Regularizing towards Causal Invariance: Linear Models with Proxies

Michael Oberst [1] **Nikolaj Thams** [2] **Jonas Peters** [2] **David Sontag** [1]

Abstract

We propose a method for learning linear models whose predictive performance is robust to causal interventions on unobserved variables, when noisy proxies of those variables are available. Our approach takes the form of a regularization term that trades off between in-distribution performance and robustness to interventions. Under the assumption of a linear structural causal model, we show that a single proxy can be used to create estimators that are prediction optimal under interventions of bounded strength. This strength depends on the magnitude of the measurement noise in the proxy, which is, in general, not identifiable. In the case of two proxy variables, we propose a modified estimator that is prediction optimal under interventions up to a known strength. We further show how to extend these estimators to scenarios where additional information about the "test time" intervention is available during training. We evaluate our theoretical findings in synthetic experiments and using real data of hourly pollution levels across several cities in China.

1. Introduction

Ideally, predictive models would generalize beyond the distribution on which they are trained, e.g., across geographic regions, across time, or across individual users. However, models often learn to rely on signals in the training distribution that are not stable across domains, causing a drop-off in predictive performance. This problem is broadly known as dataset shift (Quiñonero-Candela et al., 2009).

Tackling this problem requires a formalization of how dataset shift arises, and how that shift impacts the conditional distribution of our target Y given features X. One way to formalize this shift is in terms of an underlying causal graph (Pearl, 2009), where changes between distributions are seen as arising from causal interventions on variables.

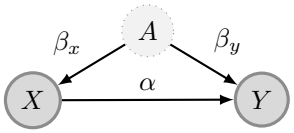

Figure 1. Conceptual Example: A represents an (unobserved) socioeconomic variable, X represents current health status, and Y represents a long-term health outcome. All relationships are assumed to be linear, and coefficients are given. We consider a broader class of graphs in this work, see Figure 2.

Conceptual example: In the causal graph given in Figure 1, the variable A serves as a confounder. In a medical setting, A could represent smoking habits or socioeconomic status, which have a causal effect on current health status (X) as well as longer-term outcomes (Y). Importantly, A may not be recorded in our training data, and the distribution of A could vary across geography and time.

In the context of this causal graph, interventions which change the distribution of A will also alter the conditional mean $\mathbb{E}(Y \mid X)$. Under the linear relationships in Figure 1, the optimal least-squares predictor $\hat{Y} = \gamma^* X$ under the test distribution depends on the test-time variance in A, in that

$$\gamma^* = \begin{cases} \alpha, & \text{if after intervention } A = 0 \\ \alpha + \frac{\beta_Y}{\beta_X}, & \text{if after intervention } \mathrm{Var}(A) \to \infty. \end{cases}$$

The first predictor encodes the direct causal effect of X on Y, but is only optimal in the setting where the correlations induced by A are removed by fixing it to a constant value of zero (the same holds when including intercepts and allowing for non-zero means). The second predictor, on the other hand, renders the distribution of the residual $Y - \hat{Y}$ independent of A, and is therefore robust to arbitrary interventions upon A. However, this is only optimal under arbitrarily strong interventions on A.

Balancing performance and invariance: Instead of seeking an invariant predictor that is robust to arbitrary interventions on A (like the second predictor above), we instead seek to minimize a worst-case loss under bounded interventions of a given strength. We contrast this with work that seeks to discover causal relationships as a route to invariance (Rojas-Carulla et al., 2018; Magliacane et al., 2018), optimize for

[1]EECS, MIT, Cambridge, USA [2]Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark. Correspondence to: Michael Oberst <moberst@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

invariance directly across environments (Arjovsky et al., 2019), or use known causal structure to select predictors with invariant performance (Subbaswamy et al., 2019).

Our proposed objective takes the form of a standard loss, plus a regularization term that encourages invariance. This builds upon Rothenhäusler et al. (2021), who introduce a similar objective, and prove that their objective optimizes a worst-case loss over bounded interventions on A, under a large class of linear structural causal models.

In contrast to Rothenhäusler et al. (2021), we do not assume that A is observed. Instead we assume that, during training, we have access to noisy proxies of A. For most of the paper, we assume that neither A nor proxies are available during testing. With this in mind, our contributions are as follows

- *Distributional robustness to bounded shifts*: In Section 3, we show that a single proxy can be used to construct estimators with distributional robustness guarantees under bounded interventions on A. However, these estimators are robust to a strictly smaller set of interventions, compared to when A is used directly, and the size of this set depends on the (unidentifiable) noise in the proxy. When two proxies are available, we propose a modified estimator that can be used to recover the same guarantees as when A is observed.
- *Targeted shifts*: In Section 4, we show how to target our loss to interventions on A contained in a specified robustness set. We show that this formulation includes Anchor Regression as a special case, but also allows for sets that are not centered around the mean of A. In this setting we give an estimator, using two proxies, that identifies the target loss.

In Section 5, we evaluate our theoretical findings on synthetic experiments, and in Section 6 we demonstrate our method on a real-world dataset consisting of hourly pollution readings across five major cities in China.

2. Preliminaries

2.1. Notation

We use upper case letters X to denote (possibly vector-valued) random variables, and lower-case letters x to denote values in the range of those random variables. Vectors are assumed to be column vectors, so that $X \in \mathbb{R}^{d_X}$ indicates that $X = (X_1, \ldots, X_{d_X})^\top$, a column vector of d_X random variables. We use $\Sigma_X \in \mathbb{R}^{d_X \times d_X}$ to denote the covariance matrix of a variable X. We use bold upper-case letters $\mathbf{X}$ to denote a data matrix in $\mathbb{R}^{n \times d_X}$, consisting of n i.i.d. observations of X, and $\mathbf{1}\{\cdot\}$ as an indicator random variable. When dealing with matrices C, D, we use $C \prec D$ and $C \preceq D$ to indicate the positive definite and positive semi-definite partial order, respectively. That is, $C \prec D$ if $D - C$ is positive definite (PD), and $C \preceq D$ if $D - C$ is positive semi-definite (PSD). We use Id to denote the identity matrix, whose dimension is given by context. All proofs are provided in the supplementary material.

