An Invariant Matching Property for Distribution Generalization under Intervened Response

Kang Du and Yu Xiang
University of Utah
50 S Central Campus Dr #2110 Salt Lake City, UT, USA
Email: {kang.du, yu.xiang}@utah.edu

***Abstract*—The task of distribution generalization concerns making reliable prediction of a response in unseen environments. The structural causal models are shown to be useful to model distribution changes through intervention. Motivated by the fundamental invariance principle, it is often assumed that the conditional distribution of the response given its predictors remains the same across environments. However, this assumption might be violated in practical settings when the response is intervened. In this work, we investigate a class of model with an intervened response. We identify a novel form of invariance by incorporating the estimates of certain features as additional predictors. Effectively, we show this invariance is equivalent to having a deterministic linear matching that makes the generalization possible. We provide an explicit characterization of the linear matching and present our simulation results under various intervention settings.**

***Index Terms*—Distribution generalization, invariance, causal structural model, LMMSE estimator**

1. Introduction

Consider the problem of predicting the response Y given its predictors $X = (X_1, ..., X_d)^\top$ in unseen environments. To model distribution changes in different environments (or training and test distributions), the common assumption is that the assignment for Y does not change across environments (or Y is not intervened). The structural causal models (SCMs) [1], [2] allow for natural formulations of the conditional distribution of Y given X [2]–[8], and the underlying principle is known as invariance, autonomy or modularity [1], [9]–[11]. For instance, in the invariance causal prediction framework [5], it is assumed that the conditional distribution of Y given a set of predictors $X_S \subseteq \{X_1, ..., X_d\}$ is invariant in all environments; a relaxed version is adopted in the stabilized regression method [12] where only the conditional mean is assumed to be invariant.

In practical settings, however, the structural assignment of Y might change across environments, i.e., Y might be intervened. We thus believe there is a need of relaxing the assumption and exploring alternative forms of invariance. To shed some light on this more challenging setting, we propose to model Y as

$$Y = f_U(X_{\mathrm{PA}(Y)}, \epsilon_Y),$$

where $\mathrm{PA}(Y)$ denotes the set of direct (causal) parents of Y, and ϵ_Y is an independent noise, and a (discrete) random variable U is introduced to capture the dependence of structural assignment on different environments (i.e., each $U = u$ corresponds to one environment). The main challenge lies in whether it is still possible to identify forms of invariance to facilitate prediction in unseen environments. In this work, we make an attempt in this direction by focusing on a mixture of linear SCM models, and the assignment for Y is

$$Y = a(U)^\top X^1_{\mathrm{PA}(Y)} + b^\top X^2_{\mathrm{PA}(Y)} + \epsilon_Y, \tag{1}$$

where $X_{\mathrm{PA}(Y)}$ is partitioned into $X^1_{\mathrm{PA}(Y)}$ and $X^2_{\mathrm{PA}(Y)}$, and coefficients $a(U)$ formalize the changing conditional distributions; furthermore, we consider a training model and a testing model, and allow $a(\cdot)$ to be arbitrarily different under the two models. To make reliable predictions on the testing model, we identify an additional class of predictors that are computed based on the linear minimum mean square error (LMMSE) estimators of X_k given X_S for any fixed $U = u$, for $k \in \{1, ..., d\}$ and $S \subseteq \{1, ..., d\} \setminus k$. Roughly speaking, these predictors (along with the original ones X) allows for a deterministic relation for predicting Y, with coefficients that are invariant for all environments. This makes the generalization task possible, as one can then reuse the coefficients for unseen environments or test data.

2. Background and Problem Formulation

We now formally introduce the model and rewrite (1) in a more compact form. Let $\mathcal{U}$ denote a set of environments where a target variable $Y \in \mathbb{R}$ and a vector of predictors $X = (X_1, \ldots, X_d)^\top \in \mathcal{X} \subseteq \mathbb{R}^{d\times 1}$ are observed, we assume $U \in \mathcal{U}$ and (X, Y) satisfy an acyclic SCM that we call the training SCM,

$$\mathcal{S}: \begin{cases} U = \varepsilon_U \\ X = \gamma Y + BX + \varepsilon_X & (2) \\ Y = (\beta + \alpha(U))^\top X + \varepsilon_Y, & (3) \end{cases}$$

