^{\text{test}}$ (right), where $\beta_1^e = 2$, $\mu_2^e = 1$, $\beta_2^{\text{test}} = 0$, and $\mu_2^{\text{test}} = -1$.

IV. THE BINARY INVARIANT MATCHING PROPERTY

A deterministic relationship such as the one in (5) has been previously referred to as *matching* [15], and can be generalized to the formulation outlined in Section II.

Definition 1. *For $k \in \{1, \ldots, m\}$, $S \subseteq \{1, \ldots, m\} \backslash k$, and $h(X_S, Y) := \mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$, the pair (k, S) satisfies the binary invariant matching property (bIMP)[1] if,*

$$\mathsf{E}_{\mathcal{P}_e}[Y|X_S] = \frac{\mathsf{E}_{\mathcal{P}_e}[X_k|X_S] - h(X_S, 0)}{h(X_S, 1) - h(X_S, 0)}, \quad (8)$$

holds for all $e \in \mathcal{E}_{obs}$, where $h(X_S, Y)$ does not depend on e.

As seen in the example, there are a variety of choices for k and S, not all of which lead to invariant representations. We

---

[1]There are *degenerate* cases when $h(X_S, 0) = h(X_S, 1)$, for which the tower property implies $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S] = \mathsf{E}_{\mathcal{P}_e}[h(X_S, Y)|X_S] = h(X_S, 0)$, and the ratio in (8) reduces to 0 divided by 0.

now detail the sufficient conditions for which a pair (k, S) satisfies the bIMP (see Appendix for the proof).

Proposition 1. *Let $k \in \{1, \ldots, m\}$ and $S = R \cup Q$ where $R, Q \subseteq \{1, \ldots, m\} \setminus k$ and $R \cap Q = \varnothing$. The pair (k, S) satisfies the bIMP if, for every $e \in \mathcal{E}_{obs}$,*
  1) $X_k^e = g(X_R^e, Y^e) + \epsilon^e$ *as in* (1) ,
  2) $X_Q^e \perp\!\!\!\perp X_k^e \mid (X_R^e, Y^e)$.

What remains is to show that the bIMP can be used to satisfy the invariance principle in (2), and thus, can be beneficial in predicting on unknown environments, as shown below.

Theorem 1. *Let $k \in \{1, \ldots, m\}$ and $S = R \cup Q$ where $R, Q \subseteq \{1, \ldots, m\} \setminus k$ and $R \cap Q = \varnothing$. When $\phi_e(X) = (X_R, X_Q, \mathsf{E}_{\mathcal{P}_e}[X_k | X_R, X_Q])$, (2) holds if the pair (k, S) satisfies the bIMP.*

Proof. Let $\ell^e(X_R, X_Q) := \mathsf{E}_{\mathcal{P}_e}[X_k | X_R, X_Q]$ and $\phi_e(X) = (X_R, X_Q, \ell^e(X_R, X_Q))$. Since (k, S) satisfies the bIMP and $\ell^e(X_R, X_Q)$ is a function of X_R and X_Q,

$$
\begin{aligned}
\mathsf{E}_{\mathcal{P}_e} & [Y | \phi_e(X) = (x_Q, x_R, z)] \\
&= \mathsf{E}_{\mathcal{P}_e}[Y | X_R = x_R, X_Q = x_q, \ell^e(X_R, X_Q) = z] \\
&= \frac{z - g(x_R, 0)}{g(x_R, 1) - g(x_R, 0)}.
\end{aligned}
\tag{9}
$$

Thus, (2) holds as (9) does not vary over $e \in \mathcal{E}_{obs}$. $\square$

Remark 1. *In this work, we focus specifically on settings where Y^e is binary. However, there does exist a corresponding matching property with sufficient conditions similar to those in Proposition 1 for cases when Y^e is multi-class or continuous. We leave the analysis for the long version of this work.*

A. A Causal Perspective

While the sufficient conditions in Theorem 1 may seem abstract, we now show that, in fact, they have a specific meaning in a causal sense. To do so, we introduce the structural causal model (SCM) [8]. Here, X^e and Y^e are part of an SCM $\mathcal{S}^e$ that varies over environment such that

$$
\mathcal{S}^e : \begin{cases} Y^e := f_Y^e(X_{PA(Y^e)}^e, \; \epsilon_Y^e), \\ X_1^e := f_1^e(X_{PA(X_1^e)}^e, \; \epsilon_1^e), \\ \quad \vdots \\ X_m^e := f_m^e(X_{PA(X_m^e)}^e, \; \epsilon_m^e). \end{cases}
\tag{10}
$$

where $\epsilon_1^e \ldots \epsilon_m^e, \epsilon_Y^e$ are independent noise variables. To simplify notation, let $X_0^e := Y^e$. Thus, $PA(X_i^e) \subseteq \{0, \ldots, 1\}$ denotes the set indexed by the direct causal parents of X_i^e for all $i \in \{0, \ldots, m\}$.