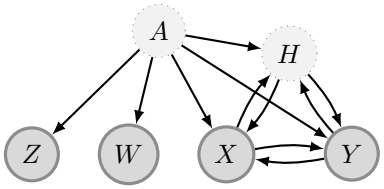

Figure 2. In contrast to Rothenhäusler et al. (2021), we assume that anchor variables (denoted A) are unobserved, but that we have access to either one or two proxies W, Z. Observed variables are shown in dark grey and unobserved variables in light grey. We do not assume knowledge of the causal structure between A, X, H, Y (except that A has no causal parents). The relationship between X, H, Y could be cyclic, but all relationships are linear.

2.2. Linear structural causal model

We assume the general class of causal graphs represented in Figure 2, where $X \in \mathbb{R}^{d_X}$ denotes observed covariates that can be used in prediction, $Y \in \mathbb{R}^{d_Y}$ is the target we seek to predict, $H \in \mathbb{R}^{d_H}$ are unobserved variables, and $A \in \mathbb{R}^{d_A}$ represents anchor variables, which are assumed to have no causal parents in the graph. We assume the linear structural causal model (SCM) given in Assumption 1.

Assumption 1 (Linear SCM)**.** *We assume the SCM*

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \coloneqq B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + M_A A + \epsilon, \tag{1}$$

where A, ϵ have zero mean, bounded covariance, and are independently distributed. We assume that $\mathbb{E}[AA^\top]$ and $\mathrm{Id} - B$ are invertible, where Id is the identity matrix. See Figure 2 for a graphical representation.

Note that we do not assume here (or anywhere in this paper) that either A or ϵ is Gaussian. The invertibility of $\mathrm{Id} - B$ is satisfied if the causal graph is a directed acyclic graph. The matrices B, M_A encode the linear causal relationships. For instance, Figure 1 can be represented in this form by $B = \begin{bmatrix} 0 & 0 \\ \alpha & 0 \end{bmatrix}$, $M = \begin{bmatrix} \beta_X \\ \beta_Y \end{bmatrix}$. In general, $\epsilon \in \mathbb{R}^D$, $B \in \mathbb{R}^{D \times D}$, and $M \in \mathbb{R}^{D \times d_A}$, where $D \coloneqq d_X + d_Y + d_H$. We assume that $d_Y = 1$ for simplicity.

2.3. Distributional robustness of anchor regression

Our goal is to learn a predictor $f^*(X)$ of Y that minimizes a worst-case risk of the following form

$$f^* = \arg\min_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Y, f(X))], \tag{2}$$

where $\mathcal{F}$ denotes a hypothesis class of possible predictors, $\mathcal{P}$ denotes a set of possible distributions, and ℓ represents our loss function. We take the class $\mathcal{P}$ to consist of distributions that arise as the result of causal interventions on A, and seek to learn a linear predictor to minimize mean-squared error.

We use $\mathbb{P}$ to refer to the observational distribution, and $\mathbb{P}_{do(A:=\nu)}$ to refer to the distribution under interventions on A, where the variable A is replaced by the random variable ν, and ν is assumed to be independent of the noise vector ϵ. We often write

$$R(\gamma) := Y - \gamma^\top X$$

as a random variable that represents the residual of a predictor $\gamma \in \mathbb{R}^{d_X}$. Importantly, Assumption 1 implies that for any γ, $\mathbb{E}[R(\gamma) \mid A]$ can be written as a linear function in A.

In this setting, Rothenhäusler et al. (2021) propose the following objective, defined here with respect to the observational distribution $\mathbb{P}$ (rather than a finite sample)

Definition 1 (Anchor Regression)**.**

$$\ell_{AR}(A;\gamma,\lambda) := \ell_{LS}(X,Y;\gamma) + \lambda \ell_{PLS}(X,Y,A;\gamma), \quad (3)$$

where $\lambda \geq -1$ is a hyperparameter and

$$\ell_{LS}(X,Y;\gamma) := \mathbb{E}\left[R(\gamma)^2\right] \quad (4)$$

$$\ell_{PLS}(X,Y,A;\gamma) := \mathbb{E}\left[\left(\mathbb{E}\left[R(\gamma) \mid A\right]\right)^2\right]. \quad (5)$$

The first term ℓ_{LS} encodes the least-squares objective, while the second term ℓ_{PLS} encodes the residual error which can be predicted from A, which we refer to as the projected least-squares error. For $\lambda > 0$, the second term adds an additional penalty (beyond that of ordinary least squares) when the bias varies across values of A. The second term (5) can also be written in the linear setting of Assumption 1 as

$$\ell_{PLS}(A;\gamma) = \mathbb{E}[R(\gamma)A^\top]\mathbb{E}[AA^\top]^{-1}\mathbb{E}[AR(\gamma)^\top], \quad (6)$$

where we drop the dependence on X, Y for notational simplicity. Under Assumption 1, Equation (3) corresponds to a worst-case loss under distributional shift caused by bounded intervention on A (Rothenhäusler et al., 2021, Theorem 1)

$$\ell_{AR}(A;\gamma,\lambda) = \sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2], \quad (7)$$

where the robustness set is given by

$$C_A(\lambda) := \{\nu : \mathbb{E}[\nu\nu^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]\}. \quad (8)$$

Since minimizing ℓ_{AR} is equivalent to ordinary least squares (OLS) regression when $\lambda = 0$, this also provides a natural robustness guarantee for the OLS estimator, where $C_{OLS} := \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top]\}$. In an identifiable instrumental variable setting, the minimizer converges against the causal parameter for $\lambda \to \infty$ (e.g. Jakobsen & Peters, 2020, eq. (71)); the ℓ_{PLS} term has therefore been referred to as 'causal regularization' (e.g. Bühlmann & Ćevid, 2020), and has also been denoted by ℓ_{IV} (Rothenhäusler et al., 2021), as $\mathrm{Cov}(A, R(\gamma)) = \mathbf{0}$ if and only if $\ell_{PLS}(\gamma) = 0$.

3. Distributional robustness to bounded shifts

We first assume the existence of a noisy proxy W, conditionally independent of (X, Y, H) given A (see Figure 2).

Assumption 2 (Single proxy with additive noise)**.** *In the context of Assumption 1, W is generated as follows*

$$W := \beta_W^\top A + \epsilon_W,$$

where ϵ_W has mean zero, bounded covariance, and is independent of (A, ϵ). In addition, we assume that the second moment matrix $\mathbb{E}[WW^\top]$ is invertible.