where $\varepsilon_U \in \mathcal{U}$, $\varepsilon_Y \in \mathbb{R}$, $\alpha(U), \beta, \gamma, \varepsilon_X \in \mathbb{R}^{d\times 1}$, $B \in \mathbb{R}^{d\times d}$, and the noise variables $\varepsilon_{X_1}, \ldots, \varepsilon_{X_d}, \varepsilon_U$, and ε_Y are jointly independent, and assume that $\alpha(U)$ is a nondegenerate random variable (i.e., not a one-point distribution). This nonlinear SCM $\mathcal{S}$ can be viewed as a mixture of linear SCMs, since $\mathcal{S}$ is linear when conditioning on $U = u$. The causal graph $\mathcal{G}(\mathcal{S})$ induced by $\mathcal{S}$ can be drawn according to the nonzero coefficients in $\mathcal{S}$. Without loss of generality, we require $\{X_j : \beta_j \neq 0\}$ and $\{X_j : \alpha_j(U) \neq 0\}$ to be two distinct sets of parents of Y.

We assume that the variable U is a root node in $\mathcal{G}(\mathcal{S})$ such that only the parameters that are functions of U may change in the testing SCM defined below. The reason behind this assumption is that if there is no evidence that a parameter is changing for a diverse set of environments in the training data, then that parameter is likely to remain invariant in the test data or any unseen environments.

Remark 1: For an intercept term in (3) that depends on U, it can be taken as the coefficient of $X_1 = 1$. For simplicity, we assume that ε_X and ε_Y have zero means, which implies $\mathsf{E}[X|U=u] = \mathsf{E}[Y|U=u] = 0$ and thus $\mathsf{E}[X] = \mathsf{E}[Y] = 0$ (the same goes for the testing SCM defined below).

Remark 2: In [13], the authors have shown that a form of varying filter connecting feature and response (as a special case of the varying coefficients in (3)) is effective for causal inference tasks, by adopting estimators from [14].

Similarly, let $\mathcal{U}^\tau$ denote a set of unseen environments ($U^\tau = u^\tau$) where the observed variables $U^\tau \in \mathcal{U}^\tau$ and $X^\tau = (X_1^\tau, \ldots, X_d^\tau)^\top \in \mathcal{X}^\tau \subseteq \mathbb{R}^{d\times 1}$ and the unobserved variable $Y^\tau \in \mathbb{R}$ follow an acyclic testing SCM,

$$\mathcal{S}^\tau : \begin{cases} U^\tau = \varepsilon_{U^\tau} \\ X^\tau = \gamma Y^\tau + B X^\tau + \varepsilon_{X^\tau} \\ Y^\tau = (\beta + \alpha^\tau(U^\tau))^\top X^\tau + \varepsilon_{Y^\tau}, \end{cases}$$

where the noise variables $\varepsilon_{X_1^\tau}, \ldots, \varepsilon_{X_d^\tau}, \varepsilon_{U^\tau}$, and ε_{Y^τ} are jointly independent, and $(\varepsilon_{X^\tau}^\top, \varepsilon_{Y^\tau})$ and $(\varepsilon_X^\top, \varepsilon_Y)$ are equal in distribution. Since we assume that only the parameters that are functions of U in $\mathcal{S}$ may change in $\mathcal{S}^\tau$, we have $\alpha_j^\tau(U^\tau) = 0$ for any $j \in \{1, \ldots, d\}$ such that $\alpha_j(U) = 0$.

In this work, we consider the setting when the only parameter in $\mathcal{S}$ that depends on U is the coefficient vector $\alpha(\cdot)$, but $\alpha(\cdot)$ and $\alpha^\tau(\cdot)$ can be arbitrarily different. By assuming the independence of ε_U and ε_Y, the distribution of ε_Y remains invariant when conditioning on $U = u$ for different u, while the case when the variance of ε_Y changes arbitrarily with respect to u can be challenging, since $\mathrm{Var}(\varepsilon_Y)$ is simply the MMSE of the estimator $\mathsf{E}[Y|X_{\mathrm{PA}(Y)}, U]$. Another setting when only the parameters in the assignments of the predictors are allowed to depend U (i.e., only X is intervened) is considered in the stabilized regression framework [12], where a weaker version of the causal invariance property [5] is assumed. Using our notation, it is assumed that there exists $S \subseteq \{1, \ldots, d\}$ such that $\mathsf{E}[Y|X_S = x, U = u] = \mathsf{E}[Y|X_S = x] \triangleq g(x)$ holds for all x and u. Since $g(x)$ does not depend on u, the above relation remains the same for both the training and testing SCMs (for instance, $\mathsf{E}[Y^\tau|X_S^\tau = x^\tau, U^\tau = u^\tau] = \mathsf{E}[Y^\tau|X_S^\tau = x^\tau] = g(x^\tau)$ holds for all x and u^τ). This assumption allows one to select predictors that provide consistent predictions of the target variable across the observed and unseen environments. In general, the assumption is violated for the SCMs $\mathcal{S}$ and $\mathcal{S}^\tau$ when Y is intervened, or equivalently, when the parameters depending on U appear in the assignment of Y.