As in Section II, Y^e is binary. Additionally, at least one structural assignment (i.e., $f_i^e(\cdot)$) in $\mathcal{S}^e$ is an additive noise function that does not vary over environment. Specifically, for some $i \in \{0, \ldots, m\}$, let $f_i^e(X_{PA(X_i^e)}^e, \epsilon_i^e) = g(X_{PA(X_i^e)}^e) + \epsilon_i^e$, where ϵ_i^e has zero mean. An intervention on a variable from $\{X_1^e, \ldots, X_m^e, Y^e\}$ occurs if the structural assignment changes for some $e \in \mathcal{E}$. Relating the SCM to the formation in Section II gives insight into the types of interventions that

may occur. While many methods [9], [14], [15] make various assumptions on the types of interventions (e.g., shifts in the mean or variance), the setting in (10) allows for very general interventions, including general interventions on Y^e, which many other approaches do not allow.

Given $\mathcal{S}^e$ for all $e \in \mathcal{E}_{obs}$, we can express the conditions of Proposition 1 in the language of SCMs, detailed below.

Corollary 1. *Let $k \in \{1, \ldots, m\}$ and $S = R \cup Q$ where $R, Q \subseteq \{1, \ldots, m\} \setminus k$ and $R \cap Q = \varnothing$. For the SCM $\mathcal{S}^e$, the pair (k, S) satisfies the bIMP for all $e \in \mathcal{E}_{obs}$ if the following cases hold.*
  1) $X_k^e = g(X_{PA(X_k^e)}^e) + \epsilon_k^e$,
  2) X_R^e *and Y^e constitute the parents of X_k^e* ,
  3) *The variables in X_Q^e can be any non-descendants of X_k^e.*

The first condition in Proposition 1 is analogous to the first and second condition above as $PA(X_k^e) = (X_S, Y)$. Additionally, in an SCM, any variable conditioned on its parents is independent of any non-descendant. As such, the set X_Q^e can be any non-descendant of X_k^e, bridging the final conditions in Proposition 1 and Corollary 1.

In many cases, the set Q can be quite inclusive despite what may seem like a strong independence condition in Proposition 1. In Corollary 1, we learn that, in a causal sense, X_Q^e can be any non-descendant of X_k^e. For example, if half of the predictors in an SCM are ancestors of Y^e, while the other half are descendants, then the set Q indexes at least half of all predictors (and potentially many more).

V. PROPOSED METHOD

For each $e \in \mathcal{E}_{train}$, we have n^e samples, represented as a matrix $\boldsymbol{X}^e \in \mathbb{R}^{n_e \times m}$, and a vector $\boldsymbol{Y}^e \in \{0, 1\}^{n_e}$ (see [17] for a discussion on the impact of different environments). Additionally, we have n_{test} samples in the test environment, and we denote $\boldsymbol{X}^{test} \in \mathbb{R}^{n_{test} \times m}$ and $\boldsymbol{Y}^{test}$ as the predictor matrix and target vector for the environment e^{test}, respectively. We denote $\boldsymbol{X}$ as the pooled predictor matrix over all $e \in \mathcal{E}_{train}$, and $\boldsymbol{X}_{Y=y}$ as the matrix comprising the rows of $\boldsymbol{X}$ in which $Y = y$, for $y \in \{0, 1\}$. Let $\boldsymbol{X}^{-e}$ be the matrix of samples indexed only by those samples not in $e \in \mathcal{E}_{train}$.

We now leverage insights gained from Theorem 1 and the bIMP to develop a practical method for estimation in unknown environments. At test time, we do not have access to Y^{test}. As such, one cannot say with definitive assurance that (2) holds for all $e \in \mathcal{E}_{obs}$. Thus, the best that can be done in such settings is to identify a ϕ_e such that (2) holds for all $e \in \mathcal{E}_{train}$, implying that $\mathcal{E}_{train}$ must have at least two environments.