Under mild identifiability conditions (e.g., that β_W is full rank) one can show (see Section C.2) that

$$\ell_{PLS}(A;\gamma) = 0 \iff \ell_{PLS}(W;\gamma) = 0, \quad (9)$$

Hence, a single proxy is enough (in the population case) to identify whether the sharp constraint $\ell_{PLS}(\gamma) = 0$ holds, representing invariance to interventions of arbitrary strength. This corresponds to the fact that if A is a valid instrumental variable, then so is W (Hernán & Robins, 2006).

However, we consider interventions on A that are not of arbitrarily large strength. With that in mind, in Section 3.1, we demonstrate that (i) when a single proxy W is used in place of A, a robustness guarantee holds, but the robustness set is reduced relative to (8), (ii) the extent of this reduction depends on the signal-to-variance relationship in W, and (iii) this relationship is not generally identifiable from the observational distribution over (X, Y, W) alone. In Section 3.2, we show that in the setting where two proxies are available, the same guarantees as for an observed A can be obtained. We do so constructively, giving a regularization term whose population version is equal to $\ell_{PLS}(A;\gamma)$.

3.1. Robustness with a single proxy

First, we establish the robustness set of Anchor Regression when a single proxy is used in place of A. We refer to this as Proxy Anchor Regression, to distinguish it from the case when A is observed, but the only difference from Definition 1 is that W is used in place of A.

Definition 2 (Proxy Anchor Regression)**.** Let ℓ_{LS}, ℓ_{PLS} be defined as in Equations (4) and (6). We define

$$\ell_{PAR}(W;\gamma,\lambda) := \ell_{LS}(\gamma) + \lambda \ell_{PLS}(W;\gamma), \quad (10)$$

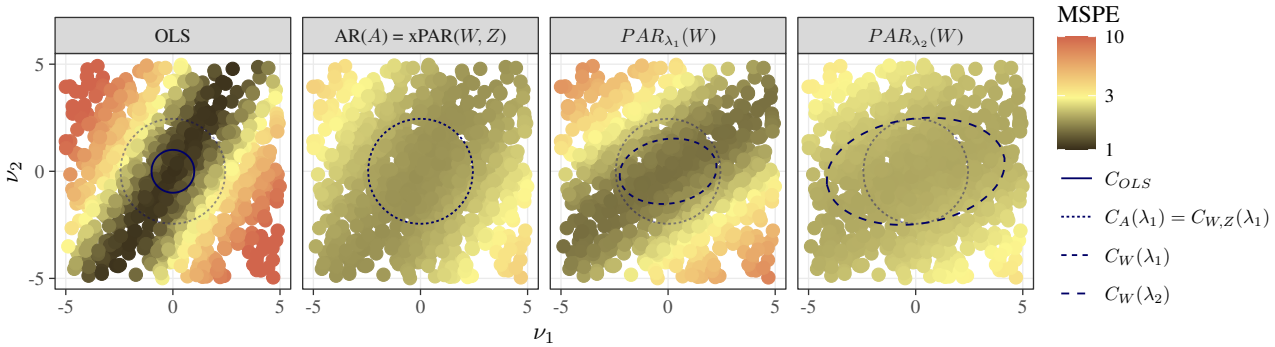

Figure 3. Test performance under interventions $do(A := (\nu_1, \nu_2))$ which give rise to different test distributions over X and Y. Each dot corresponds to a different intervention (i.e., test distribution on X, Y), and the color gives the resulting mean squared prediction error (MSPE). **(Far Left)** OLS performs well for interventions in the set C_{OLS} (solid circle), corresponding to the training covariance of A. However, it performs poorly under interventions far from this region (e.g., top left). **(Middle Left)** Anchor Regression (AR) minimizes the worst-case loss over interventions on A within the region $C_A(\lambda_1)$ (cf., (8)), a re-scaling of C_{OLS}. There is a trade-off, with better performance than OLS under large interventions, but worse performance under small interventions. Given two proxies W, Z, we introduce Cross-Proxy Anchor Regression (xPAR, cf., (14)) and prove that it minimizes the same worst-case loss. **(Middle Right)** When only a single proxy W is used in place of A, the result is a weaker guarantee, in the form of a smaller robustness set $C_W(\lambda_1)$ (cf., (11)) for the same value of λ_1. The shape of this set depends on the noise in the proxy along different dimensions. **(Far Right)** As a result, there does not generally exist a λ_2 such that $C_W(\lambda_2) = C_A(\lambda_1)$. If we choose some $\lambda_2 > \lambda_1$ such that $C_A(\lambda_1) \subset C_W(\lambda_2)$, we enforce a stronger constraint than intended, resulting in an unwanted trade-off between performance and robustness.

where $\lambda \geq -1$ is a hyperparameter and we suppress the dependence on X, Y in the notation.

Theorem 1. *Under Assumptions 1 and 2, for all $\gamma \in \mathbb{R}^{d_X}$ and for all $\lambda \geq -1$*

$$\ell_{PAR}(W; \gamma, \lambda) = \sup_{\nu \in C_W(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2],$$

where the robustness set is given by

$$C_W(\lambda) := \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\Omega_W\} \tag{11}$$

and where Ω_W is defined as

$$\Omega_W := \mathbb{E}[AW^\top]\big(\mathbb{E}[WW^\top]\big)^{-1}\mathbb{E}[WA^\top]. \tag{12}$$

Intuitively, Ω_W defines a signal-to-variance relationship in W, and this determines the robustness guarantee. In the case where both $A, W \in \mathbb{R}$ are one-dimensional, and A has unit variance, the robustness sets simplify to

$$\begin{aligned} C_{OLS} &= \{\nu : \mathbb{E}[\nu^2] \leq 1\} \\ C_W(\lambda) &= \{\nu : \mathbb{E}[\nu^2] \leq 1 + \lambda \cdot \rho_W\} \\ C_A(\lambda) &= \{\nu : \mathbb{E}[\nu^2] \leq 1 + \lambda\}, \end{aligned}$$

where $\rho_W := \beta_W^2/(\beta_W^2 + \mathbb{E}\epsilon_W^2) < 1$ is the signal-to-variance ratio of W, also referred to as the reliability ratio in the measurement error literature (Fuller, 1987). Thus, in the one-dimensional case, the robustness set using W is strictly smaller than the one obtained by using A when $\lambda > 0$, except in the case where $\epsilon_W = 0$ a.s. This result generalizes to higher dimensions.

