Learning Invariant Representations Under General Interventions on the Response

Kang Du and Yu Xiang, Member, IEEE

Abstract—It has become increasingly common nowadays to collect observations of feature and response pairs from different environments. As a consequence, one has to apply learned predictors to data with a different distribution due to distribution shifts. One principled approach is to adopt the structural causal models to describe training and test models, following the invariance principle which says that the conditional distribution of the response given its predictors remains the same across environments. However, this principle might be violated in practical settings when the response is intervened. A natural question is whether it is still possible to identify other forms of invariance to facilitate prediction in unseen environments. To shed light on this challenging scenario, we focus on linear structural causal models (SCMs) and introduce invariant matching property (IMP), an explicit relation to capture interventions through an additional feature, leading to an alternative form of invariance that enables a unified treatment of general interventions on the response as well as the predictors. We analyze the asymptotic generalization errors of our method under both the discrete and continuous environment settings, where the continuous case is handled by relating it to the semiparametric varying coefficient models. We present algorithms that show competitive performance compared to existing methods over various experimental settings including a COVID dataset.

Index Terms—Multi-environment domain adaptation, invariance, structural causal models, semiparametric varying coefficient model.

I. INTRODUCTION

HOW TO make reliable predictions in unseen environments that are different from training environments is a challenging problem, which is fundamentally different from the classical machine learning settings [1], [2], [3]. Modeling these distribution shifts in a principled way is of great importance in many fields including robotics, medical imaging, and environmental science. Apparently, this problem is ill-posed without any constraints on the relationship between training and test distributions, as the test distribution may be arbitrary. Consider the problem of predicting the response $Y$ given its predictors $X = (X_1, \ldots, X_d)^T$ in unseen environments. To model distribution changes across different environments (or training and test distributions), we follow the approach of using structural causal models (SCMs) [4], [5] to model different data-generating mechanisms. The common assumption is that the assignment for $Y$ does not change across environments (or $Y$ is not intervened), which allows for natural formulations of the invariant conditional distribution of $Y$ given a subset of $X$ [5], [6], [7], [8], [9], [10], [11], [12], [13]. The underlying principle is known as invariance, autonomy or modularity [4], [14], [15], [16].

Following this principle, the invariance-based causal prediction initiated in [8] (also see [17] and [11] and references therein) assumes that the conditional distribution of $Y$ given a set of predictors $X_S \subseteq \{X_1, \ldots, X_d\}$ is invariant in all environments, i.e., $P_e(Y|X_S) = P_h(Y|X_S)$ for environments $e$ and $h$, where $(X, Y)$ is generated according to the joint distribution $P_e \coloneqq P_{X, Y}^e$. Focusing on linear SCMs, it assumes the existence of a linear model that is invariant across environments, with an unknown noise distribution and arbitrary dependence among predictors (see extensions to nonlinear [18] and time series [19] settings). Following this framework, theoretical guarantees for domain adaptation have been developed in [9], [20]. More recently, a multi-environment robustification method for domain adaptation called the stabilized regression [21] explicitly enforces stability (based on a weaker version of invariance $E_{P_e}[Y|X_S = x_s] = E_{P_h}[Y|X_S = x_s]$) by introducing the stable blanket, which is a refined version of the Markov blanket to promote generalization. The tradeoff between predictive performance on training and test data has been studied via regularization under shift interventions [22]. Motivated by [8], the invariant risk minimization (IRM) [23] only uses data from the training environments (i.e., the out-of-distribution generalization setting), and imposes $P_e(Y|\phi(X)) = P_h(Y|\phi(X))$, where $\phi$ is invariant across environments, leading to a bi-leveled optimization problem that is not practical. Several relaxed versions of IRM have been proposed in [23], but they behave very differently from the original IRM (see, e.g., [24], [25]). For a framework of the out-of-distribution setting from a causal perspective with a focus on minimizing the worst-case risk, see [26] and references therein. In this line of invariance-based work, the fundamental assumption is that interventions on the target variable $Y$ are not allowed. In practical settings, however, the structural assignment of $Y$ might change across environments, namely, $Y$ might be intervened. How to relax this assumption in a principled way is one of the main motivations in our work. We propose to explore alternative forms of invariance and make an attempt in this direction by focusing on linear SCMs. Concretely, the assignment for $Y$ allows general interventions $Y^e = (\beta^e)^TX^e + \epsilon^e_Y$, where $Y$ can be intervened through coefficient $\beta^e$ and/or the noise $\epsilon^e_Y$, to capture the dependence of structural assignment
across different environments (preliminary results have been reported by the same authors in [27]). We consider a multi-environment regression setting for domain adaptation: There are multiple training data \((X_e, Y_e)\) for \(e \in \mathcal{E}^{\text{train}}\) that are generated from a training model and one test data (indexed by \(e^{\text{test}}\)) from a test model; we assume the training model and test model follow SCMs with the same but unknown graph structure, but we allow \(\beta_e\) and the mean and variance of \(e_Y\) to be arbitrarily different under the two models. To avoid the setting to be ill-posed, a key necessary condition is that \(Y^e\) needs to have at least one child in the SCMs, as prediction is not possible otherwise given that \(Y^e\) may change arbitrarily over environments. The main challenge lies in whether it is still possible to identify other forms of invariance to facilitate prediction in the test environment. We propose an alternative form of invariance \(\mathcal{P}_p(Y|\phi_e(X)) = \mathcal{P}_h(Y|\phi_h(X))\) that is enabled by a family of conditional invariant transforms \(\Phi \ni \phi_e, \phi_h\). Under general interventions on \(Y\), we provide explicit constructions of such transforms by developing invariant matching property (IMP), a deterministic relation between an estimator of \(Y\) and \(X\), along with an additional predictor constructed from \(X\). To enable a systematic way of constructing the IMP, we provide a natural decomposition of it and demonstrate this when only \(Y\) is intervened or both \(X\) and \(Y\) are intervened. The IMP comes with several appealing features: (1) it does not vary over environments, making it applicable in unseen environments, and (2) the identification of the IMP follows directly from the fact that the training data contains multiple environments. We study the asymptotic generalization error for both the discrete environment setting and continuous counterpart, which is the more challenging setting. Interestingly, we reveal a connection between the continuous environment setting with the semiparametric varying models, which makes the asymptotic generalization analysis possible. We believe that our results open up new possibilities for multi-environment regression methods for domain adaptation under the structure causal models.

The role of causality in facilitating domain adaptation problems is first articulated in [6], focusing on causal and anticausal predictions. Reweighting methods have been extensively studied for covariate shift [1], [35], [36], which assumes that only the feature distribution changes over environments while the conditionals remain the same. The label shift, which aligned with the anticausal setting, has attracted much attention recently [37], [38], [39]. Many other interesting domain adaptation methods have been developed but they are less related to this work. The performance bounds using Vapnik-Chervonenkis (VC) theory have been initiated in [40]. There are fundamental works from the robust statistics perspective including distributional robust learning [41], [42], [43], [44], [45], [46] and adversarial machine learning [47], [48].

### B. Contribution and Structure

There are four main contributions in this work. (I) We formulate a general invariance property and the corresponding conditional invariant transforms for analyzing general interventions on linear SCMs (Section II). (II) We tackle this problem by introducing the invariant matching property (IMP) and providing a systematic approach for establishing explicit characterization of it (Sections III and IV). (III) To handle the continuous environment setting, we bridge our framework with the profile likelihood estimators developed in the semiparametric literature, leading to asymptotic performance guarantees under this challenging setting (Section VI). (IV) Motivated by our theoretical results, we develop efficient algorithms that show competitive empirical performance over a variety of simulation settings including a COVID dataset (Sections V and VII). All the technical details are deferred to Appendix in the supplementary material.

