Supplementary Materials

The supplementary materials are organized as follows

- (Section A): First, we give a simple 1D example to build intuition for the theoretical results.
- (Section B): In the context of Section 3.1, we give a concrete example to demonstrate the non-identifiability of Ω_W, defined in (12). We focus on the simple case when W is one dimensional, and the matrix Ω_W reduces to a single number $\rho_W := \beta_W^2/(\beta_W^2 + \sigma_W^2)$, indicating the signal-to-variance ratio of W. We give an example of an observed distribution for which ρ_W is not identified, and moreover, the optimal predictor with respect to the robustness set $C_A(\lambda)$ is not identified (see Figure 10).
- (Section C): Proofs for results stated in the main paper.
- (Section D): Additional results (and proofs) for Proxy Targeted Anchor Regression (PTAR) and Cross-Proxy TAR, deferred from the main paper.
- (Section E): Details for implementation of all experiments
- (Section F): Additional synthetic experimental results

A. An example for building intuition

To illustrate the problem, consider the following setup, where we observe A, X, Y at training time, and wish to learn a predictor $\hat{y} = \alpha + \gamma x$ that will generalize to a new environment where $\mathbb{P}_{te}(A) \neq \mathbb{P}_{tr}(A)$.

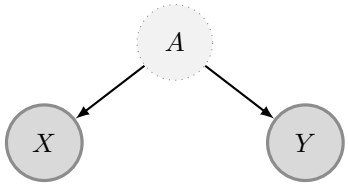


Figure 9. Simple example where $X, Y, A \in \mathbb{R}$.

Suppose that our data is generated under $\mathbb{P}_{tr}$ as follows

$$\begin{aligned} A &= \epsilon_A, & \epsilon_A &\sim \mathcal{N}(0,1) \\ X &= A + \epsilon_X, & \epsilon_X &\sim \mathcal{N}(0, \sigma_X^2) \\ Y &= A + \epsilon_Y, & \epsilon_Y &\sim \mathcal{N}(0, \sigma_Y^2), \end{aligned}$$

where $\epsilon_A, \epsilon_X, \epsilon_Y$ are jointly independent. This simple example demonstrates a few concepts:

- Assuming $\sigma_X^2 > 0$, the conditional expectation $\mathbb{E}[Y \mid X]$ changes as the distribution of A changes.
- We can write the residuals $Y - \hat{Y}$ as a linear function in A and the noise variables. This holds, even if the errors are non-Gaussian.
- The test population MSE is a convex function of α, γ.

In particular, we will see that the parameters α, γ trade off between the variance of A and ϵ_X: There exists an invariant solution, where $\alpha = 0, \gamma^* = 1$, such that the MSE is completely independent of A, but this is only optimal in the setting where $\text{Var}(A) \to \infty$.

Conditional Expectation depends on A Starting with the assumption that A, X, Y are multivariate Gaussian, we can write down the optimal predictor in the target environment, supposing that at test time $\mathbb{P}_{te}(A) \overset{(d)}{=} \mathcal{N}(\mu_A, \sigma_A^2)$.

$$\begin{aligned} \mathbb{E}_{te}[Y \mid X = x] &= \mathbb{E}_{te}[Y] + \frac{\text{Cov}_{te}(X,Y)}{\text{Var}_{te}(X)} \cdot (x - \mathbb{E}_{te}[X]) \\ &= \mu_A + \underbrace{\frac{\sigma_A^2}{\sigma_A^2 + \sigma_X^2}}_{\gamma} \cdot (x - \mu_A) \\ &= \mu_A(1-\gamma) + \gamma x, \end{aligned}$$

where if $\epsilon_X = 0$, then $\gamma = 1$ and the optimal solution does not depend on the parameters of A, and is given by

$$\mathbb{E}_{te}[Y|X = x] = x. \tag{18}$$

However, for any $\sigma_x^2 > 0$, the optimal solution under $\mathbb{P}_{te}(A)$ depends on μ_A, σ_A^2.