Proposition 1. *Assume Assumptions 1 and 2 and that $\mathbb{E}[\epsilon_W \epsilon_W^\top] \in \mathbb{R}^{d_W \times d_W}$ is positive definite. Then for $\lambda > 0$*

$$C_{OLS} \subseteq C_W(\lambda) \subset C_A(\lambda),$$

and the set $C_W(\lambda)$ increases monotonically when $\mathbb{E}[\epsilon_W \epsilon_W^\top]$ decreases w.r.t. the partial matrix ordering. If $d_W = d_A$, β_W is full rank, and $\epsilon_W = 0$ a.s., then $C_W(\lambda) = C_A(\lambda)$.

If Ω_W were known, we could choose a larger λ^* such that $C_A(\lambda) \subseteq C_W(\lambda^*)$. In contrast to the one-dimensional case, where we could choose $\lambda^* = \lambda/\rho_W$ to obtain an equality $C_A(\lambda) = C_W(\lambda^*)$, we cannot generally achieve equality in higher dimensions (see Figure 3).

However, Ω_W is not generally identifiable from the observed distribution over (X, Y, W) alone. Moreover, SCMs compatible with the observed distribution react differently under interventions on A and yield different coefficients that are optimal w.r.t. interventions in $C_A(\lambda)$. Consequently, in this setting, it is not possible to recover the guarantees of Anchor Regression without further assumptions (e.g., on Ω_W). See Supplement B for an example.

Note that these results apply regardless of whether or not β_W is full rank. However, if β_W is not full rank, then there will be directions of variation in A that are not reflected in W, and we will not be able to achieve additional robustness (beyond that of OLS) against interventions along these directions.

3.2. Robustness with two proxies

We now show that if we have two (sufficiently different) proxies for A, then it is possible to recover the original robustness set using a different regularization term. We denote these proxies by W, Z, as shown in Figure 2. In this setting, the structural causal model over (X, Y, H, A) can still be written in the form of Equation (1), where we make the following additional assumptions.

Assumption 3 (Proxies with additive noise)**.** *In the context of Assumption 1, Z, W are generated as follows*

$$W := \beta_W^\top A + \epsilon_W \qquad \text{and} \qquad Z := \beta_Z^\top A + \epsilon_Z,$$

where ϵ_W, ϵ_Z are mean-zero with bounded covariance, and $\epsilon_W, \epsilon_Z, \epsilon, A$ are jointly independent.

Assumption 4. *The dimensions of A, W, Z are equal, $d_A = d_W = d_Z$, and β_W, β_Z are full-rank.*

Note that Assumption 4 also implies that the second moment matrix $\mathbb{E}[ZW^\top]$ is invertible.

To build intuition, note that this assumption is trivially satisfied in the setting where $W = A + \epsilon_W$ and $Z = A + \epsilon_Z$, i.e., where W and Z are two noisy observations of A. More generally, Assumption 4 rules out directions of variation in A that are undetectable in W or Z.

In this setting we introduce the following loss, and prove that it is equal to the worst-case loss obtained when A is observed (c.f., (7))

Definition 3 (Cross-Proxy Anchor Regression)**.**

$$\ell_{\times PAR}(W, Z; \gamma, \lambda) := \ell_{LS}(X, Y; \gamma) + \lambda \ell_\times(W, Z; \gamma),$$

where we refer to

$$\ell_\times(W, Z; \gamma) := \mathbb{E}[R(\gamma)W^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZR(\gamma)^\top], \tag{13}$$

as the cross-proxy regularization term.

Theorem 2. *Under Assumptions 1, 3 and 4, for any $\gamma \in \mathbb{R}^{d_X}$ and any $\lambda \geq -1$*

$$\ell_{\times PAR}(W, Z; \gamma, \lambda) = \sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2], \tag{14}$$

where $C_A(\lambda) = \{\nu : \mathbb{E}[\nu\nu^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]\}$.

$\ell_{\times PAR}$ is convex in γ and has a closed form solution for its minimizer based only on the population moments of X, Y, W and Z (see Proposition A4 in the supplement).

To build intuition for why Assumption 4 is required for this result, consider an example where W, Z are both scalars ($d_W = d_Z = 1$) and A has two independent dimensions (A_1, A_2). In this example, if both proxies measure the same dimension A_1, then variation in A_2 is not detectable in either proxy, and we cannot optimize for robustness to interventions on A_2. On the other hand, if W only measures A_1 (e.g., $W = A_1 + \epsilon_W$), and Z only measures A_2 (e.g., $Z = A_2 + \epsilon_Z$), then we cannot use Z to identify the signal-to-variance ratio of W, and vice-versa. In this case, (W, Z) is effectively a single two-dimensional proxy in the framework of Section 3.1, where we showed that recovering the guarantees of Anchor Regression is not generally possible. Intuitively, we need all directions of variation in A to have some influence on both proxies (i.e., β_W, β_Z full rank), and hence require that W, Z have sufficiently large dimension.

4. Targeted anchor regression: Incorporating additional shift information

We now generalize Anchor Regression to an estimator that is targeted to be robust against particular shifts, and demonstrate that we can similarly handle this setting when only proxies of A are observed. In Section 2.3 we showed that Anchor Regression minimizes the worst-case loss over the set $C_A(\lambda)$ of all interventions $do(A := \nu)$ where $\mathbb{E}[\nu\nu^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]$. For deterministic ν, $C_A(\lambda)$ is an ellipsoid centered at 0, and its width in each direction is proportional to the variation of A in that direction. However, we may desire a different robustness set: For instance, if we anticipate a particular shift μ_ν in the mean of A, or if we want to add extra protection against particular directions of variation in A. This can be formalized as a robustness set defined by an ellipsoid that may not be centered at 0, nor be proportional to $\mathbb{E}[AA^\top]$. The estimator developed in this section can incorporate such prior beliefs.

More formally, instead of considering robustness against interventions $do(A := \nu)$ over the set $\nu \in C_A(\lambda)$, we now assume that we have additional information on the nature of ν, which is specified in the form of a vector μ_ν and a symmetric PSD matrix Σ_ν. We introduce a new method, Targeted Anchor Regression, minimizing what we refer to as the *targeted loss*. We prove in Propositions 2 and 3 that minimizing this objective can be interpreted in two ways: First, as minimizing an expected loss over interventions ν with a known mean and covariance, or minimizing a worst-case loss over deterministic interventions ν contained in an ellipsoid robustness set (as discussed above). This is visualized in Figure 4.