### II. BACKGROUND AND PROBLEM FORMULATION

Consider a response \(Y \in \mathbb{R}\) and a vector of predictors \(X = (X_1, \ldots, X_d)^\top \in \mathcal{X} \subseteq \mathbb{R}^d\) following an acyclic linear SCM,

\[
\mathcal{M} : \begin{cases}
X = \gamma Y + BX + \varepsilon_X \\
Y = \beta^\top X + \varepsilon_Y,
\end{cases}
\]

where \(\beta, \gamma \in \mathbb{R}^d\), \(B \in \mathbb{R}^{d \times d}\), \(\varepsilon_X = (\varepsilon_{X_1}, \ldots, \varepsilon_{X_d})^\top\), and the noise variables \(\{\varepsilon_{X_1}, \ldots, \varepsilon_{X_d}\}\) and \(\varepsilon_Y\) are jointly independent. We use \(\mathcal{G}(\mathcal{M})\) to denote the directed acyclic graph induced by \(\mathcal{M}\), with edges determined by the non-zero coefficients in \(\mathcal{M}\). We denote the parents, children, descendants, and Markov blanket of a variable \(Z \in \{X_1, \ldots, X_d, Y\}\) as \(PA(Z), CH(Z), DE(Z),\) and \(MB(Z)\), respectively. When \((X, Y)\) is observed in different environments (e.g., different experiment settings for data collection), the parameters \((\beta, \gamma, B)\) and the distributions of \(\{\varepsilon_X, \varepsilon_Y\}\) may change. In the following, we use interventions on the SCM \(\mathcal{M}\) to model such changes.

Let \(\mathcal{E}^{\text{all}}\) denote the set of all possible environments,\(^1\) which are partitioned into multiple training environments \(\mathcal{E}^{\text{train}}\) and one test environment \(\{e^{\text{test}}\}\) such that \(\mathcal{E}^{\text{all}} = \mathcal{E}^{\text{train}} \cup \{e^{\text{test}}\}\). In each environment \(e \in \mathcal{E}^{\text{all}},\) \((X, Y)\) is generated according to

\(^1\) We use training (or test) environments and observable (or unseen) environments interchangeably.
the joint distribution $P_e := P^{X,Y}_e$, and to simplify notation, we write $P_{\text{test}} := P^{X,Y}_{\text{test}}$. A variable from $\{X_1, \ldots, X_d, Y\}$ is intervened if the parameters or noise distribution in its assignment changes over different $e \in E^{\text{all}}$. For instance, the changes of $\beta$ and/or the distribution of $\epsilon_Y$ correspond to the intervention on $Y$ (see different intervention cases below). Importantly, we allow both $Y$ and any subset of $\{X_1, \ldots, X_d\}$ to be intervened in $E^{\text{all}}$.

Now, we introduce the linear SCM $\mathcal{M}$ with parameters that change with environments. For each $e \in E^{\text{all}}$, the linear SCM $\mathcal{M}$ is modified to be

$$\mathcal{M}^e : \begin{align*}
X^e &= y^e Y^e + B^e X^e + \epsilon^e_X \\
Y^e &= (\beta^e)^\top X^e + \epsilon^e_Y.
\end{align*}$$

This formulation is fairly general. From the structural perspective, this consists of causal, anticausal, and mixed-causal-anticausal settings [6]. It should be noted that we only adopt the linear SCM rather than the fully specified SCMs as in [49], since learning the functional forms can be more complicated than the prediction problem we aim to solve. Regarding the intervention types, we discuss several special cases to put them into perspective.

1) **Shift interventions on $X$ or $Y$:** A variable $X_j$ is intervened through a shift if the mean of the noise variable $\epsilon^e_{X,j}$ changes with $e \in E^{\text{all}}$. For the shift intervention on $Y$, the mean of $\epsilon^e_Y$ changes.

2) **Interventions on the coefficients of $X$ or $Y$:** A variable $X_i$ is intervened through coefficients if the coefficients $\{\gamma^e_{i,j}, B^e_{i,j}\}$ change with $e \in E^{\text{all}}$. For $Y$, the change is on the coefficient vector $\beta^e$.

3) **Interventions on the noise variance of $X$ or $Y$:** Similar to shift interventions, a variable $X_j$ or $Y$ is intervened if its noise variance changes.

We observe $n^e$ i.i.d. samples $\{(x^1_1, y^1), \ldots, (x^e_m, y^e)\}$ from each training environment distribution $P_e$ for $e \in E^{\text{train}}$, but in the test environment $e^{\text{test}}$ we only observe $m$ i.i.d. samples $\{x^1_1, \ldots, x^m_1\}$ from $P^{X}_{\text{test}}$. The goal is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that works well on $e^{\text{test}}$ in the sense that it minimizes the test population loss

$$L_{\text{test}}(f) := E_{(X,Y) \sim P_{\text{test}}}[l(Y, f(X))],$$

where $l$ is the square loss function $l(y, \hat{y}) = (y - \hat{y})^2$. The optimal function is $f(x) = E_{P_{\text{test}}}[Y|X = x]$, which cannot be learned from the observed data in general when $X$ and/or $Y$ is intervened. Without any constraints on the relationship between $P_{\text{test}}$ and $P_e$ for $e \in E^{\text{train}}$, the test population loss can be arbitrarily large. To make this problem tractable, we assume that $(X, Y)$ under $P_{\text{test}}$ and $P_e$ are generated according to the SCMs described in (2) but we do not assume that the causal graph is known and we allow for general types of interventions.

It is well-known that if $Y$ is not intervened, a general form of invariance principle applies, assuming the existence of some subset $S \subseteq \{1, \ldots, d\}$ such that

$$P_e(Y|X_S) = P_h(Y|X_S)$$

holds for any $e, h \in E^{\text{all}}$. Under this assumption, the causal function $E_{P_e}[Y|X_{PA}(Y)]$ is invariant across different environments and minimax under the class of all possible interventions on $X$ [9]. If not every predictor is intervened arbitrarily, previous works that are motivated by (4) (as mentioned in Section I) aim to improve upon the causal function. The main challenge in our setting comes from the general interventions on $Y$, making the traditional invariance principle not applicable. Importantly, the causal function can change arbitrarily with environments. In this work, we propose to exploit an alternative form of invariance to tackle this problem.

**Definition 1:** A function $\phi : (E^{\text{all}}, \mathbb{R}^d) \rightarrow \mathbb{R}^q$ is called a conditional invariant transform if the following invariance property holds for any $e, h \in E^{\text{all}}$

$$P_e(Y|\phi_e(X)) = P_h(Y|\phi_h(X)).$$

Under general intervention settings, we denote this class of conditional invariant transforms as $\Phi$, and we provide explicit characterizations of it via the invariant matching property (IMP) (see Definition 2). For each $\phi \in \Phi$, the invariance property (5) enables us to compute

$$f_{\phi_e}(x) = g \circ \phi_e(x) = E_{P_e}[Y|\phi_e(x)],$$

for any $e \in E^{\text{all}}$, where the function $g : \mathbb{R}^q \rightarrow \mathbb{R}$ is invariant across environments and is nonlinear in general. Equivalently, this solves a relaxed version of (3) by minimizing $L_{\text{test}}(f_{\phi})$ over $\{\phi \in \Phi\}$. This formulation allows us to treat the general mixed-causal-anticausal problem under general interventions on $Y$ in a unified manner.

### III. INVARIANT MATCHING PROPERTY AND THEORETICAL GUARANTEES

Among the intervention settings below (2), the interventions on only $Y$ are important but rarely studied, which concerns the changes of $\beta^e$ and the distribution of $\epsilon^e_Y$. Our method can be motivated in this setting by the following observation: If $X$ includes any descendants of $Y$, then $\beta$ and $\epsilon_Y$ (that may change arbitrarily in the unseen environments) will be passed on to the descendants. Thus, the changes might be revealed by the changes of certain statistical properties of $X$, leading to our proposed invariant matching property detailed in this section. We start with a toy example.