Our invariance property relies on the LMMSE estimators of a target variable $Y \in \mathbb{R}$ given a vector of predictors $X \in \mathbb{R}^{p\times 1}$, denoted by $\mathsf{E}_l[Y|X] = (\theta^{\mathrm{ols}})^\top (X - \mathsf{E}[X]) + \mathsf{E}[Y]$, where $\theta^{\mathrm{ols}} \triangleq \mathrm{Cov}(X,X)^{-1}\mathrm{Cov}(X,Y)$ is also called the population ordinary least squares (OLS) estimator. For the SCM $\mathcal{S}$, we denote the LMMSE estimator of Y given X when conditioning on $U = u$ as $\mathsf{E}_l[Y|X; U=u] \triangleq (\theta^{\mathrm{ols}}(u))^\top X$ with its OLS estimator $\theta^{\mathrm{ols}}(u) \in \mathbb{R}^{d\times 1}$. And correspondingly, we define $\mathsf{E}_l[Y|X;U] \triangleq (\theta^{\mathrm{ols}}(U))^\top X$ that is linear in X but with coefficients depending on U. Equivalently, one can define $\mathsf{E}_l[Y|X;U]$ by

$$\mathsf{E}_l[Y|X;U] = \operatorname*{argmin}_{l^\top(U)X \in \mathcal{L}} \mathsf{E}\left[|Y - l^\top(U)X|^2\right], \tag{4}$$

where $\mathcal{L} = \{l^\top(U)X \mid l : \mathcal{U} \to \mathbb{R}^{d\times 1}\}$ is a class of functions that are linear in X but with coefficients depending on U. This function class is introduced as it is compatible with the form of the assignment of Y in (3). Similarly, we have the function class $\mathcal{L}^\tau$ for the testing SCM $\mathcal{S}^\tau$. It is important to note that even though $\mathsf{E}_l[Y|X;U]$ achieves the minimum prediction error for Y (as in (4)), it may not be applicable for predicting Y^τ since $\mathcal{U}$ and $\mathcal{U}^\tau$ may differ in general. And in fact, the prediction error of using $\mathsf{E}_l[Y|X;U]$ for Y^τ can be arbitrarily high as we do not restrict the forms of $\alpha(\cdot)$ and $\alpha^\tau(\cdot)$. In the next section, we show that this issue can be resolved via our invariance property.

3. Invariant Matching Property

A. One Motivating Example

Example 1: Consider $(Y, X^\top, U) \triangleq (Y, X_1, X_2, X_3, U)$ satisfying the following acyclic SCM (illustrated in Fig. 1),

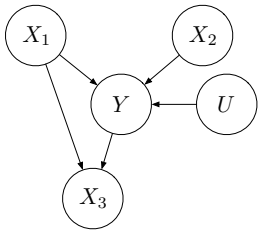

Fig. 1. Directed acyclic graph $\mathcal{G}(\mathcal{S}_{\text{toy}})$.

$$\mathcal{S}_{\text{toy}} : \begin{cases} Y = a(U)X_1 + X_2 + N_Y \\ X_3 = Y + X_1 + N_3, \end{cases} \tag{5}$$

where U, X_1, X_2, N_3, and N_Y are jointly independent, and X_1, X_2, N_Y, and N_3 are $\mathcal{N}(0,1)$-distributed. The testing SCM $\mathcal{S}_{\text{toy}}^\tau$ over $(Y^\tau, X_1^\tau, X_2^\tau, X_3^\tau, U^\tau)$ can be defined similarly, where X_1^τ has a coefficient $a^\tau(U^\tau)$. Since (Y, X) is multivariate Gaussian given $U = u$, the MMSE estimator of Y using X given $U = u$ is

$$\begin{aligned} \mathsf{E}[Y|X, U=u] =& X^\top \left(\mathsf{E}[XX^\top|U=u]\right)^{-1} \mathsf{E}[XY|U=u] \\ =& \frac{1}{2}(a(u)-1)X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3, \end{aligned}$$

which implies $\mathsf{E}[Y|X,U] = \frac{1}{2}(a(U)-1)X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3$. Similarly, one can compute $\mathsf{E}[X_3|X_1, X_2, U] = (1 + a(U))X_1 + X_2$. Observe that $\mathsf{E}[Y|X_1, X_2, X_3, U]$ and $\mathsf{E}[X_3|X_1, X_2, U]$ are two linear combinations of