4.1. Targeting when A is observed

We first consider the case when A is observed during training, and the mean and covariance of ν are known, given by μ_ν, Σ_ν. Importantly, for a given γ we have $\mathbb{E}[R(\gamma) \mid A = a] = b_\gamma^\top a$, where, writing $\Sigma_A := \mathbb{E}[AA^\top]$,

$$b_\gamma^\top := \mathbb{E}[R(\gamma)A^\top]\Sigma_A^{-1}. \tag{15}$$

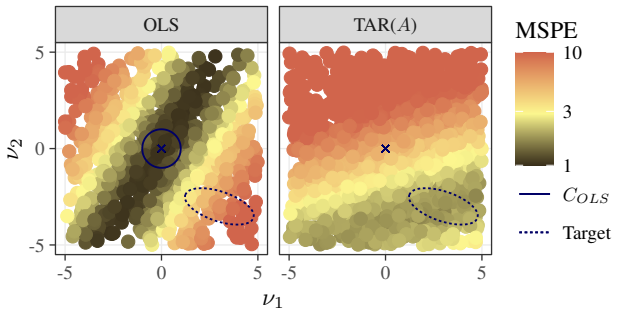

Figure 4. Targeted Anchor Regression allows for minimizing the worst-case loss in regions (dashed ellipse) that may differ in location, size, and shape from the regions in Figure 3 (OLS copied for reference). Every point ν represents a test distribution $do(A := \nu)$, the color indicating the mean squared prediction error in this distribution. Cross marks the origin. The TAR estimator achieves its minimal test loss at the center of the targeted region.

Definition 4 (Targeted Anchor Regression)**.** Let $\mu_\nu \in \mathbb{R}^{d_A}$, and $\Sigma_\nu \in \mathbb{R}^{d_A \times d_A}$, where Σ_ν is a symmetric PSD matrix.

$$\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha) \\ := \ell_{LS}(\gamma) + b_\gamma^\top (\Sigma_\nu - \Sigma_A) b_\gamma + (b_\gamma^\top \mu_\nu - \alpha)^2, \quad (16)$$

where b_γ is defined in (15), and Σ_A is the covariance of A.

Proposition 2. *Under Assumption 1, and the assumption that $\nu \perp\!\!\!\perp \epsilon$, we have, for all $\gamma \in \mathbb{R}^{d_X}, \alpha \in \mathbb{R}$,*

$$\ell_{TAR}(A; \mu_\nu, \Sigma_\nu; \gamma, \alpha) = \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2],$$

where $\mu_\nu = \mathbb{E}[\nu]$ and Σ_ν is the covariance matrix of ν.

Importantly, the objective in Equation (16) is convex in (γ, α), and has a closed-form solution (see Proposition A5 in the supplement). If ν is a known constant, then this corresponds to performing OLS using both X and A as predictors during training, and using the known value of ν for A for prediction (see Supplement C.3.2). However, if for example ν exhibits more variance than A along certain directions, and less variance along others, then the targeted regression parameter differs from standard solutions. Optimizing the objective in Equation (16) can also be interpreted as optimizing a worst-case loss over interventions $do(A := \nu)$ in a certain set.

Proposition 3. *Under Assumption 1, we have, for all $\mu_\nu \in \mathbb{R}^{d_A}$ and $\Sigma_\nu \in \mathbb{R}^{d_A \times d_A}$ being a symmetric positive definite matrix, that*

$$\arg\min_{\gamma,\alpha} \ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha) \\ = \arg\min_{\gamma,\alpha} \sup_{\nu \in T(\mu_\mu, \Sigma_\nu)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2],$$

where the supremum is taken over (deterministic or random) shifts ν of the form $\nu = \mu_v + \delta$, where δ satisfies the constraint that $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$. If δ is random, we require that it is independent of all other random variables. In other words, we can write that ν lies in the set

$$T(\mu_\nu, \Sigma_\nu) := \{\nu : \mathbb{E}[(\nu - \mu_\nu)(\nu - \mu_\nu)^\top] \preceq \Sigma_\nu\}.$$

Note that the expectation in the constraint T is with respect to the random variable ν. This covers the case in which ν (and hence δ) is deterministic, in which case it is equal to a fixed value with probability one.

Proposition 3 shows that Targeted Anchor Regression generalizes Anchor Regression to a broader class of robustness sets, that need not depend explicitly on $\mathbb{E}[AA^\top]$. In particular, Anchor Regression can be viewed as a special case, where $\Sigma_\nu = (1 + \lambda)\Sigma_A$ and $\mathbb{E}[\nu] = 0$, in which case the objectives are equal for $\alpha = 0$. In the following, we adopt the interpretation of μ_ν, Σ_ν as specifying a mean and covariance of ν (Proposition 2).

4.2. Targeting with proxies

In the single-proxy setting, we define Proxy Targeted Anchor Regression as using W in place of A in Equation (16). We assume a known mean and covariance of W under $\mathbb{P}_{do(A:=\nu)}$, used in place of μ_ν, Σ_ν. By similar arguments to those in Section 3.1, this approach does not generally yield the optimal predictor, in a way that depends on the (unidentified) signal-to-variance relationship in W. Given the similarity, we defer details to Supplement D.

When two proxies W, Z are available, we can recover the statement from Proposition 2 using a modified estimator, by similar arguments to those in Section 3.2. The core observation is that we can construct a linear term

$$a_\gamma^\top := \mathbb{E}[R(\gamma)Z^\top](\mathbb{E}[WZ^\top])^{-1}, \quad (17)$$

which, if $\beta_Z = \beta_W = \mathrm{Id}$ can be seen as a linear IV estimate of $b_\gamma^\top$ in Equation (15), an estimator used in the measurement error literature given repeated noisy measurements of a single variable (Fuller, 1987). In our case, Equation (17) identifies $b_\gamma^\top$ only up to the linear transformation β_W, but this is sufficient to identify the targeted loss.

Definition 5 (Cross-Proxy Targeted Anchor Regression)**.** Let $\tilde{\mu} \in \mathbb{R}^{d_W}$, and $\tilde{\Sigma}_W \in \mathbb{R}^{d_W \times d_W}$, where $\tilde{\Sigma}_W$ is a symmetric positive semi-definite matrix. We define

$$\ell_{\times TAR}(W, Z; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha) \\ := \ell_{LS}(\gamma) + a_\gamma^\top \left(\tilde{\Sigma}_W - \Sigma_W\right) a_\gamma + \left(a_\gamma^\top \tilde{\mu} - \alpha\right)^2,$$

where a_γ is defined in (17).