**A. Motivating Example**

**Example 1:** Consider $(X^e, X^e_0), e \in E^{\text{toy}} = \{1, 2\}$, with $X^e := (X^e_1, X^e_2, X^e_3)^\top$ satisfying the following linear acyclic SCM (illustrated in Fig. 1),

$$M_{\text{toy}} : \begin{align*}
Y^e &= a^e X^e_1 + X^e_2 + N^e_Y \\
X^e_3 &= Y^e + X^e_1 + N^e_3,
\end{align*}$$

![Fig. 1. Directed acyclic graph $G(M_{\text{toy}})$.](Image 393x651 to 481x738)
where \( X_1^e, X_2^e, N_2^e \), and \( N_2^e \) are independent and \( \mathcal{N}(0,1) \)-distributed for every \( e \in \mathcal{E}^{\text{toy}} \). Since \((Y^e, X^e)\) is multivariate Gaussian, the MMSE estimator of \( Y^e \) given \( X^e \) is

\[
\mathbb{E}_{\mathcal{P}_e}[Y|X] = X^e \left( \mathbb{E}_{\mathcal{P}_e}[XX^\top] \right)^{-1} \mathbb{E}_{\mathcal{P}_e}[XY],
\]

which is not directly applicable for predicting \( Y^{\text{test}} \) as \( \alpha^e \) can change arbitrarily. Similarly, one can compute \( \mathbb{E}_{\mathcal{P}_e}[X_1|X_1, X_2] \) and \( \mathbb{E}_{\mathcal{P}_e}[X_1|X_1, X_2] \) can each be generated by a linear combination of \( \mathbb{E}_{\mathcal{P}_e}[Y|X_1, X_2] = \alpha^e X_1^e + X_2^e \) and \( \{X_1^e, X_2^e, X_3^e\} \), and we will highlight this observation by introducing two invariant relations (see Def. 4 and Def. 5). As a result, there exists a deterministic linear relation, which we refer to as matching,

\[
\mathbb{E}_{\mathcal{P}_e}[Y|X] = \lambda \mathbb{E}_{\mathcal{P}_e}[X_1|X_1, X_2] + \eta^T X^e,
\]

with coefficients \( \lambda = 1/2 \) and \( \eta = (-1, 0, 1/2)^\top \) that are invariant with respect to the environment (illustrated in Fig. 2(a)-(b)). For simplicity of notation, we denote \( X_3^e := \mathbb{E}_{\mathcal{P}_e}[X_2|X_1, X_3] \) in Fig. 2. Moreover, one can verify that \( \mathcal{P}_e(Y|X) \) is invariant since \( \mathbb{E}_{\mathcal{P}_e}[Y|X, X_1] \) and \( \text{Var}_{\mathcal{P}_e}(Y|X, X_4) \) are invariant. Thus \( \phi_e(X) = (X^T, X_4)^\top \) satisfies the invariance property (5). A prediction model in (8) with invariant relations is often not unique when it exists. One can show that

\[
\mathbb{E}_{\mathcal{P}_e}[Y|X] = -X_1^e + 1/2 X_2^e + X_3^e - 3/2 \lambda X_4^e,
\]

with \( X_3^e := \mathbb{E}_{\mathcal{P}_e}[X_2|X_1, X_3] \). However, invariant relations in (8) and (9) do not hold for \( X_5^e := \mathbb{E}_{\mathcal{P}_e}[X_1|X_1, X_2, X_3] \), since

\[
X_5^e = -\frac{\alpha^e + 1}{(\alpha^e + 1)^2} X_2^e + \frac{\alpha^e + 1}{(\alpha^e + 1)^2} X_4^e
\]

depends on \( e \) in a more complicated form so that there is no linear invariant relation between \((X_5^e, X^e)\) and \( \mathbb{E}_{\mathcal{P}_e}[Y|X] \), as illustrated in Fig. 2(c).

In Appendix A in the supplementary material, we extend Example 1 to allow for interventions on \( X_1, X_2, \) and \( Y \) through the means and/or variances of the noise variables.

Remark 1: In [50], the authors have shown that a form of varying filter connecting the features and response (as a special case of the varying coefficients captured by \( \beta^e \) in (2)) is effective for causal inference tasks, by adopting estimators from [51].

B. Invariant Matching Property

In this section, we generalize the invariant relations observed in Example 1 to a class of such relations for \( \mathcal{M}^e, e \in \mathcal{E}^{\text{all}} \) without assuming joint Gaussian distributions, and connect this with the invariance property in (5). For the identification of such relations, we show that even two training environments suffice (see Proposition 1).

To handle non-Gaussian cases (beyond Example 1), we choose to adopt the linear MMSE (or LMMSE) estimators for constructing linear invariant relations. For a target variable \( Y \in \mathbb{R} \) given a vector of predictors \( X \in \mathbb{R}^p \), the LMMSE estimator is defined as

\[
\mathbb{E}_l[Y|X] := \left( \theta^{\text{ols}} \right)^\top (X - \mathbb{E}[X]) + \mathbb{E}[Y],
\]

where \( \theta^{\text{ols}} := \text{Cov}(X, X)^{-1} \text{Cov}(X, Y) \) is called the population ordinary least squares (OLS) estimator. With a slight abuse of notation, we write \( \mathbb{E}_{l, \mathcal{P}_e}[Y|X] \) to denote the LMMSE of \( Y \) given \( X \) with respect to \( (X, \mathcal{P}_e) \). To simplify presentation, we focus on \((X^e, Y^e)\) with zero means for each \( e \in \mathcal{E}^{\text{all}} \) (or equivalently all the noise variables have zero means), while the non-zero mean settings can be handled by introducing the constant one as an additional predictor.

Definition 2: For \( k \in \{1, \ldots, d\} \), \( R \subseteq \{1, \ldots, d\} \backslash k \), and \( S \subseteq \{1, \ldots, d\} \), we say that the tuple \((k, R, S)\) satisfies the invariant matching property (IMP) if, for every \( e \in \mathcal{E}^{\text{all}}, \)

\[
\mathbb{E}_{l, \mathcal{P}_e}[Y|X_S] = \lambda \mathbb{E}_{l, \mathcal{P}_e}[X_k|X_R] + \eta^T X^e,
\]

for some \( \lambda \in \mathbb{R} \) and \( \eta \in \mathbb{R}^d \) that do not depend on \( e \). We denote \( \mathcal{I}_M := \{(k, R, S) : (10) \text{ holds}\} \) for model \( \mathcal{M} \), and we call \( (\eta^T, \lambda)^\top \) the matching parameters.

For a tuple \((k, R, S)\) such that the IMP does not hold, we simply call it a non-IMP. Observe that \( \mathbb{E}_{l, \mathcal{P}_e}[Y|X_S] \) is not directly applicable to the test environment due to its components depending on \( e \), but those components are fully captured by \( \mathbb{E}_{l, \mathcal{P}_e}[X_k|X_R] \). If the matching parameters are identified from the training environments, the IMP is applicable to the test environment since \( \mathbb{E}_{l, \mathcal{P}_e}[X_k|X_R] \) is completely determined by the distribution of \( X^e \) without the need of the target \( Y^{\text{test}} \). Since computing the additional feature \( \mathbb{E}_{l, \mathcal{P}_e}[X_k|X_R] \) is simply the prediction of \( X_k \) (as if \( X_k \) is not observed), the IMP indicates that the prediction of \( Y \) can benefit from the predictions of certain predictors. We formally define this class of additional features as follows.

Definition 3: For any \( k \in \{1, \ldots, d\} \) and \( R \subseteq \{1, \ldots, d\} \backslash k \), we call \( \mathbb{E}_{l, \mathcal{P}_e}[X_k|X_R] \) a prediction module. If a prediction module satisfies an IMP for some \( S \subseteq \{1, \ldots, d\} \), we call it a matched prediction module for \( S \). Now we discuss the relationship between the IMP and the invariance property \( \mathbb{P}_e(Y|\phi_e(X)) = \mathbb{P}_h(Y|\phi_h(X)) \) in (5), we rewrite (10) in a compact form as

\[
\mathbb{E}_{l, \mathcal{P}_e}[Y|X_S] = \theta^\top \tilde{X}^e,
\]

where

\[
\tilde{X}^e := (X_1^e, \ldots, X_d^e, \mathbb{E}_{l, \mathcal{P}_e}[X_k|X_R])^\top,
\]

and \( \theta = (\eta^T, \lambda)^\top \) denotes the matching parameter. Define

\[
\phi_e^{(k, R, S)}(X^e) := (X_5^e, \mathbb{E}_{l, \mathcal{P}_e}[X_k|X_R])^\top,
\]

where \( X_5^e \) is a row vector for some \( S' \subseteq \{1, \ldots, d\} \). For notational convenience, we introduce the shorthand \( \phi_e(X^e) := \phi_e^{(k, R, S)}(X^e) \). Note that \( \phi_e(X^e) \) is a linear transform of \( X^e \).