Thus, our goal in a practical setting is to identify (k, S) pairs that may satisfy the bIMP overall $e \in \mathcal{E}_{train}$. Simply put, we test whether $\mathsf{E}_{\mathcal{P}_e}[X_k | X_S, Y]$ is invariant. To do so, we consider a special form of the model in (1) where $X_k^e = g(X_S^e, Y^e) + \epsilon^e$ with $\epsilon^e \sim \mathcal{N}(0, (\sigma^e)^2)$ is assigned a different nonlinear additive noise function for each value of Y^e. Specifically,

$$
g(X_S^e, Y^e) = \begin{cases} g_1(X_S^e), & \text{if } Y^e = 1, \\ g_0(X_S^e), & \text{if } Y^e = 0. \end{cases}
\tag{11}
$$

As X_k^e can be split into two models, one for each value of Y^e, we can perform an invariance test on each model. If both are found to be invariant, we can consider $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$ as a whole to be invariant. Invariance tests on additive noise models have been widely studied: Various tests have been proposed for linear [9] and nonlinear [10] models. We adopt one such approximate test from [10] known as the *residual distribution test* for our setting, as further detailed in Algorithm 1.

Algorithm 1 Binary Invariant Residual Distribution Test

Input: Y^e and X^e, for each $e \in \mathcal{E}_{\text{train}}$, significance level α, and the pair (k, S)

Output: *accepted* or *rejected*

Regress $X_{k,Y=i}$ on $X_{S,Y=i}$ to get $\hat{g}_i$, for $i \in \{0, 1\}$
for each $e \in \mathcal{E}_{\text{train}}$ and $i \in \{0, 1\}$ **do**
  $R_i^e = X_{k,Y=i}^e - \hat{g}_i(X_{S,Y=i}^e)$
  $R_i^{-e} = X_{k,Y=i}^{-e} - \hat{g}_i(X_{S,Y=i}^{-e})$
  $\text{pval}_i^e = t\text{-test}(R_i^e, R_i^{-e})$
Combine p-values in pval_1^e and pval_0^e via Bonferroni correction
if $\min_{e \in \mathcal{E}_{\text{train}}} \text{pval}_1^e > \alpha$ **and** $\min_{e \in \mathcal{E}_{\text{train}}} \text{pval}_2^e > \alpha$ **then**
  return *accepted*
else return *rejected*

We use Algorithm 1 as an approximate test for whether $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$ is invariant over environments. We now employ this test to develop a practical method for estimating Y^{test} which we refer to as bIMP. We adopt a similar approach to that of [14] and [15] in which we test the invariance of $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$ for all possible pairs (k, S). We then train models using the X_k^e and X_S^e which are *accepted* according to Algorithm 1. Our bIMP models are a combination of two separate models trained to estimate both $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$ and $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S]$. Given both of these estimates, we compute an estimate of Y^{test} using (8). As it is likely that more than one pair is accepted, the final estimate of Y^{test} is the average estimate over all accepted pairs.

While we can guarantee invariance via the bIMP, there is no guarantee that the estimation will predict well on e^{test}. As such, in addition to filtering pairs based on invariance, bIMP also filters based on a prediction score. Invariant pairs $\mathcal{T}_{\text{inv}}$ computed using (8) are filtered using the mean squared prediction error. The threshold by which the pairs are filtered is identical to the procedure proposed in [14].

The bIMP method proposed gives freedom to the user to select the underlying models with which to estimate $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S]$ and $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$. In the case of $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S]$, we have complete freedom to select whichever model suits the data, be it linear or nonlinear. For $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$, we are restricted by the additive noise of (1). In addition, we have chosen to model X_k using two sub-models, one for each value of Y as in (11). This, however, is not the only option and depends on the invariance test used. When estimating each model, ordinary least squares (OLS) could be used for linear models, and a generalized additive model (GAM) or Gaussian

---

Algorithm 2 bIMP

Input: Y^e, for each $e \in \mathcal{E}_{\text{train}}$, and X^e, for each $e \in \mathcal{E}_{\text{obs}}$
Output: Estimate $\hat{Y}^{\text{test}}$