In Theorem 3 (Supplement D) we prove, analogous to Theorem 2, that this population objective is equal to that of Targeted Anchor Regression (16).

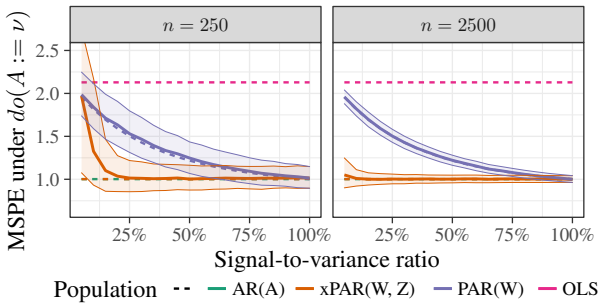

Figure 5. Mean squared prediction error (MSPE) under interventions $do(A := \nu)$ for estimators PAR and xPAR. We display population losses for the population parameters as dashed lines, and median empirical MSPE when fit from data as solid lines, with shaded regions covering the 25% to 75% quantiles.

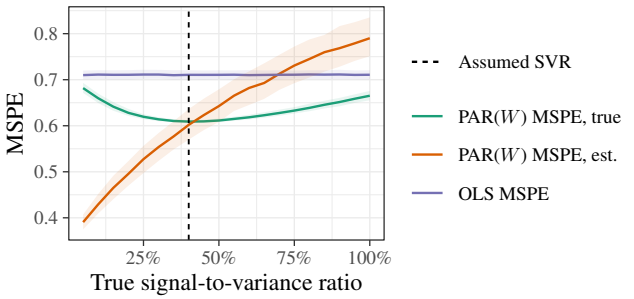

Figure 6. Estimates of worst-case mean squared prediction error (MSPE) over a robustness set C. PAR is applied assuming that the signal-to-variance ratio is 0.4, which gives an estimate of the worst-case MSPE over C (orange). Green line shows actual worst-case MSPE over C at different underlying signal-to-variance ratios.

5. Synthetic experiments

In Section 5.1, we show that Cross-Proxy Anchor Regression (xPAR) outperforms Proxy Anchor Regression (PAR) in settings with noisy proxies. As the noise increases, xPAR continues to match Anchor Regression (AR) test performance under intervention, while PAR approaches OLS. In Section 5.2, we demonstrate the risks of attempting to correct for this noise by assuming a certain signal-to-variance ratio. In Section 5.3 we demonstrate another benefit of xPAR over PAR, giving an example where it places more weight on causal predictors relative to PAR. Finally, in Section 5.4, we highlight the trade-off between using Targeted Anchor Regression (TAR) vs. OLS and AR, showing that TAR improves performance under the targeted shift, at the cost of incurring additional error on the training distribution. Code for experiments is available at `https://github.com/clinicalml/proxy-anchor-regression`.

5.1. Mean squared prediction error under intervention

We demonstrate on synthetic data that xPAR recovers similar test performance to AR, while the performance of PAR degrades as the signal-to-variance ratio (SVR) of the proxies decreases. We simulate training data (at different levels of signal-to-variance) from an SCM with the structure given in Figure 2, fix $\lambda := 5$ and fit PAR and xPAR. We then choose a fixed intervention ν, and simulate test data under the intervened distribution, evaluating our learned predictors.

In Figure 5, we see that the test errors for xPAR and AR coincide (see Theorem 2) while PAR interpolates between OLS and AR, depending on the signal-to-variance ratio (see Proposition 1). Section E gives additional implementation details on this and remaining experiments.

5.2. Misspecified signal-to-variance ratio

In Section 3.1, we noted that if the (unidentified) signal-to-variance ratio (SVR) were known, we could correct for it when using PAR with a single proxy. Here we demonstrate the implications of incorrectly specifying this correction. We simulate data from the same SCM as in Section 5.1, with varying (true) signal-to-variance ratio.

In Figure 6, for the predictor chosen by PAR, we plot the estimated worst-case MSPE (in orange), using a correction factor assuming that the signal-to-variance ratio is 0.4, against the true worst-case MPSE (in green). We observe that if the true signal-to-variance ratio is smaller than our assumption of 0.4, then our estimate is too conservative, and vice versa if the true signal-to-variance ratio is larger.

5.3. Causal and anti-causal predictors

We demonstrate the ability of xPAR to select causal predictors, in a synthetic setting where predictors X may contain both causal and anti-causal predictors. We simulate data from an SCM (Figure 7 [top]), where one anchor, A_1, is a parent of the causal predictors, while the other, A_2, is a parent of the anti-causal predictors. We consider two identically distributed noisy proxies W, Z of $A := (A_1, A_2)$. The challenge is that A_2 is measured with significantly more noise than A_1, across both proxies.

As seen in Figure 7 [bottom] PAR places more weight on anti-causal features. In effect, the noise in the measurement of A_2 causes $X_{\text{anti-causal}}$ to appear less sensitive to shifts in A_2. This is an ideal scenario for xPAR, as it is designed to deal with additional noise by leveraging both proxies. Consequently, when two proxies W, Z are available, xPAR places more weight on the causal predictors, relative to PAR.

5.4. Targeted shift

We demonstrate the trade-off made by Targeted Anchor Regression (TAR) versus Anchor Regression (AR), considering the case when A is observed for simplicity. We simulate training data and fit estimators γ_{OLS}, γ_{AR} and γ_{TAR},

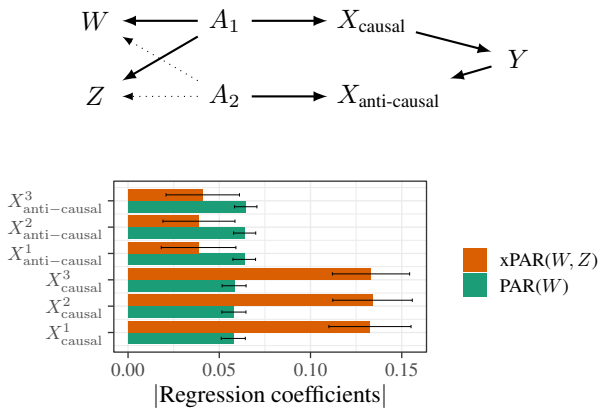

Figure 7. Top: SCM with A_1, A_2 (unobserved), target Y and predictor variables $X_{\text{causal}}, X_{\text{anti-causal}} \in \mathbb{R}^3$. Dotted lines indicate higher noise. *Bottom*: Absolute value of regression coefficients. PAR places more weight on anti-causal predictors, while xPAR places more weight on causal predictors.