In general, however, the invariance of the matching parameter \( \theta \) does not imply the invariance property (5). In Section IV,
we will characterize a class of IMPs that each satisfies (5). When the invariance property holds, one can apply the general conditional expectation $f_{\theta_e}(x) = E_{P_e}[Y|\phi_e(x)]$ as in (6), since the linear estimator from the IMP is in general sub-optimal for the non-Gaussian cases. We will focus on linear estimators in this work as the extension can be handled via nonlinear regression methods in a straightforward manner.

It is noteworthy that since $E_{l,P_e}[X_k|X_R]$ is a linear function of $X_R$, the matching parameter $\theta$ is not unique given a single environment $e \in \mathcal{E}^{\text{train}}$. This causes issues when one aims to identify the possible IMPs given the distribution of $(X^e, Y^e)$. However, we show that two environments in the training data are sufficient to identify $\theta$ under a mild assumption on the matched prediction module.

**Proposition 1:** For a tuple $(k, R, S)$ that satisfies an IMP, the matching parameter $\theta$ can be uniquely identified in $\mathcal{E}^{\text{train}}$ if $|\mathcal{E}^{\text{train}}| \geq 2$ and

$$E_{l,P_e}[X_k|X_R = e] \neq E_{l,P_e}[X_k|X_R = h]$$

for some $e, h \in \mathcal{E}^{\text{train}}$ and $x \in \mathcal{X}_R$.

This proposition shows how the heterogeneity of the data generating process can be helpful for identifying important invariant relations.

**C. A Decomposition of the IMP**

In our toy examples, recall that the IMPs are derived by first computing $E_{P_e}[Y|S]$ and $E_{P_e}[X_k|X_R]$ separately and then fitting a linear relation from $(E_{P_e}[X_k|X_R], S)$ to $E_{P_e}[Y|S]$. These two steps reveal a natural decomposition of the IMP, which we term as the first and second matching properties below.

**Definition 4:** We say that $S \subseteq \{1, \ldots, d\}$ satisfies the first matching property if, for every $e \in \mathcal{E}^{\text{all}}$,

$$E_{l,P_e}[Y|S] = \lambda S E_{l,P_e}[Y|PA(Y)] + \eta Y^e,$$

for some $\lambda S \in \mathbb{R}$ and $\eta Y \in \mathbb{R}^d$ that do not depend on $e$.

First, observe that the first matching property holds for $S = PA(Y)$ since

$$E_{l,P_e}[Y|PA(Y)] = E_{l,P_e}[Y|PA(Y)] = (\beta^e)^\top X^e.$$

The first matching property concerns the set $S$ such that the components in $E_{l,P_e}[Y|S]$ that depend on $e$ are fully captured by the causal function $E_{P_e}[Y|PA(Y)]$. However, this invariant relation is not directly useful for the prediction of $Y^{\text{true}}$, since the causal function can change arbitrarily with $e$. To this end, we identify another invariant relation from $M^e$ which is called the second matching property.

**Definition 5:** For $k \in \{1, \ldots, d\}$ and $R \subseteq \{1, \ldots, d\} \setminus k$, we say that a tuple $(k, R)$ satisfies the second matching property if, for every $e \in \mathcal{E}^{\text{all}}$,

$$E_{l,P_e}[X_k|R] = \lambda X E_{l,P_e}[Y|PA(Y)] + \eta_{X} X^e,$$

for some $\lambda X \in \mathbb{R}$ and $\eta_{X} \in \mathbb{R}^d$ that do not depend on $e$.

It is straightforward to see that, if $\lambda X \neq 0$ in the second matching property, the first and second matching properties imply the IMP as follows,

$$E_{l,P_e}[Y|S] = \frac{\lambda Y}{\lambda X} E_{l,P_e}[X_k|R] + \left(\eta Y - \frac{\lambda Y}{\lambda X}\right)^\top X^e \quad := \lambda E_{l,P_e}[X_k|R] + \eta^e X^e.$$

For prediction tasks under SCMs, the causal function often plays a central role. Our first and second matching properties show how the LMMSE estimator $E_{l,P_e}[Y|S]$ and the matched prediction module $E_{l,P_e}[X_k|R]$ are connected with the causal function, respectively. Together, the two individual connections make up the IMP (illustrated in Fig. 3).
IV. CHARACTERIZATION OF INVARIANT MATCHING PROPERTIES

A. Interventions on the Response

First, we consider model $\mathcal{M}^e$ with interventions only on $Y$ through the coefficients, i.e.,

$$
\mathcal{M}^{e,1} : \begin{cases} 
X^e = \gamma Y^e + BX^e + \varepsilon_X^e \\
Y^e = (\alpha^e + \beta) X^e + \varepsilon_Y^e.
\end{cases}
$$

To distinguish the parents of $Y$ with varying and invariant coefficients, we decompose $\beta^e$ in $\mathcal{M}^e$ into two parts $\alpha^e$ and $\beta$. Without loss of generality, we assume that $\alpha^e_f \neq 0$ if and only if $\alpha^e$ is a non-constant function of $e$, and we define the following subset of parents of $Y$,

$$
PE = \{ j \in [1, \ldots, d] : \alpha^e_f \neq 0 \}.
$$

Remark 2: Note that, in the non-zero mean settings, this model covers the shift intervention on $Y$ through the varying coefficient of the predictor which is a constant one.

Recall that prediction modules do not rely on the response $Y$ but on the relations between the predictors for each environment. When $Y$ is unobserved (or equivalently, substituting $Y$ in (18) into (17)), the relations between the predictors are as follows,

$$
X^e = (\gamma (\alpha^e + \beta)^\top + B) X^e + \gamma \varepsilon_Y^e + \varepsilon_X^e,
$$

where $\gamma \varepsilon_Y + \varepsilon_X$ a vector of dependent random variables when $\gamma$ is not a zero vector. If $\alpha^e$ vanishes from (19), the distribution of $X^e$ becomes invariant with respect to environments. As a consequence, the condition (14) in Proposition 1 will not be satisfied. Observe that $\alpha^e$ is non-vanishing in (19) only if $\gamma$ is not a zero vector, which brings up the following key assumption.

Assumption 1: When $Y$ is intervened, we assume that $Y$ has at least one child.

Note that if $Y$ has no children, by the Markov property of SCMs [4], the test data sampled from $\mathcal{T}^Y_{test}$ provides no information about $\mathcal{T}^Y_{test}$ and thus the observed $X_{test}^e$ may correspond to two arbitrarily different $Y_{test}^e$’s, which makes the problem ill-posed.

The first and second matching properties enable us to characterize the tuples $(k, R, S)$’s that satisfy IMFs through the characterizations of $S$ (for the first matching property) and $(k, R)$ (for the second matching property) separately. In the following theorem, we show that a class of IMFs implied by the first and second invariant matching properties satisfy the invariance property (5).

Theorem 1: For model $\mathcal{M}^{e,1}$, the first and second matching properties hold in the following cases.

1) On the first MP: For each $S \subseteq \{1, \ldots, d\}$ such that $PE \subseteq S$, the first matching property holds.

2) On the second MP: For each $k \in \{1, \ldots, d\} \setminus PE$ and $R \subseteq \{1, \ldots, d\} \setminus k$ such that $PE \subseteq R$, the second matching property holds.