Identify the set of all invariant pairs $\mathcal{T}_{\text{inv}}$ using Algorithm 1
Filter pairs from $\mathcal{T}_{\text{inv}}$ based on prediction score
for each (k, S) in $\mathcal{T}_{\text{inv}}$ **do**
  Estimate $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S, Y]$ by regressing X_k on (X_S, Y)
  Estimate $\mathsf{E}_{\mathcal{P}_{\text{test}}}[X_k|X_S]$ by regressing X_k^{test} on X_S^{test}
  Using (8), compute $\hat{Y}_{k,S}^{\text{test}}$ for the pair (k, S)
$\hat{Y}^{\text{test}} = \frac{1}{|\mathcal{T}_{\text{inv}}|} \sum_{(k,S) \in \mathcal{T}_{\text{inv}}} \hat{Y}_{k,S}^{\text{test}}$

---

process regression could be used for nonlinear models. In practice, we found estimating each model using OLS to be the most efficient, as fitting two nonlinear models for all possible (k, S) pairs can be computationally expensive.

Remark 2. *There are several challenges with this approach that we leave for future work. We observe that nonlinear implementations of the invariance test (Algorithm 1) may lead to erroneously accepted invariant pairs. In addition to this, the complexity of training a nonlinear model for all possible (k, S) pairs can be high. Finally, the effects of model misspecification can be challenging to analyze.*

VI. Experiments

We provide one synthetic and two real datasets to test the effectiveness of bIMP and compare with the following two baselines: (1) a binary adaptation of Method II from [9] (ICP), and (2) logistic regression (LR). While we do not expect LR to perform well on unknown environments, it serves as a natural baseline. While ICP can handle the binary response setting via logistic regression, SR is specific to regression settings and thus not reported. In all experiments, we set $\alpha = 0.1$.
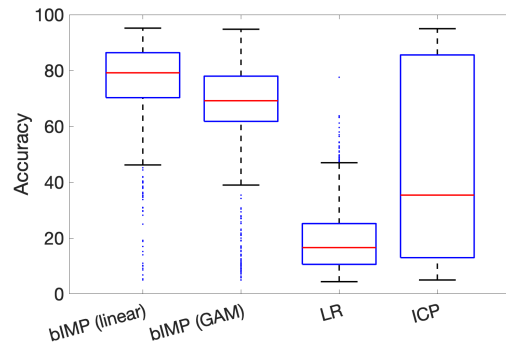

Fig. 2: Simulation accuracy over 1000 simulated datsets.

As there is some degree of freedom in selecting how the sub-models in bIMP are trained, we explore two variants of bIMP: bIMP (linear) and bIMP (GAM). For both variants, we follow the invariance test in Algorithm 1 and estimate g_1 and g_0 using OLS. We estimate $\mathsf{E}_{\mathcal{P}_e}[X_k|X_S]$ using OLS for bIMP (linear), and a GAM for bIMP (GAM).

Synthetic data. The simulated dataset is generated as follows. We generate data from three environments, $e^1, e^2 \in \mathcal{E}_{\text{train}}$, and

e^{test}. The number of predictors m is randomly selected from $\{3, \ldots, 7\}$. For each $i \in \{2, \ldots, m\}$ and $e \in \mathcal{E}_{\text{train}}$, $X_i^e \sim \mathcal{N}(\mu_i^e, 1)$, and μ_i^e is randomly selected on the interval $[-2, 0]$ for $e = e^1$, $[0, 2]$ for $e = e^2$, and $[0, 3]$ for $e = e^{\text{test}}$. Then, where $S_1 = \{2, \ldots, m\}$, $Y^e | X_{S_1}^e$ follows a logistic model such that $\mathcal{P}_e(Y = 1 | X_{S_1}) = 1/(1 + e^{-X_{S_1}\beta^e})$ for $e \in \mathcal{E}_{\text{train}}$. For e^{test}, $Y^{\text{test}} | X_{S_1}^{\text{test}}$ follows a probit model such that $Y^{\text{test}} = 1$, if $X_{S_1}^{\text{test}}\beta^{\text{test}} + \epsilon < 0$, where $\epsilon \sim \mathcal{N}(0,1)$. For all $e \in \mathcal{E}_{\text{obs}}$, randomly select β^e as $\beta^e \sim \text{Unif}[0,1]$. The coefficients are then scaled such that they sum to one. For all $e \in \mathcal{E}_{\text{obs}}$, the variable X_1^e is then simulated similarly to X_k^e in (11). Specifically, $g_1(X_{S_1}^e) = X_{S_1}^e \eta_1$ and $g_0(X_{S_1}^e) = X_{S_1}^e \eta_0$. The noise term associated with X_1^e is a standard normal. The coefficients $\eta_{1,i} \sim \text{Unif}[0,1]$ and $\eta_{0,i} \sim \text{Unif}[0,1]$ do not vary over the environment. The number of samples per environment is fixed to 1000.