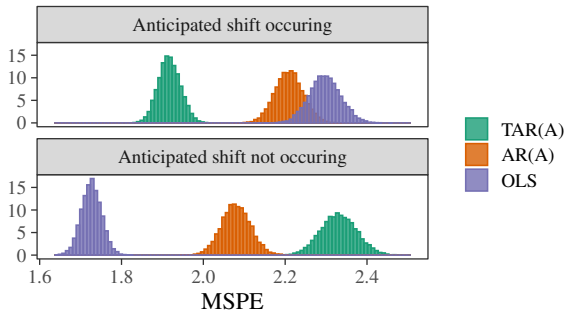

Figure 8. Empirical mean squared prediction error of TAR, OLS and AR under the shifted distribution and the training distribution.

where γ_{TAR} is targeted to a particular mean and covariance of a random intervention ν, and we select λ for γ_{AR} such that this intervention is contained within $C_A(\lambda)$.

We then simulate test data from two distributions: $\mathbb{P}_{do(A:=\nu)}$ (i.e., the shift occurs), and $\mathbb{P}$ (where it does not), and evaluate the mean squared prediction error (MSPE). The results are shown in Figure 8, and demonstrated that TAR performs better than AR and OLS in the first scenario, but this comes at the cost of worse performance on the training distribution.

6. Real-data experiment: Pollution

We test our approach on a real-world heterogeneous dataset of hourly pollution readings in five cities in China, taken over several years (Liang et al., 2016), with most data available from 2013-15. Our prediction target is PM2.5 concentration, a measure of pollution, and covariates are primarily weather-related, including dew point, temperature, humidity, pressure, wind direction / speed, and precipitation.

Real-World Proxy (Temperature): Pollution tends to be seasonal in this dataset, and so we construct our training and test environments using seasons: For each of the four seasons, we train only on the other three seasons, and evaluate on the held-out season. We do this for each city, treating each city and held-out season as a separate evaluation. This leads to 20 separate scenarios.

With this variation in mind, we use temperature as a real-world proxy, and treat it as unavailable at test time. We also construct two noisier copies of temperature, which we refer to as W, Z, adding independent Gaussian noise while controlling the signal-to-variance ratio (in the training distribution) at $\text{Var}(\text{Temp})/\text{Var}(W) = 0.9$.

Estimators / Benchmarks: For Proxy and Cross-Proxy AR (PAR, xPAR, see Section 3), we choose $\lambda \in [0, 40]$ by leave-one-group-out cross-validation on the three training seasons, using the first year (2013) of data. For instance, if "winter" is the test season, then we choose the value of λ that performs best on average across combinations of the other seasons e.g., training on the fall & summer data and evaluating on the spring data.

When using temperature as a single proxy in PAR, we observe that in 9 out of 20 scenarios, $\lambda = 40$ is chosen, but in the remaining 11, $\lambda = 0$ is chosen, which is equivalent to OLS. For comparability, we use the same values of λ for PAR(W) and xPAR(W, Z). For Proxy Targeted AR and Cross-Proxy Targeted AR (PTAR, xPTAR, see Section 4), we use the mean and variance of the relevant variables (e.g., temperature, W, Z) in the held-out season to target our predictors.

Our primary benchmark is OLS (without temperature). We also compare to (a) OLS that uses temperature during train and test [OLS (TempC)], and (b) OLS that includes the temperature during training, and uses the mean test value for temperature during prediction [OLS + Est. Bias]. We present the results for the 9 scenarios where $\lambda > 0$ in Table 1, since PAR with $\lambda = 0$ is equivalent to OLS (aggregate results in Table 2 in the supplement).

Results: For both PAR and PTAR, we see improvement over OLS on average across scenarios, with limited downside (e.g., in the worst scenario for PTAR relative to OLS, the additional MSE incurred is 0.001). In Figure 12 (Supplement), we observe that PAR and PTAR achieve gains in two different ways: PAR increases the coefficients of humidity and dew point relative to OLS, while PTAR reduces them and incorporates a correction into the intercept.

7. Discussion and related work

Learning a predictive model that performs well under arbitrarily strong causal interventions is an ambitious goal. In this work, we have argued that even if causal invariance is

Table 1. Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536	5	-0.028	0.026
OLS + Est. Bias	0.569	4	-0.072	0.150
PAR (TempC)	0.531	6	-0.041	0.006
PAR (W)	0.531	6	-0.037	0.006
xPAR (W, Z)	0.531	6	-0.039	0.007
PTAR (TempC)	0.525	8	-0.061	0.001
PTAR (W)	0.529	8	-0.038	0.001
xPTAR (W, Z)	0.526	7	-0.059	0.001

achievable, it may not be desirable: A model whose performance is invariant to arbitrarily strong interventions may have poor performance when the test distribution does not differ too much from the training distribution.

There is a large body of work that seeks to learn causal models as a route to achieving invariance (Rojas-Carulla et al., 2018; Magliacane et al., 2018), or that uses knowledge of the causal graph to select predictors with invariant performance under a set of known interventions (Subbaswamy et al., 2019). Similarly, invariant risk minimization (IRM) seeks a predictor Φ such that $\mathbb{E}(Y \mid \Phi(X))$ is invariant across a set of discrete environments (Arjovsky et al., 2019; Xie et al., 2020; Krueger et al., 2020; Bellot & van der Schaar, 2020). Recent work has pointed to the theoretical and practical difficulty of learning such a predictor for IRM (Rosenfeld & Risteski, 2020; Kamath et al., 2021; Guo et al., 2021), in part due to the fact that recovering a truly invariant model, even in linear settings, requires a large number of environments. Generalization in non-linear settings requires sufficient overlap between environments and strong restrictions on the model class (e.g., Christiansen et al., 2020). In contrast to all of the above, we trade off between in-distribution performance and invariance explicitly, instead of seeking invariance as a primary goal. Moreover, since we allow for A to influence Y directly and through hidden variables, invariance may not even be achievable, but we can still formulate a worst-case loss for bounded interventions.