For any tuple $(k, R, S)$ above such that $R \subseteq S$, if $\lambda_X \neq 0$ in the second matching property, then $\phi_e(X^e) = (X^e_k, \mathcal{E}_{X^e}(X^e_k, X^e_R))^\top$ satisfies (5). Furthermore, $\mathcal{L}_{test}(f_0)$ is minimized by any $\phi$ with $S = \{1, \ldots, d\}$.

As a concrete example, recall (8) from our motivating example, where $R \subseteq S$ and $\lambda_X \neq 0$ are satisfied, $\mathbb{E}_{\mathcal{P}_E[Y|X]}$ can be represented using invariant coefficients. It is noteworthy that Assumption 1 is a necessary condition for $\lambda_X \neq 0$, and we provide a sufficient condition for $\lambda_X \neq 0$ in a concrete setting with $S = \{1, \ldots, d\}$ below.

Corollary 1: For model $\mathcal{M}^{e,1}$, the first and second matching properties hold in the following cases.

1) On the first MP: The first matching property holds for $S = \{1, \ldots, d\}$.

2) On the second MP: For each $k \in \{j \in MB(Y) : \alpha^e_f = 0\}$ and $R = -k := \{1, \ldots, d\}\setminus k$, the second matching property holds.

Proposition 2: Under Assumption 1, if $\alpha^e_k \neq 0$ in the second matching property if $B_{-k, k}$ is not in the following hyperplane,

$$
w^\top x + b = 0,
$$

where $w \in \mathbb{R}^{d-1}$ and $b \in \mathbb{R}$ are determined by the parameters in $\mathcal{M}^{e,1}$ other than $B_{-k, k}$.

The explicit expressions of $w$ and $b$ using the parameters in $\mathcal{M}^{e,1}$ are provided in the proof of Proposition 2. For generic choices of the parameters, the second matching property holds with $\lambda_X \neq 0$ since $B_{-k, k}$ is not necessarily on the hyperplane described in (20).

When $Y$ is additionally intervened through the noise variance, the proof of Theorem 1 will break down in general (see Remark 8 in Appendix D in the supplementary material). However, recall that the first matching property holds for $S = \mathcal{P}_{\alpha}(Y)$ by definition. In this case, we provide an example for the second matching property in the following corollary.

Corollary 2: Under Assumption 1, if $Y$ is intervened through the noise variance in model $\mathcal{M}^{e,1}$, the second matching property holds for $k \in CH(Y)$ such that $k \notin DE(i)$ for any $i \in CH(Y) \setminus k$, and $R = \{1, \ldots, d\} \setminus DE(Y)$.

The resulting IMFs no longer satisfy the invariance property (5), but we can use the IMP directly for the prediction of $Y_{test}^e$.

Remark 3: To sum up, the class of IMFs constructed under interventions on $Y$ only through the coefficients and shifts will in general imply the invariance property (5), but it is not the case under interventions on $Y$ through the noise variance. For the characterizations of IMP, we focus on sufficient conditions that are relatively simple to evaluate, while they are not necessary conditions in general. One way to find necessary and sufficient conditions for the first/second IMP is through explicit (but tedious) calculations as in the proof of Proposition 2. However, such conditions are hard to evaluate and provide limited insights into our methodology. We thus do not pursue them in this work. When $Y$ is intervened through the coefficients of all its parents, i.e., $PE = \mathcal{P}_{\alpha}(Y)$, a recent work [52] shows that $\mathcal{P}_{\alpha}(Y) \subseteq S$ and $\mathcal{P}_{\alpha}(Y) \subseteq R$ are also necessary for the IMP under mild assumptions, leading to a novel algorithm for identifying direct causes of $Y$ in the multi-environment setting.

B. Interventions on Both Predictors and Response

To generalize the setting when only $Y$ is intervened to the general setting when $X$ and $Y$ are both intervened, an idea is
to merge the setting when only Y is intervened with the one when only X is intervened. The latter setting has been studied in the stabilized regression framework [21]. The following set of predictors is identified (see Definition 3.4 therein),

\[ X^{\text{int}}(Y) = CH(Y) \]

\[ \cup \{j \in \{1, \ldots, d\} \mid \exists i \in CH(Y) \text{ such that } j \in DE(X_i)\}, \]

which contains the intervened children of Y (denoted by \( CH(Y) \)) and the descendants of such children. This useful notion can be defined for each \( X_j \in \{X_1, \ldots, X_d\} \), denoted by \( X^{\text{int}}(X_j) \) for each \( X_j \). When only X is intervened, the invariance principle (4) holds for \( S^* = \{1, \ldots, d\} \setminus X^{\text{int}}(Y) \); the Markov blanket of Y defined with respect to \( S_X \) is called the stable blanket of \( Y \) in [21]. In other words, by excluding the predictors in \( X^{\text{int}}(Y) \), the target \( Y \) is blocked from the interventions on \( X \) when conditioning on \( S_X \). This holds when \( Y \) is additionally intervened, as if only \( Y \) is intervened given \( S_X \). In order for \( S^* \) to include as least one child of \( Y \) as in Assumption 1, we need the following assumption.

Assumption 2: When \( Y \) is intervened, we assume that \( Y \) has at least one child that is not intervened and that child is not a descendant of some intervened child of \( Y \).

Based on the observation above, we identify an important class of IMPs for the general setting in the following Theorem.

Theorem 2: For the training model \( \mathcal{M}^e \) without the intervention on the noise variance of \( Y \), the first and second matching properties hold in the following cases.

1) On the first MP: For \( S \subseteq \{1, \ldots, d\} \setminus X^{\text{int}}(Y) \) such that \( PA(Y) \subseteq S \), the first matching property holds.

2) On the second MP: For each \( k \in \{1, \ldots, d\} \setminus \{PE \cup X^{\text{int}}(Y)\} \), and \( R \subseteq \{1, \ldots, d\} \setminus \{k, X^{\text{int}}(X_k) \cup X^{\text{int}}(Y)\} \) such that \( PA(X_k) \cup R \subseteq S \), the second matching property holds.

Furthermore, if \( \lambda_X \neq 0 \) in the second matching property, then \( \phi_e(X^\tau) = (X^\tau, E_{i \in \mathcal{R}}[X_i|X_R])^\top \) satisfies (5).

Remark 4: In general, there exist subsets of \( S \) and \( R \) from Theorem 2 such that the first and second matching properties hold, respectively. In particular, \( PA(Y) \subseteq S \) and \( PA(X_k) \subseteq R \) are not necessarily satisfied. Due to the Markov property of SCMs, \( Y \) is independent of its ancestors when conditioning on \( PA(Y) \), thus we focus on the IMPs that are more predictive by including \( PA(Y) \) in \( S \).

When \( Y \) is intervened through its noise variance, recall again that the first matching property always holds for \( S = PA(Y) \) by definition. The following proposition follows from Remark 9 in Appendix F in the supplementary material.

Proposition 3: When \( Y \) is additionally intervened through the noise variance in Theorem 2, the second matching property in Theorem 2 still holds.

Similar to the argument in the proof of Theorem 1, the class of \( \phi \)'s from Theorem 2 will lead to the same test population loss, as they depend on the same \( S \) that is fixed in this setting. Assumption 2 is necessary for \( \lambda_X \neq 0 \), while sufficient conditions for \( \lambda_X \neq 0 \) can be found similarly as in Proposition 2.