Simulation results on both accuracy and mean squared error (MSE) indicate that bIMP can generalize to the test environment while LR and ICP are not (Fig 2). In addition, bIMP (linear) slightly outperforms bIMP (GAM). While we expect LR to behave poorly, ICP also performs poorly as all parents of Y are intervened in every simulation.

	bIMP (linear)	bIMP (GAM)	LR
Environment	Accuracy		
born in US	85.0	84.9	78.2
overtime	68.4	59.1	77.0
caucasian	85.0	85.2	78.1

TABLE I: *census*: performance and training environments.

Two real-world data. We also include experiments on two real datasets: *census* [18] and *mushroom* [19]. The census dataset is data gathered from the 1994 US census and contains 14 societal and demographic variables such as age, education, marital status, and working class. The target variable used is whether or not an individual's income exceeded 50k/yr. The data is first split into test and training data by whether or not a person graduated from a college. Thus, we train only on those who did not graduate college with the aim of extending our trained model to those who did. We further split the training data and run the methods on each set of training environments. The variables used to split the training data into environments are "was the person born in the US", "do they regularly work more than 40hr/week", and "does the person identify as Caucasian". The experiment shows that bIMP outperforms LR and ICP in all environments aside from the *overtime* environment (Table I). The ICP method returns no invariant predictors for any environment, thus no predictions can be made and no accuracy is reported; this is also the case for the *mushroom* data below.

	bIMP (linear)	bIMP (GAM)	LR
Environment	Accuracy		
meadows	76.0	87.5	46.2
paths	88.1	90.9	11.8

TABLE II: *mushroom*: performance and training environments.

The *mushroom* dataset contains 16 features related to naturally growing mushrooms' size, shape, and color and showcases how the proposed approach can handle discrete and categorical data. We aim to predict whether or not a mushroom is edible based on these factors. The environments on which we predict are the habitats in which the mushrooms grow. Specifically, we train on mushrooms that grow in grass or urban habitats and test on mushrooms that grow in meadows or paths. Results in Table II indicate that bIMP outperforms ICP and LR for both the linear and GAM variants, while the GAM variant performed the best.

VII. ACKNOWLEDGEMENTS

APPENDIX

Proof of Proposition 1. First, we show that (8) holds for any $e \in \mathcal{E}_{\text{obs}}$. Without loss of generality, let X_i^e be continuous for all $i \in \{1, \ldots, m\}$. The pdf of $X_k^e | X_S^e$ for any $e \in \mathcal{E}_{\text{obs}}$ is

$$f_{X_k^e | X_S^e}(x_k | x)$$
$$= f_{X_k^e | X_S^e, Y^e}(x_k | x, 1) \cdot p_{Y^e | X_S^e}(1|x)$$
$$\quad + f_{X_k^e | X_S^e, Y^e}(x_k | x, 0) \cdot p_{Y^e | X_S^e}(0|x)$$
$$= f_{X_k^e | X_S^e, Y^e}(x_k | x, 1) \cdot p_{Y^e | X_S^e}(1|x)$$
$$\quad + f_{X_k^e | X_S^e, Y^e}(x_k | x, 0) \cdot \left[1 - p_{Y^e | X_S^e}(1|x)\right]$$
$$= p_{Y^e | X_S^e}(1|x) \left[f_{X_k^e | X_S^e, Y^e}(x_k | x, 1) - f_{X_k^e | X_S^e, Y^e}(x_k | x, 0)\right]$$
$$\quad + f_{X_k^e | X_S^e, Y^e}(x_k | x, 0). \tag{12}$$