We argue for incorporating prior knowledge about potential shifts by (1) identifying proxies for relevant factors of variation (i.e., anchor variables), and (2) specifying plausible sets of interventions on these factors of variation. We build upon the causal framework of Anchor Regression (Rothenhäusler et al., 2021), extending it in two important ways.

To start, we relax the assumption that the anchor variables are directly observed. Instead, we only assume access to proxies, and prove that identification of the worst-case loss is feasible with two proxies. The challenge of identifying the worst-case loss is related to the problem of identifying causal effects with noisy proxies of unmeasured confounders (Tchetgen Tchetgen et al., 2020; Miao & Tchetgen, 2018; Shi et al., 2018; Kuroki & Pearl, 2014), and the challenge of learning under classical measurement error (Fuller, 1987; Hyslop & Imbens, 2001; Bound et al., 2001). Our observation that a single proxy will underestimate the worst-case loss is related to the well-known problem of regression dilution bias (Frost & Thompson, 2000), where performing linear regression under measurement error leads to bias in parameter estimation. In contrast, we are not concerned with causal / structural parameter estimation, which is generally not possible in the models we consider, but rather estimating a worst-case loss under a class of interventions. Srivastava et al. (2020) also consider distributional shift in unmeasured variables for which proxies are available, and apply techniques for handling worst-case sub-populations from DRO (Duchi et al., 2020). In contrast, we consider causal interventions on A that could lie outside the support of the training data, which cannot be represented as a sub-population. Moreover, they consider the single-proxy case, and give a generalization bound that incorporates the impact of noise, while under our assumptions we are able to recover guarantees as if A were observed, using two proxies.

We then introduce Targeted Anchor Regression, a method for incorporating additional prior knowledge on the strength and direction of shifts in anchor variables. This method can be interpreted as allowing for specification of a broader class of robustness sets, beyond those considered in Rothenhäusler et al. (2021), or as specifying the mean and covariance of the anchors at test time. We prove analogous results with proxies in this setting, and evaluate this strategy empirically in Section 6, targeting our loss to a particular mean and variance over temperature in the held-out season.

Our work contributes to a growing body of literature that seeks to generalize Anchor Regression to new settings, whether allowing for unobserved anchors and a broader class of robustness sets (as in our work), or generalizing to discrete and censored outcomes, as in Kook et al. (2021).

Acknowledgements

We thank Hussein Mozannar, Chandler Squires, Hunter Lang, Zeshan Hussain, and other members of the ClinicalML lab for feedback and insightful discussions. This work was supported in part by Office of Naval Research Award No. N00014-17-1-2791. NT and JP are supported by a research grant (18968) from VILLUM FONDEN, and JP, in addition, is supported by Carlsberg Foundation.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. arXiv (1907.02893), 2019.

Bellot, A. and van der Schaar, M. Accounting for Unobserved Confounding in Domain Generalization. arXiv (2007.10653), July 2020.

Bound, J., Brown, C., and Mathiowetz, N. Chapter 59: Measurement Error In Survey Data. Handbook of Econometrics, 5:3705–3843, 2001.

Bühlmann, P. and Ćevid, D. Deconfounding and causal regularisation for stability and external validity. International Statistical Review, 88(S1):S114–S134, 2020.

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. The Difficult Task of Distribution Generalization in Nonlinear Models. arXiv, pp. 1–48, 2020. URL http://arxiv.org/abs/2006.07433.

Duchi, J. C., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures. arXiv (2007.13982), pp. 1–39, 2020.

Frost, C. and Thompson, S. G. Correcting for Regression Dilution Bias: Comparison of Methods for a Single Predictor Variable. Journal of the Royal Statistical Society: Series A, 163(2):173–189, 2000.

Fuller, W. A. Measurement error models. John Wiley and Sons Inc., 1987.

Guo, R., Zhang, P., Liu, H., and Kiciman, E. Out-of-distribution Prediction with Invariant Risk Minimization: The Limitation and An Effective Fix. arXiv (2101.07732), January 2021.

Hernán, M. A. and Robins, J. M. Instruments for causal inference: An epidemiologist's dream? Epidemiology, 17(4):360–372, 2006.

Hyslop, D. R. and Imbens, G. W. Bias from classical and other forms of measurement error. Journal of Business and Economic Statistics, 19(4):475–481, 2001.

Jakobsen, M. E. and Peters, J. Distributional Robustness of K-class Estimators and the PULSE. arXiv (2005.03353), 2020.

Kamath, P., Tangella, A., Sutherland, D. J., and Srebro, N. Does Invariant Risk Minimization Capture Invariance? arXiv (2101.01134), January 2021.

Kook, L., Sick, B., and Bühlmann, P. Distributional Anchor Regression. arXiv (2101.08224), January 2021.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-Distribution Generalization via Risk Extrapolation (REx). arXiv (2003.00688), March 2020.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. Biometrika, 101(2):423–437, 2014.

Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Lovell, M. C. Seasonal adjustment of economic time series and multiple regression analysis. Journal of the American Statistical Association, 58(304):993–1010, 1963.

Lovell, M. C. A simple proof of the FWL theorem. Journal of Economic Education, 39(1):88–91, 2008.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS), 2018.

Miao, W. and Tchetgen, E. T. A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects. arXiv (1808.04945), 2018.

Pearl, J. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2nd edition, 2009.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (eds.). Dataset Shift in Machine Learning. MIT Press, 2009.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant Models for Causal Transfer Learning. Journal of Machine Learning Research, 19(36):1–34, 2018.

Rosenfeld, E. and Risteski, A. The Risks of Invariant Risk Minimization. arXiv (2010.05761v1), pp. 1–36, 2020.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen, E. J. T. Multiply Robust Causal Inference with Double Negative Control Adjustment for Categorical Unmeasured Confounding. arXiv (1808.04906), 2018.

Srivastava, M., Hashimoto, T., and Liang, P. Robustness to Spurious Correlations via Human Annotations. 37th International Conference on Machine Learning, 2020.

Subbaswamy, A., Schulam, P., and Saria, S. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. An Introduction to Proximal Causal Learning. arXiv (2009.10982), 2020.

Xie, C., Ye, H., Chen, F., Liu, Y., Sun, R., and Li, Z. Risk Variance Penalization. arXiv (2006.07544), June 2020.

Zhang, F. The Schur complement and its applications, volume 4. Springer Science & Business Media, 2006.