V. ALGORITHMS

For each \( e \in \mathcal{E}^{\text{train}} \), we are given the i.i.d. training data \( X_e \in \mathbb{R}^{n \times d} \), \( Y_e \in \mathbb{R}^n \), and we observe the i.i.d. test data \( X^\tau \in \mathbb{R}^{m \times d} \) and aim to predict \( Y^\tau \in \mathbb{R}^m \). Let \( X \in \mathbb{R}^{n \times d} \)

Algorithm 1: Invariant Prediction Using the IMP (Discrete)

Procedure IDENTIFY IMPs FROM THE TRAINING DATA

for \( k \in \{1, \ldots, d\}, S \subseteq \{1, \ldots, d\}, R \subseteq S \setminus k \) do

Compute the IMP score \( s_{\text{imp}} \) and the prediction score \( s_{\text{pred}} \) for \( i = (k, R, S) \)

Regress \( Y \) on \( [X_S, \hat{L}_2] \) to obtain \( f_i \)

Identify \( \hat{T} \) and \( \hat{T}_{\text{pred}} \)

Procedure PREDICTION ON THE TESTING DATA

\[ \hat{Y}^\tau = \frac{1}{|\mathcal{T}_{\text{pred}}|} \sum_{i \in \mathcal{T}_{\text{pred}}} f_i(X^\tau, \hat{L}_2) \]

with \( n = \sum_{e \in \mathcal{E}^{\text{train}}} n_e \) denote the pooled data of \( X_e \)'s. In this section, we present the implementation of our method starting with the case when \( e \) is sampled from a discrete distribution with a finite support. In this setting, we expect to have \( n_e \gg 1 \) for every \( e \in \mathcal{E}^{\text{train}} \) in general, thus it is possible to do estimation based on the data from each single environment. The challenging setting of continuous environments will be handled afterward.

A. Discrete Environments

To implement our method, the main task is to identify the set of IMPs \( \mathcal{I}_{\mathcal{M}} \) from the training data. For each tuple \( (k, R, S) \) in Algorithm 1, we test the following null hypothesis

\[ \mathcal{H}_0 : \text{There exists } \theta \in \mathbb{R}^{d+1} \text{ such that (11) holds.} \]

We propose two test procedures.

1) Test of the Deterministic Relation: Since the IMPs are linear and deterministic (i.e., noiseless), we test whether the residual vector \( R \in \mathbb{R}^n \) of fitting an IMP on \( (k, R, S) \) is a zero vector or not using the test statistics,

\[ T = \frac{1}{n} R^\top R, \]

where \( R \) is a pooled data vector of \( R_e \)'s defined below. To fit an IMP, we first estimate the two LMMSE estimators in (10) using OLS for each environment,

\[ \hat{L}_{e,1} := \left( X^\tau_{e,S} X_{e,S} \right)^{-1} X^\tau_{e,S} Y_e, \]

\[ \hat{L}_{e,2} := \left( X^\tau_{e,R} X_{e,R} \right)^{-1} X^\tau_{e,R} X_{e,k}. \]

Let \( \hat{L}_1 \in \mathbb{R}^d \) and \( \hat{L}_2 \in \mathbb{R}^d \) denote the pooled data of \( \hat{L}_{e,1} \)'s and \( \hat{L}_{e,2} \)'s, respectively. It is noteworthy that \( \hat{L}_2 \) only depends on \( X \), thus \( \hat{L}_2 \) for the test data can be computed similarly using \( X^\tau \). Next, we estimate the matching parameter using OLS on the pooled data (recall that the matching parameter cannot be identified using the data from a single environment). The OLS estimator of the matching parameter is

\[ \hat{\theta} := \left( \hat{\eta}^\top, \hat{\lambda} \right) = \left( \left[ X_S, \hat{L}_2 \right]^\top \left[ X_S, \hat{L}_2 \right] \right)^{-1} \left[ X_S, \hat{L}_2 \right]^\top \hat{L}_1. \]

For each \( e \in \mathcal{E}^{\text{train}} \), we obtain the residual vector of fitting an IMP

\[ R_e = \hat{L}_{e,1} - \hat{\lambda} \hat{L}_{e,2} - X_{e,S} \hat{\eta}. \]

2) Approximate Test of Invariant Residual Distributions: According to the invariance property (6), we test whether
the residual when regressing $Y$ on $[X_3, L_2]$ has constant mean and variance. Specifically, we use the t-test and F-test with corrections for multiple hypothesis testing from [8] (see Section II-A Method II). The test yields a p-value.

The test statistic from the first procedure and the p-value from the second procedure quantify how likely an IMP holds (i.e., the smaller the more likely), and thus we will refer to either one of them as an IMP score denoted by $s_{\text{IMP}}$. Let $\hat{I} = \{(k, R, S) : s_{\text{IMP}}(k, R, S) < c_{\text{IMP}}\}$ denote the set of IMPs identified from the training data, where $c_{\text{IMP}}$ is some cutoff parameter. Then, since IMPs are not equally predictive in general, we focus on the most predictive ones by introducing the mean squared prediction error as a prediction score $s_{\text{pred}}$, and we select the set of IMPs that are more predictive $\hat{I}_{\text{pred}} = \{(k, R, S) \in \hat{I} : s_{\text{pred}}(k, R, S) < c_{\text{pred}}\}$ with some cutoff parameter $c_{\text{pred}}$. For the second IMP score that is a p-value, the cutoff parameter $c_{\text{IMP}}$ is simply a significance level that is fixed to 0.05 in this work. For choosing the rest of the cutoff parameters, we follow a bootstrap procedure from [19] with one subtle difference: We sample the same amount of bootstrap samples from each environment rather than sampling over the pool data as in [19] since our procedure involves estimations using the data from each environment.

In practice, there can be spurious IMPs that have extremely small IMP scores but have large prediction scores, e.g., when $Y$ is independent of $X_3$ and $X_k$ is independent of $X_R$. To this end, we will pre-select $(k, R, S)$’s with prediction scores smaller than the median of all the computed prediction scores before identifying $\hat{I}$.

If the regression function $f_i$ in Algorithm 1 is chosen to be linear, one can use the IMP directly, i.e.,

$$\hat{Y}_{\text{IMP}}(k, R, S) = [X_S, \hat{L}_2] \hat{\theta},$$

which we call the discrete IMP estimator denoted by IMP. To make use of all the IMPs selected in $\hat{I}_{\text{pred}}$, we use an averaging step for the prediction of $Y$ in Algorithm 1.

Now we discuss the computation complexity of our method. For a given graph of size $d \geq 3$, the number of $(k, R, S)$’s with nonempty $(R, S)$’s is given by

$$\sum_{j=1}^{d} \binom{d}{j} (k - 1)^{d-j} (d-j) (d! - 1) = d! (2 \cdot 3^{d-1} - 2^d),$$

which follows by first choosing a set $S$ with $j$ elements, and then considering two settings $k \in S$ and $k \notin S$ since $R$ has to satisfy $R \subseteq S \setminus k$. This calculation implies that the exhaustive search step in our method is not applicable for relatively large graphs (e.g., $d \geq 15$) due to the exponential time complexity. It is noteworthy that in the ICP framework [8], a similar issue occurs due to an exhaustive search over $S \subseteq \{1, \ldots, d\}$. We discuss several options to alleviate this issue. (1) One can adopt a preprocessing step for feature selection to reduce the dimension $d$, using Lasso [53] or Boosting [54], as proposed in ICP [8]. (2) When prior information about the graph structure (e.g., the maximum number of parents of the nodes) is available or the graph structure can be estimated, the search space can be reduced. For instance, one can first adopt existing causal discovery methods (e.g., [55]) for graph structure estimation and then apply our methods according to the sufficient conditions in Theorems 1 and 2. (3) One can also exploit some intrinsic sparsity regarding IMPs, observe that there is only one matched prediction module (recall Definition 3) in the IMP, while the number of prediction modules grows as $d \cdot 3^d$. This sparsity has been studied when only $Y$ is intervened [56], leveraging a variant of Lasso.

### B. Continuous Environments

To model continuous environments, we introduce an environmental variable $U$ that is a continuous random variable with support $\mathcal{U}$. Apparently, this is a much more challenging setting compared with the discrete environment case, as we only have one training data sample for each $u \in \mathcal{U}$, making the OLS a poor estimate of $M_{\text{pred}}(Y|X_S)$. Fortunately, it turns out that we can leverage the semi-parametric varying coefficient (SVC) models [57] (see Appendix H in the supplementary material) to remedy this issue. In particular, we estimate $\mathbb{E}_{I_{\text{pred}} \mid u}[Y|X_S]$ by fitting,

$$Y = M + \beta^T Z + N \quad \text{with} \quad M = \alpha^T (U) W,$$

where $N$ is independent of $U$ and the two vectors of predictors $W \in \mathbb{R}^r$ (for the varying coefficient) and $Z \in \mathbb{R}^q$ (for the invariant coefficients) with $p + q = |S|$. Since we assume $N \perp U$, we focus on the settings when $Y$ is not intervened through the noise variance.