$\{a(U)X_1, X_1, X_2, X_3\}$; $a(U)X_1$ can not be linearly represented by $\{X_1, X_2, X_3\}$. Thus, there exists a deterministic linear relation

$$\mathsf{E}[Y|X,U] = \lambda\,\mathsf{E}[X_3|X_1,X_2,U] + \eta^\top X, \tag{6}$$

with unique coefficients $\lambda = 1/2$ and $\eta = (-1, 0, 1/2)^\top$ that *do not depend on* U. Furthermore, since the right-hand side of (6) is a linear function of $\mathsf{E}[X_3|X_1, X_2, U]$ and X, and it equals to the MMSE estimator $\mathsf{E}[Y|X,U]$ among all functions of X and U, we obtain an invariant relation

$$\begin{aligned}\mathsf{E}[Y|X,U] &= \mathsf{E}_l\left\{Y \,\middle|\, \mathsf{E}[X_3|X_1,X_2,U], X\right\} \\ &= \frac{1}{2}\,\mathsf{E}[X_3|X_1,X_2,U] - X_1 + \frac{1}{2}X_3 \end{aligned} \tag{7}$$

for the training SCM. Since λ and η are not functions of U, they will remain invariant for the testing SCM. Note that $\mathsf{E}[X_3^\tau|X_1^\tau, X_2^\tau, U^\tau]$ is determined by the distribution of (X^τ, U^τ). A prediction model like (7) with invariant coefficients is often not unique when it exists. One can show that $\mathsf{E}[Y|X,U] = -\frac{3}{2}\,\mathsf{E}[X_2|X_1,X_3,U] - X_1 + \frac{1}{2}X_2 + X_3$, however, this does not hold for $\mathsf{E}[X_1|X_2,X_3,U]$.

Remark 3: In general, an invariance relation (7) may not hold for the MMSE estimator $\mathsf{E}[Y|X,U]$ if (X,Y) is not Gaussian when conditioning on $U = u$ for each $u \in \mathcal{U}$. In this work, we do not require Gaussianity, and we focus on the LMMSE estimator detailed in the next section.

B. Invariance Matching Property

Our invariant matching property is motivated by the following observation: If X includes any descendants of Y, then $\alpha(\cdot)$ (that may change in the unseen environments) will be passed on to the descendants. In other words, if Y has at least one child, there will be certain dependency between the mechanism that generates Y and certain statistical properties of X, which is also true for Y^τ and X^τ. Thus, the change of the function $\alpha(\cdot)$ can be revealed by the changes of certain statistical properties of X. As illustrated in the motivating example, we have identify features of the form $\mathsf{E}_l[X_k|X_S;U]$ to be useful for prediction.

Formally, we say that a model $\mathcal{S}$ satisfies the *invariant matching property* if there exists $k \in \{1,\dots,d\}$ and $S \subseteq \{1,\dots,d\} \setminus k$ such that

$$\mathsf{E}_l[Y|X\,;U] = \mathsf{E}_l\left\{Y \,\middle|\, X, \mathsf{E}_l[X_k|X_S;U]\right\} \tag{8}$$

$$= \lambda\,\mathsf{E}_l[X_k|X_S;U] + \eta^\top X, \tag{9}$$

where parameters λ and η do not depend on U, and the same holds for the testing SCM $\mathcal{S}^\tau$, i.e.,

$$\mathsf{E}_l[Y^\tau|X^\tau;U^\tau] = \lambda\,\mathsf{E}_l[X_k^\tau|X_S^\tau;U^\tau] + \eta^\top X^\tau.$$

In general, due to the difference between $\alpha(\cdot)$ and $\alpha^\tau(\cdot)$,

$$\mathsf{E}_l[Y|X=x;U=u] = \mathsf{E}_l[Y^\tau|X^\tau=x;U^\tau=u]$$

does not hold for $x \in \mathcal{X} \cup \mathcal{X}^\tau$ and $u \in \mathcal{U} \cup \mathcal{U}^\tau$ even if U and U^τ are equal in distribution. By introducing some feature $\mathsf{E}_l[X_k|X_S;U]$, the invariant matching property bridges $\mathsf{E}_l[Y|X=x;U=u]$ and $\mathsf{E}_l[Y^\tau|X^\tau=x;U^\tau=u]$ (for the same (x,u)) with a linear relation that remains invariant across all observed and unseen environments.