Then using (12), we can write $\mathsf{E}_{\mathcal{P}_e}[X_k | X_S = x]$ as

$$\int_{-\infty}^{\infty} x_k \cdot f_{X_k^e | X_S^e}(x_k | x) \, dx_k$$
$$= \mathsf{E}_{\mathcal{P}_e}[Y | X_S = x] \cdot \mathsf{E}_{\mathcal{P}_e}[X_k | X_S = x, Y = 1]$$
$$\quad - \mathsf{E}_{\mathcal{P}_e}[Y | X_S = x] \cdot \mathsf{E}_{\mathcal{P}_e}[X_k | X_S = x, Y = 0]$$
$$\quad + \mathsf{E}_{\mathcal{P}_e}[X_k | X_S = x, Y = 0]. \tag{13}$$

Thus, $\mathsf{E}_{\mathcal{P}_e}[Y | X_S]$ can be written as

$$\frac{\mathsf{E}_{\mathcal{P}_e}[X_k | X_S] - \mathsf{E}_{\mathcal{P}_e}[X_k | X_S, Y = 0]}{\mathsf{E}_{\mathcal{P}_e}[X_k | X_S, Y = 1] - \mathsf{E}_{\mathcal{P}_e}[X_k | X_S, Y = 0]}. \tag{14}$$

We now show (I) $\mathsf{E}_{\mathcal{P}_e}[X_k | X_S = x, Y = y]$ does not depend on e and (II) the denominator of (14) is non-zero. Since $X_S^e = (X_R^e, X_Q^e)$,

$$\mathsf{E}_{\mathcal{P}_e}[X_k | X_S, Y] = \mathsf{E}_{\mathcal{P}_e}[X_k | X_R, X_Q, Y] \stackrel{(a)}{=} \mathsf{E}_{\mathcal{P}_e}[X_k | X_R, Y]$$
$$\stackrel{(b)}{=} \mathsf{E}_{\mathcal{P}_e}[g(X_R, Y) + \epsilon | X_R, Y] = g(X_R^e, Y^e), \tag{15}$$

where (a) follows since $X_Q^e \perp\!\!\!\perp X_k^e | X_R^e, Y^e$, (b) follows from the assumption $X_k^e = g(X_R^e, Y^e) + \epsilon^e$, and (c) follows since ϵ has zero mean. Thus, the $\mathsf{E}_{\mathcal{P}_e}[X_k | X_S = (x_Q, x_R), Y = y]$ does not depend on e as $\mathsf{E}_{\mathcal{P}_e}[X_k | X_S = (x_Q, x_R), Y = y] = g(x_R, y)$. As the output of the function g is not constant with regards to any of its inputs as in (1), the denominator of (14) is non-zero. $\square$

REFERENCES

[1] N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann, "Methods for causal inference from gene perturbation experiments and validation," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7361–7368, 2016.

[2] A. V. Goddard, Y. Xiang, and C. J. Bryan, "Invariance-based causal prediction to identify the direct causes of suicidal behavior," *Frontiers in psychiatry*, p. 2598, 2022.

[3] T. Haavelmo, "The probability approach in econometrics," *Econometrica: Journal of the Econometric Society*, vol. 12, pp. 1–115, 1944.

[4] J. Aldrich, "Autonomy," *Oxford Economic Papers*, vol. 41, no. 1, pp. 15–34, 1989.

[5] K. D. Hoover, "The logic of causal inference: Econometrics and the conditional analysis of causation," *Economics & Philosophy*, vol. 6, no. 2, pp. 207–234, 1990.

[6] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," *arXiv preprint arXiv:1206.6471*, 2012.

[7] A. P. Dawid and V. Didelez, "Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview," *Statistics Surveys*, vol. 4, pp. 184–231, 2010.

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

[9] 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)*, pp. 947–1012, 2016.

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

[11] 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.

[12] 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.

[13] 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.

[14] N. Pfister, E. G. Williams, J. Peters, R. Aebersold, and P. Bühlmann, "Stabilizing variable selection and regression," *The Annals of Applied Statistics*, vol. 15, no. 3, pp. 1220–1246, 2021.

[15] K. Du and Y. Xiang, "Learning invariant representations under general interventions on the response," *IEEE Journal on Selected Areas in Information Theory*, 2023.

[16] ——, "Generalized invariant matching property via lasso," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[17] A. Goddard, Y. Xiang, and I. Soloveychik, "Error probability bounds for invariant causal prediction via multiple access channels," *Asilomar Conference on Signals, Systems, and Computers*, 2023.

[18] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[19] "Mushroom," UCI Machine Learning Repository, 1987, DOI: https://doi.org/10.24432/C5959T.