**Remark 5:** Our estimation procedure for the discrete environments can also be formulated under the SVC model with a discrete random variable $U$, where we treat all the coefficients as varying coefficients (i.e., $\beta = 0$), and IMP becomes an estimate of $M$.

An SVC model over $(Y^\tau, W^\tau, Z^\tau, N^\tau)$ for the test data can be defined similarly, where $\sigma^2 = \mathbb{E}[(N^\tau)^2]$ is the population generalization error of the IMP estimator. Observe that the linear SCM $M_u$ in (2) can be viewed as a collection of SVC models parameterized by $U = u$. Thus the estimation tasks for the linear SCMs from continuous environments can greatly benefit from the existing theories developed for SVC models. More precisely, we employ the following estimate

$$\mathbb{E}_{I_{\text{pred}} \mid u}[Y|X_S] = \hat{M}_{\mid U = u} + \hat{\beta}^T Z,$$

where the profile least-squares estimation of $\beta$ and $M$ proposed in [57] can be found in Appendix H in the supplementary material. Similarly, $\mathbb{E}_{I_{\text{pred}} \mid u}[Y|X_R]$ can be estimated by fitting another semi-parametric varying coefficient model

$$V = M_V + \beta_V^T Z_V + N_V \quad \text{with} \quad M_V = \alpha_V^T (U) W,$$

where $V$ denotes any $X_k$, and $X_R$ is divided into $Z_V \in \mathbb{R}^r$ and $W$.

It is noteworthy that the two SVC models share the same set of predictors with varying coefficients, which we explain below. A challenge for fitting such models is that the vector of predictors with varying coefficients, namely $W$, needs to be known. For continuous environments, we focus on discovering...
The following corollary considers the setting when the amount of unlabeled training and test data grows in a higher order than that of labels in the training data. The generalization error due to the estimation on the test data disappears.

**Corollary 3:** Given i.i.d. training data of size \( n \) with \( 0 < l_u < n \) labels and test data of size \( m \), if \( \max (\frac{m}{n}, \frac{1}{n}) \rightarrow 0 \) as \( \min (m, n) \rightarrow \infty \), under the technical assumptions in Appendix I in the supplementary material, the asymptotic generalization error of the \( \text{IMP}_c \) estimator is given by

\[
\frac{1}{m} \sum_{i=1}^{m} (\hat{Y}_i - Y_i)^2 = \sigma^2 + O_p(c_n \vee n^{-1/2}).
\]

The setting of discrete environments can be viewed as a special case of continuous environments, where the error term \( c_n \vee n^{-1/2} \) due to the kernel estimation procedure is replaced by an error term from multiple OLS estimations.

**Corollary 4:** For any \( (k, R, S) \in \mathcal{I}_M \), under the technical assumptions in Appendix I in the supplementary material, the asymptotic generalization error of the \( \text{IMP}_d \) estimator is given by

\[
\frac{1}{m} \sum_{i=1}^{m} (\hat{Y}_i - Y_i)^2 = \sigma^2 + O_p(a_n) + O_p(a_m),
\]

where \( a_n = (\min_{\gamma \in \gamma_{\text{num}} n_m})^{-1/2} \) and \( a_m = m^{-1/2} \).

This asymptotic generalization error heavily depends on the environment with the smallest sample size, which also supports the fact that \( \text{IMP}_d \) should not be employed for continuous environment settings.

**VII. Experiments**

The prediction performance is measured by the mean residual sum of squares (RSS) on the test environments. We compare our method with several baseline methods: Ordinary Least Squares (OLS), stabilized regression (SR) [21], anchor regression (AR) [22]. We have also compared with domain invariant projection (DIP) [29], conditional invariance penalty (CIP) [34], conditional invariant residual matching (CIRM) [28], and invariant risk minimization (IRM) [23]; it turns out that the empirical performance of these methods is not as competitive as the other baselines in our experimental settings, thus we do not report them below.

The two IMP scores lead to two versions of our algorithm, and we refer to the first one as IMP and the second one as \( \text{IMP}_{\text{inv}} \) (since it tests the invariance of the noise mean and variance). We focus on linear functions \( f_i \)'s in Algorithm 1, namely, we use the IMP estimators. For the profile likelihood estimation, we adopt the Epanechnikov kernel \( k(u) = 0.75 \max (1 - u^2, 0) \) with the bandwidth fixed to be 0.1. We test DIP, CIP, CIRM, and their variants provided in [28] with the default parameters. For the anchor regression, we use a 5-fold cross-validation procedure to select the hyper-parameter \( \gamma \) from \( \{0, 0.05, 0.1, \ldots, 0.5\} \). The significance levels are fixed to 0.05 for all methods. We randomly simulate 500 data sets for each experiment, if not mentioned otherwise.
steps of IMP, IMP inv, and SR, they have smaller variances through shifts sampled from Unif$[-2,2]$. For each test environment, the perturbations are sampled from Unif$[-10,10]$. The shift intervention on $X$ is the same as the shift interventions on $X$ in Section VII-A1. In this setting, since none of the baseline methods allow interventions on $X$ through the coefficients, they cannot even improve upon OLS. In Fig. 5, IMP performs slightly better than IMP inv, which may be due to the fact that the IMP method aims to find all possible IMPs, but the IMP inv only looks for IMPs that imply invariance. To further examine our test procedures for identifying IMPs, we check if the sufficient conditions in Theorem 1 are satisfied for the estimated IMPs ($k$, $R$, $S$), summarized in Table I. We observe that the conditions on $(k, S)$ are satisfied for the majority of cases, while the condition on $R$ holds with a noticeably lower empirical probability. Note that our sufficient conditions might be conservative. Moreover, due to the randomly selected coefficients, the intervention strength can be quite weak in some cases, making it challenging to distinguish true IMPs and non-IMPs. Fortunately, the averaging step helps to mitigate inaccurate predictions made by non-IMPs.

2) Interventions on $Y$: We consider the response $Y$ to be intervened through both the coefficients and shifts. We randomly select $n_p \sim \text{Unif}[1,\ldots,|PA(Y)|]$ of parents of $Y$ to have varying coefficients. For each training environment, we add perturbation terms sampled from Unif$[-2,2]$ to the original coefficients. For each test environment, the perturbations are sampled from Unif$[-10,10]$. The shift intervention on $Y$ is the same as the shift interventions on $X$ in Section VII-A1. In this setting, since none of the baseline methods allow interventions on $Y$ through the coefficients, they cannot even improve upon OLS. In Fig. 5, IMP performs slightly better than IMP inv, which may be due to the fact that the IMP method aims to find all possible IMPs, but the IMP inv only looks for IMPs that imply invariance. To further examine our test procedures for identifying IMPs, we check if the sufficient conditions in Theorem 1 are satisfied for the estimated IMPs ($k$, $R$, $S$), summarized in Table I. We observe that the conditions on $(k, S)$ are satisfied for the majority of cases, while the condition on $R$ holds with a noticeably lower empirical probability. Note that our sufficient conditions might be conservative. Moreover, due to the randomly selected coefficients, the intervention strength can be quite weak in some cases, making it challenging to distinguish true IMPs and non-IMPs. Fortunately, the averaging step helps to mitigate inaccurate predictions made by non-IMPs.

3) Interventions on Both $X$ and $Y$: The setting of interventions on both $X$ and $Y$ is simply a combination of the two settings above. In this challenging setting, our method outperforms the baselines by a large margin. We provide some additional observations on this experiment in Appendix K in the supplementary material. An experiment with continuous environments is provided in Appendix L in the supplementary material. Overall, the performance of our IMP algorithm in the continuous setting is similar to that in the discrete setting. In Appendix M in the supplementary material, we examine the robustness of IMP by intervening on every variable through every parameter.