In our invariant matching property, note that (9) follows from (8) by the definition of linear MMSE. Now we show the other direction is also true in the following technical lemma.

Lemma 1: For some $k \in \{1,\dots,d\}, S \subseteq \{1,\dots,d\} \setminus k$,

$$\mathsf{E}_l[Y|X;U] = \mathsf{E}_l\left\{Y \,\middle|\, X, \mathsf{E}_l[X_k|X_S;U]\right\} \tag{10}$$

if and only if there exists $\lambda \in \mathbb{R}$ and $\eta \in \mathbb{R}^{d\times 1}$ such that

$$\mathsf{E}_l[Y|X;U] = \lambda\,\mathsf{E}_l[X_k|X_S;U] + \eta^\top X. \tag{11}$$

holds for all $u \in \mathcal{U}$.

C. Characterization of the Features

An important fact about a feature of the form $\mathsf{E}_l[X_k|X_S;U]$ is that it does not depend on Y, so the corresponding feature $\mathsf{E}_l[X_k^\tau|X_S^\tau;U^\tau]$ for the testing SCM will not depend on the unobservable variable Y^τ. In other words, extracting the features only requires exploring the relations between the predictors while taking the target variables Y and Y^τ as unobserved. The consequence of Y being unobserved is that $(U, X_1, \dots, X_d)^\top$ no longer follows a mixture of linear SCMs, but a mixture of linear models with a set of dependent noise variables. Specifically, when Y is unobserved (or equivalently, substitute Y in (3) into (2)), then the relations between the predictors are as follows,

$$X = \left(\gamma(\beta + \alpha(U))^\top + B\right) X + \gamma\varepsilon_Y + \varepsilon_X, \tag{12}$$

where $\gamma\varepsilon_Y + \varepsilon_X$ is a vector of dependent random variables when γ non-zero. Observe that the function $\alpha(\cdot)$ is captured by the relations of the predictors only if γ is not a zero vector in (12), which brings up the following key assumption.

Assumption 1: Y has at least one child.

The equivalent definition of our invariant matching property in (11) allows for simpler evaluation through computation. In order to verify (11) for a particular feature $\mathsf{E}_l[X_k|X_S;U]$, we compute the LMMSE estimators $\mathsf{E}_l[X_k|X_S;U=u]$ and $\mathsf{E}_l[Y|X;U=u]$ and check whether there exists coefficients for (11) to hold. In the following theorem, we show that it holds for a wide class of $k \in \{1,\dots,d\}$ and $S \subseteq \{1,\dots,d\} \setminus k$.

Theorem 1: There exists $\lambda_Y \in \mathbb{R}$ and $\eta_Y \in \mathbb{R}^{d\times 1}$ such that

$$\mathsf{E}_l[Y|X;U=u] = (\lambda_Y\alpha(u) + \eta_Y)^\top X$$

holds for every $u \in \mathcal{U}$. For each $k \in \{j : \alpha_j(u) = 0\}$ and $S \subseteq \{1,\dots,d\} \setminus k$ such that $\{j : \alpha_j(u) \neq 0\} \subseteq S$, there exists $\lambda_{k,S} \in \mathbb{R}$ and $\eta_{k,S} \in \mathbb{R}^{d\times 1}$ such that

$$\mathsf{E}_l[X_k|X_S;U=u] = (\lambda_{k,S}\alpha(u) + \eta_{k,S})^\top X$$

holds for every $u \in \mathcal{U}$. If $\lambda_{k,S} \neq 0$, then the relation

$$\mathsf{E}_l[Y|X;U=u]$$

$$= \frac{\lambda_Y}{\lambda_{k,S}} \mathsf{E}_l[X_{k,S}|X_S; U = u] + \left(\eta_Y - \frac{\lambda_Y}{\lambda_{k,S}}\eta_{k,S}\right) X,$$

holds for every $u \in \mathcal{U}$.

In this theorem, we provide a complete characterization of the invariant matching property (9) for $\mathsf{E}_l[Y|X; U = u]$ and $\mathsf{E}_l[X_k|X_S; U = u]$, which allows us to use $\mathsf{E}_l[X_k|X_S; U]$ as a predictor for Y (see Algorithm 1 for implementation details).

Remark 4: Observe that $\lambda_{k,S}$ is nonzero if and only if there exists different $u_1, u_2 \in \mathcal{U}$ such that

$$\mathsf{E}_l[X_k|X_S; U = u_1] \neq \mathsf{E}_l[X_k|X_S; U = u_2]. \tag{13}$$

This is true in generic cases when $X_k \not\perp\!\!\!\perp U|X_S$. Note that if Assumption 1 is not satisfied (Y has no children), then $X_k \perp\!\!\!\perp U|X_S$ since U is a root node and has only one descendent Y.