Rewriting residuals Regardless of whether the Gaussian assumption holds, for a given predictor $\hat{Y} = \alpha + \gamma x$, we can write the error $Y - \hat{Y}$ as a function that is linear in A and the noise variables

$$\begin{aligned} Y - \hat{Y} &= (A + \epsilon_Y) - \gamma(A + \epsilon_X) - \alpha \\ &= A(1-\gamma) + (\epsilon_Y - \gamma\epsilon_X - \alpha). \end{aligned}$$

Optimizing for a known target distribution The mean squared error $\mathbb{E}[(Y - \hat{Y})^2]$ can be written as a function of α, γ, and the mean and variance of A under $\mathbb{P}_{te}(A)$. Here, all expectations are taken with respect to the test distribution.

$$\begin{aligned} \mathbb{E}_{te}[(y - \hat{y})^2] &= \mathbb{E}_{te}[\mathbb{E}_{te}[(y - \hat{y})^2 \mid A]] \\ &= \alpha^2 - 2\alpha\mathbb{E}_{te}[A](1-\gamma) \\ &\quad + (1-\gamma)^2\mathbb{E}_{te}[A^2] + \gamma^2\sigma_x^2 + \sigma_y^2. \end{aligned} \tag{19}$$

By first-order conditions, this expression is minimized by

$$\alpha^* = \mu_A(1 - \gamma^*) \qquad \gamma^* = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_X^2}. \tag{20}$$

When $\sigma_A^2 \to \infty$, then $\gamma^* \to 1$ from Equation (20). This is intuitive, because in Equation (19), $\gamma = 1$ renders the MSE functionally independent of the distribution of A.

Optimizing for a worst-case distribution Equation (20) shows the optimal solution under a known target distribution, if μ_A, σ_A^2 were known in advance. However, a similar intuition applies to the case where $\mathbb{P}_{te}(A)$ is unknown, but we expect it to lie in a particular class. Consider interventions of the form $do(A := \nu)$, where we constrain ν to lie in the set of random variables $C(\lambda) := \{\nu : \mathbb{E}[\nu^2] \leq \lambda\}$. In this case, our worst-case loss is given by

$$\begin{aligned}&\sup_{\nu \in C(\lambda)} \mathbb{E}_\nu[(Y - \hat{Y})^2] \\ &= \sup_{\nu \in C(\lambda)} (1-\gamma)\left[-2\alpha\mathbb{E}[\nu] + (1-\gamma)\mathbb{E}[\nu^2]\right] \\ &\quad + \alpha^2 + \gamma^2\sigma_X^2 + \sigma_Y^2,\end{aligned}$$

where the last line does not depend on ν. We observe that $\alpha^* = 0$, by analyzing two cases. First, if $\gamma = 1$, then the first term is eliminated, and the only term that depends on α is α^2. Second, if $\gamma \neq 1$, then $(1-\gamma)^2 > 0$, the first term is partially maximized when $\mathbb{E}[\nu^2] = \lambda$, and if $\alpha \neq 0$, then the expression can be made even larger by choosing a deterministic $\nu = \pm\sqrt{\lambda}$ (instead of e.g., a random $\nu \sim \mathcal{N}(0, \lambda^2)$), depending on the sign of $\alpha(1-\gamma)$. From this (and the presence of the α^2 term in the second line) it follows that $\alpha^* = 0$, in this case as well. When $\alpha = 0$, the supremum is obtained by any random or deterministic ν such that $\mathbb{E}[\nu^2] = \lambda$.

With $\alpha^* = 0$ and taking $\mathbb{E}[\nu^2] = \lambda$ in the supremum, this expression simplifies to

$$\begin{aligned}&\sup_{\nu \in C(\lambda)} \mathbb{E}_\nu[(Y - \hat{Y})^2] \\ &= (1-\gamma)^2\lambda + \gamma^2\sigma_X^2 + \sigma_Y^2.\end{aligned}$$

Differentiating with respect to γ, we obtain

$$\gamma^* = \frac{\lambda}{\sigma_X^2 + \lambda}.$$

Here, λ trades off accuracy and stability; As $\lambda \to \infty$, we recover the solution where $\gamma^* = 1$, but for situations where σ_X^2 is large and λ is bounded, we are better off choosing $\gamma^* < 1$.