B. COVID Data Set

For observational data, there is often no ground truth regarding the set of variables that are intervened. If the data is not collected under a carefully designed experimental setting (e.g., the gene perturbation experiments from [8]), it is reasonable to assume that every variable is more or less intervened, especially the response. To examine our methods under real-world distribution shifts, we consider a COVID data set [58] collected at 3142 U.S. counties from January 22, 2020, to June 10, 2021. The data set consists of 46 predictive features that are relevant to the number of COVID cases. Some of the features are treated as fixed or time-invariant, e.g., population.

A. Discrete Environments

First, we generate linear SCMs $\mathcal{M}^e$’s without interventions. For each $e \in \mathcal{E}^{\text{train}} = \{1, \ldots, 5\}$ or $e \in \mathcal{E}^{\text{test}} = \{6, \ldots, 10\}$, we randomly generate a linear SCM with 9 variables as follows. The graph $G(\mathcal{M}^e)$ is specified by a lower triangular matrix of i.i.d. Bernoulli$(1/2)$ random variables. The response $Y$ is randomly selected from the 9 variables and we require that $Y$ has at least one parent and one child in $G(\mathcal{M}^e)$. When $X$ and $Y$ are both intervened, we randomly choose a child of $Y$ to not be intervened. For each linear SCM, the non-zero coefficients are sampled from Unif$[-1.5,-0.5] \cup [0.5,1.5]$ and the noise variables are standard normal. For each training or test environment, we simulate i.i.d. data of sample size 300.

1) Interventions on $X$: Since the baseline methods have been examined extensively under shift interventions, we focus on shift interventions on $X$ for comparison. The general interventions on $X$ will be considered in Appendix in the supplementary material. Specifically, for each training environment, we randomly selected 4 predictors to be intervened through shifts sampled from Unif$[-2,2]$. For each test environment, the shifts are sampled from Unif$[-10,10]$. In Fig. 4, IMP inv performs similarly to SR since they share a similar idea when only $X$ is intervened. Due to the averaging steps of IMP, IMP inv, and SR, they have smaller variances compared with OLS and AR (a similar result has been reported in [19]).

2) Interventions on $Y$: We consider the response $Y$ to be intervened through both the coefficients and shifts. We randomly select $n_p \sim \text{Unif}[1,\ldots,|PA(Y)|]$ of parents of $Y$ to have varying coefficients. For each training environment, we add perturbation terms sampled from Unif$[-2,2]$ to the original coefficients. For each test environment, the perturbations are sampled from Unif$[-10,10]$. The shift intervention on $Y$ is the same as the shift interventions on $X$ in Section VII-A1. In this setting, since none of the baseline methods allow interventions on $Y$ through the coefficients, they cannot even improve upon OLS. In Fig. 5, IMP performs slightly better than IMP inv, which may be due to the fact that the IMP method aims to find all possible IMPs, but the IMP inv only looks for IMPs that imply invariance. To further examine our test procedures for identifying IMPs, we check if the sufficient conditions in Theorem 1 are satisfied for the estimated IMPs ($k$, $R$, $S$), summarized in Table I. We observe that the conditions on $(k, S)$ are satisfied for the majority of cases, while the condition on $R$ holds with a noticeably lower empirical probability. Note that our sufficient conditions might be conservative. Moreover, due to the randomly selected coefficients, the intervention strength can be quite weak in some cases, making it challenging to distinguish true IMPs and non-IMPs. Fortunately, the averaging step helps to mitigate inaccurate predictions made by non-IMPs.

3) Interventions on Both $X$ and $Y$: The setting of interventions on both $X$ and $Y$ is simply a combination of the two settings above. In this challenging setting, our method outperforms the baselines by a large margin. We provide some additional observations on this experiment in Appendix K in the supplementary material. An experiment with continuous environments is provided in Appendix L in the supplementary material. Overall, the performance of our IMP algorithm in the continuous setting is similar to that in the discrete setting. In Appendix M in the supplementary material, we examine the robustness of IMP by intervening on every variable through every parameter.

| $P(k \notin PE)$ | $P(PE \subseteq \hat{R})$ | $P(PE \subseteq \hat{S})$ |
|------------------|------------------|------------------|
| IMP             | 0.9610 (0.0796)  | 0.7503 (0.2511)  | 0.9376 (0.1079)  |
| IMP inv         | 0.9603 (0.0850)  | 0.7060 (0.2617)  | 0.9373 (0.1090)  |
TABLE II
EXPERIMENT RESULTS OF THE COVID DATA SET

| Cities/Counties (total cases) | IMP  | OLS  | AR   | SR   | CIP  | DIP  | CRM  |
|-----------------------------|------|------|------|------|------|------|------|
| Baltimore (18k)             | 0.1292 | 0.0937 | 0.0379 | 0.0978 | 0.0184 | 0.0199 | 0.3503 |
| Boston (26k)                | 0.0238 | 0.0503 | 0.0351 | 0.0543 | 0.0192 | 0.5820 | 0.4605 |
| Philadelphia (32k)          | 0.1352 | 0.1191 | 0.0701 | 0.0970 | 0.0333 | 1.0407 | 0.4217 |
| NY County (34k)             | 0.1611 | 0.4088 | 0.5606 | 0.3449 | 0.3537 | 2.4055 | 2.6372 |
| average mean RSS            | 0.1168 | 0.1680 | 0.1759 | 0.1485 | 0.1066 | 1.2370 | 0.9674 |
| Queens County (73k)         | 0.1126 | 0.1020 | 0.0882 | 0.1109 | 0.0673 | 0.9389 | 0.9647 |
| Houston (143k)              | 1.0884 | 1.2042 | 1.2200 | 1.1673 | 1.3705 | 1.7802 | 1.4982 |
| Chicago (146k)              | 0.3687 | 0.4489 | 0.4026 | 0.4099 | 0.4652 | 0.6328 | 0.3499 |
| Miami (171k)                | 0.4534 | 0.4621 | 0.5170 | 0.4025 | 0.6011 | 1.4465 | 0.8004 |
| average mean RSS            | 0.5058 | 0.5543 | 0.5548 | 0.5226 | 0.6238 | 1.1996 | 0.9033 |

VIII. DISCUSSION

To deal with general interventions on the response, we introduce the IMP that allows for an alternative form of invariance when the traditional invariance principle fails. We provide explicit characterizations of the IMP under different intervention settings and provide asymptotic generalization error analysis for both the discrete environment and the more challenging continuous environment settings, supported by our empirical studies.

Our work is motivated by allowing general interventions on the response, and it can be extended into several directions. First, we observe a connection between the IMP and self-supervised learning schemes (e.g., [59] from natural language processing). Specifically, the prediction of $X_k$ using $X_R$ corresponds to the pretext (or pretrained) task, and the linear relation in the IMP solves the downstream prediction task. We believe that methods developed through causal modeling including our IMP method will bring new opportunities for self-supervised learning. We have discussed several heuristic approaches to alleviate the computation complexity issue; it would be worthwhile to analyze such methods and provide theoretical guarantees (e.g., exploiting sparsity beyond the setting in [56]). Our asymptotic analysis of the generalization error is developed given that IMPs are correctly identified. It would be interesting to further analyze the consistency of the set of IMPs $\hat{I}$ identified by the proposed algorithms or new efficient algorithms.

REFERENCES

[1] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Dataset Shift in Machine Learning, Cambridge, MA, USA: Massachusetts Inst. Tech., 2008.
[2] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” J. Big Data, vol. 3, no. 1, pp. 1–40, 2016.
[3] G. Csurka, “Domain adaptation for visual applications: A comprehensive survey,” 2013, arXiv:1702.05374.
[4] J. Pearl, Causality. Cambridge, U.K.: Cambridge Univ., 2009.