4. Algorithm

For each $k \in \{1, \ldots, d\}$, $S \subseteq \{1, \ldots, d\} \setminus k$, we estimate the following LMMSE estimator for the prediction of Y,

$$\begin{aligned} Y(k, S) &= \mathsf{E}_l\left\{ Y \,\middle|\, X, \mathsf{E}_l[X_k|X_S; U] \right\} \\ &= \lambda(k, S)\, \mathsf{E}_l[X_k|X_S; U] + \eta^\top(k, S)X. \end{aligned} \tag{14}$$

According to Lemma 1, a feature $\mathsf{E}_l[X_{k_*}|X_{S_*}; U]$ that satisfies the invariant matching property if and only if $Y(k_*, S_*)$ achieves the minimum prediction error for Y among the prediction errors of all possible $Y(k, S)$'s. Since such feature is not unique in general (shown by the toy example in Section 3-A), we do not choose the feature that leads to the lowest prediction error. Rather, we look for features with prediction errors below a certain threshold ε (see the end of this section for the determination of ε).

For the training SCM $\mathcal{S}$ and testing SCM $\mathcal{S}^\tau$, let $\mathcal{U} = \{u_1, \ldots, u_p\}$ and we denote $\mathcal{U}^\tau \triangleq \mathcal{V} = \{v_1, \ldots, v_q\}$ for simplicity of notation. For each $u_i \in \mathcal{U}$, we are given the i.i.d. training data $\boldsymbol{X}^{u_i} \in \mathbb{R}^{n(u_i)\times d}, \boldsymbol{Y}^{u_i} \in \mathbb{R}^{n(u_i)\times 1}$, and for each $v_i \in U^\tau$, we observe the i.i.d. testing data $\boldsymbol{X}^{v_i} \in \mathbb{R}^{m(v_i)\times d}$. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with $n \triangleq \sum_{i=1}^p n(u_i)$ denote the pooled data matrix of all $\boldsymbol{X}^{u_i}, u_i \in \mathcal{U}$. Similarly, we define the pooled data matrices $\boldsymbol{Y} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{X}^\tau \in \mathbb{R}^{m\times d}$ with $m \triangleq \sum_{i=1}^q m(u_i)$.

In Algorithm 1, for $k \in \{1, \ldots, d\}, S \in \{1, \ldots, d\} \setminus k$, we adopt the OLS estimator to estimate the feature vectors

$$\hat{\mathsf{E}}_l[\boldsymbol{X}_k^{u_i}|\boldsymbol{X}_S^{u_i}] = \boldsymbol{X}_S^{u_i}\left((\boldsymbol{X}_S^{u_i})^\top \boldsymbol{X}_S^{u_i}\right)^{-1}(\boldsymbol{X}_S^{u_i})^\top \boldsymbol{X}_k^{u_i}, \tag{15}$$

$$\hat{\mathsf{E}}_l[\boldsymbol{X}_k^{v_i}|\boldsymbol{X}_S^{v_i}] = \boldsymbol{X}_S^{v_i}\left((\boldsymbol{X}_S^{v_i})^\top \boldsymbol{X}_S^{v_i}\right)^{-1}(\boldsymbol{X}_S^{v_i})^\top \boldsymbol{X}_k^{v_i}, \tag{16}$$

for the training data and the testing data, respectively.

Let $\tilde{\boldsymbol{X}}(k, S) \triangleq \left(\hat{\mathsf{E}}_l[\boldsymbol{X}_k|\boldsymbol{X}_S], \boldsymbol{X}\right) \in \mathbb{R}^{n\times(d+1)}$ denote the augmented design matrix. For feature selection on the training data, we compute the prediction residuals of Y as follows

$$\boldsymbol{R}(k, S) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}(k, S), \tag{17}$$

with an estimate of $\boldsymbol{Y}(k, S)$ (the vector form of $Y(k, S)$) as

$$\begin{aligned} \hat{\boldsymbol{Y}}(k, S) &= \tilde{\boldsymbol{X}}(k, S)\left(\tilde{\boldsymbol{X}}^\top(k, S)\tilde{\boldsymbol{X}}(k, S)\right)^{-1}\tilde{\boldsymbol{X}}^\top(k, S)\boldsymbol{Y} \\ &\triangleq \tilde{\boldsymbol{X}}(k, S)\beta(k, S), \end{aligned}$$

where the OLS estimator $\beta(k, S) \in \mathbb{R}^{d\times 1}$ can be reused for predicting $\boldsymbol{Y}^\tau$. That is, on the testing data we compute

$$\hat{\boldsymbol{Y}}^\tau(k, S) = \tilde{\boldsymbol{X}}^\tau(k, S)\beta(k, S), \tag{18}$$

where $\tilde{\boldsymbol{X}}^\tau(k, S) \triangleq \left(\hat{\mathsf{E}}_l[\boldsymbol{X}_k^\tau|\boldsymbol{X}_S^\tau], \boldsymbol{X}^\tau\right) \in \mathbb{R}^{m\times(d+1)}$.

Algorithm 1 Generalizable Prediction via Invariant Matching

procedure Select Features on the Training Data
  for $k \in \{1, \ldots, d\}$ **do**
    for $S \subseteq \{1, \ldots, d\} \setminus k$ **do**
      (i) Compute the feature vector $\hat{\mathsf{E}}_l[\boldsymbol{X}_k^{u_i}|\boldsymbol{X}_S^{u_i}]$ for each u_i by (15), and combine the feature vectors into one vector $\hat{\mathsf{E}}_l[\boldsymbol{X}_k|\boldsymbol{X}_S]$
      (ii) Compute $\boldsymbol{R}(k, S)$ in (17) and check whether $||\boldsymbol{R}(k, S)||_2^2 \leq \varepsilon$

procedure Extract The Selected Features on the Testing Data
  for every (k, S) such that $||\boldsymbol{R}(k, S)||_2^2 \leq \varepsilon$ **do**
    (i) Compute the feature vector $\hat{\mathsf{E}}_l[\boldsymbol{X}_k^{v_i}|\boldsymbol{X}_S^{v_i}]$ for each v_i by (16), and combine the feature vectors into one vector $\hat{\mathsf{E}}_l[\boldsymbol{X}_k^\tau|\boldsymbol{X}_S^\tau]$
    (ii) Predict $\boldsymbol{Y}^\tau$ using the feature $\hat{\mathsf{E}}_l[\boldsymbol{X}_k^\tau|\boldsymbol{X}_S^\tau]$ by computing $\hat{\boldsymbol{Y}}^\tau(k, S)$ according to (18)
  Output $\hat{\boldsymbol{Y}}^\tau$ as the average of all computed $\hat{\boldsymbol{Y}}^\tau(k, S)$

We determine the parameter ε using the residuals $\boldsymbol{R}(k, S)$ defined above. First, we run the first procedure in Algorithm 1 with a sufficiently large ε to compute the training prediction error $||\boldsymbol{R}(k, S)||_2^2$ for all (k, S)'s. Then, we rank all the prediction errors, and set ε be the $(100\alpha)\%$-quantile of the all prediction errors, where α controls the proportion of the features that will be selected. For the experiments in the next section, α is fixed to be 0.05.

5. Experiments

We compare our method with three baseline methods: Ordinary Least Squares (OLS), stabilized regression (SR) [12], and anchor regression (AR) [15]. For the anchor regression, we use a 5-fold cross-validation procedure to select the hyper-parameter γ from $\{0.2, 0.4, \ldots, 1\} \cup \{2, 3, \ldots, 5\}$.

Experiment A: Regular setting.

We randomly simulate 500 models generated as follows. Consider the training SCM $\mathcal{S}$ with 10 predictors and $\mathcal{U} = \{1, 2, \ldots, 5\}$. The acyclic graph $\mathcal{G}(\mathcal{S})$ is randomly generated with each edge existing with probability 0.5. In the generated graph, we require Y to have at least one parent and one child. For the parameters in the training SCM, all the coefficients in $\mathcal{S}$ that do not depend on U are randomly sampled from $\mathrm{Unif}[-1.5, -0.5] \cup [0.5, 1.5]$, and the noise variables ε_X and ε_Y are jointly independent standard normal random variables. For the intervention on Y, we choose $n_p \sim \mathrm{Unif}\{1, \ldots, |\mathrm{PA}(Y)|\}$ of the parents of Y to have a coefficient vector $\alpha(u) + \beta \in \mathbb{R}^{n_p\times 1}, u \in \mathcal{U}$. The coefficient vectors $\alpha(1), \alpha(2), \ldots, \alpha(5)$ are vectors of i.i.d. random variables

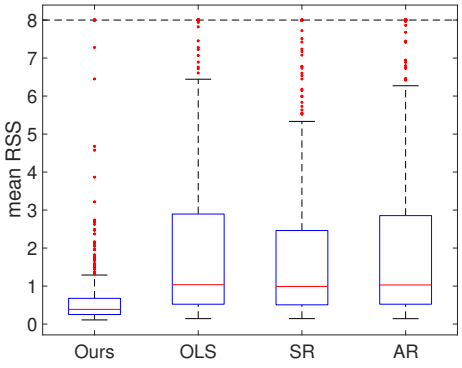

Fig. 2. Experiment A.

following $\mathrm{Unif}[-2, 2]$, and the elements of β are sampled from $\mathrm{Unif}[-1.5, -0.5] \cup [0.5, 1.5]$.

The testing SCM $\mathcal{S}^\tau$ has the same graph and parameters as $\mathcal{S}$ except that $\mathcal{U}^\tau = \{6, 7, \ldots, 10\}$ and the new coefficient vectors $\alpha^\tau(6), \alpha^\tau(7), \ldots, \alpha^\tau(10)$ are drawn from i.i.d. $\mathrm{Unif}[-10, 10]$. For each $u \in \mathcal{U}$ or $u^\tau \in \mathcal{U}^\tau$, the sample size is 300. Overall, Fig. 2 shows our method outperforms all three baseline methods by having smaller median and variance for the mean residual sum of squares (RSS).

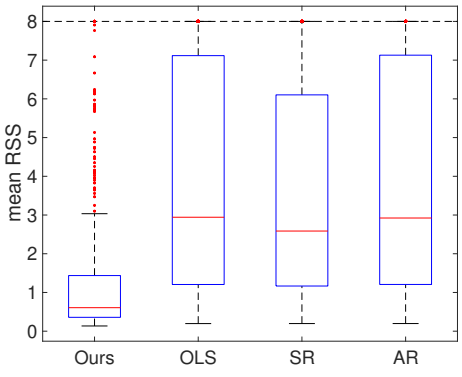

Fig. 3. Experiment B-1.

Experiment B: Reduced intervention on training data.

We consider cases when interventions on the training data are reduced from that in Experiment A, while the data generating process for the testing data remains the same as before.

1) Smaller variation of coefficient vectors: The coefficient vectors $\alpha(1), \ldots, \alpha(5)$ are vectors of i.i.d. entries according to $\mathrm{Unif}[-1, 1]$, reducing the variation of the coefficient vector.

2) Less number of environments: The support of the variable U is reduced to $\mathcal{U} = \{1, 2\}$. Accordingly, the sample size of the pooled training data is now $2 * 300 = 600$.

In Fig. 3 and Fig. 4, our method has a smaller median compared with the three baseline methods, while they have similar medians. Due to the averaging procedure over multiple prediction models in Algorithm 1, our method has smaller variances than OLS and AR. The averaging procedure of SR fails since their assumption that Y is not intervened is violated. Compared with Experiment A, the median and variance of our method are slightly larger, but our method is less sensitive with respect to the reduced interventions in comparison with the baseline methods.

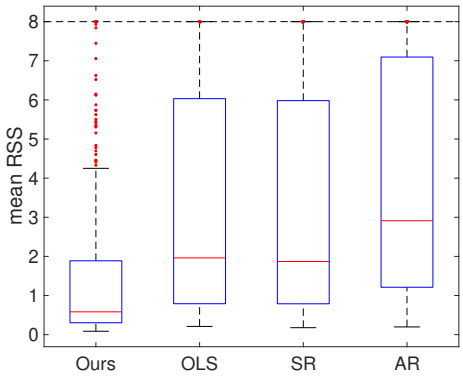

Fig. 4. Experiment B-2.

References

[1] J. Pearl, *Causality*. Cambridge University Press, 2009.

[2] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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

[4] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *International Conference on Machine Learning*. PMLR, 2013, pp. 819–827.

[5] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction: identification and confidence intervals," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

[6] 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.

[7] C. Heinze-Deml and N. Meinshausen, "Conditional variance penalties and domain shift robustness," *arXiv preprint arXiv:1710.11469*, 2017.

[8] P. Bühlmann, "Invariance, causality and robustness," *Statistical Science*, vol. 35, no. 3, pp. 404–426, 2020.

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

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

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

[12] 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.

[13] K. Du and Y. Xiang, "Causal inference from slowly varying nonstationary processes," *arXiv preprint arXiv:2012.13025*, 2020.

[14] Y. Xiang, J. Ding, and V. Tarokh, "Estimation of the evolutionary spectra with application to stationarity test," *IEEE Transactions on Signal Processing*, vol. 67, no. 5, pp. 1353–1365, 2019.

[15] 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.