B. Example: Non-identifiability of Ω_W

Overview In the context of Section 3.1, we give a concrete example to demonstrate the non-identifiability of Ω_W, defined in (12). We focus on the simple case when W is one dimensional, and the matrix Ω_W reduces to a single number $\rho_W := \beta_W^2/(\beta_W^2 + \sigma_W^2)$, indicating the signal-to-variance ratio of W. We give an example of an observed distribution for which ρ_W is not identified, and moreover, the optimal predictor with respect to the robustness set $C_A(\lambda)$ is not identified (see Figure 10).

Setup If $(X, Y, W) \in \mathbb{R}^3$ is distributed multivariate normal with zero mean, then their covariance matrix fully determines the observed distribution. Let that covariance matrix be denoted by $\Sigma_{(X,Y,W)} \in \mathbb{R}^{3\times 3}$, which gives us six observed moments of the distribution

$$\Sigma_{(X,Y,W)} := \begin{pmatrix} \mathbb{E}[X^2] & \cdot & \cdot \\ \mathbb{E}[XY] & \mathbb{E}[Y^2] & \cdot \\ \mathbb{E}[WX] & \mathbb{E}[WY] & \mathbb{E}[W^2] \end{pmatrix},$$

where we only show the lower triangular portion, since the matrix is symmetric. Suppose that we knew that this observed distribution was generated by the following SCM, but that we do not know the values for the parameters $(\beta_W, \beta_X, \beta_Y, \alpha, \sigma_W^2, \sigma_X^2, \sigma_Y^2)$

$$\begin{aligned} A &:= \epsilon_A & \epsilon_A &\sim \mathcal{N}(0,1) \\ W &:= \beta_W A + \epsilon_W & \epsilon_W &\sim \mathcal{N}(0, \sigma_W^2) \\ X &:= \beta_X A + \epsilon_X & \epsilon_X &\sim \mathcal{N}(0, \sigma_X^2) \\ Y &:= \alpha X + \beta_Y A + \epsilon_Y & \epsilon_Y &\sim \mathcal{N}(0, \sigma_Y^2), \end{aligned}$$

where $\epsilon_A, \epsilon_W, \epsilon_X, \epsilon_Y$ are jointly independent. We can attempt to identify the parameters using the following relationships implied by the SCM, and matching these to the moments that we observe

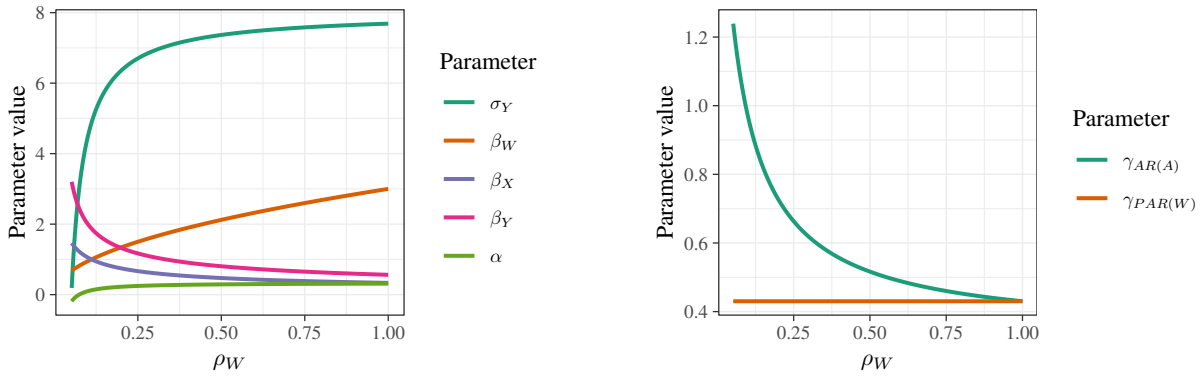
$$\begin{aligned} \mathbb{E}[WX] &= \beta_W\beta_X \\ \mathbb{E}[XY] &= \beta_Y\beta_X + \alpha\mathbb{E}[X^2] \\ \mathbb{E}[WY] &= \beta_W(\beta_Y + \alpha\beta_X) \\ \mathbb{E}[W^2] &= \beta_W^2 + \sigma_W^2 \\ \mathbb{E}[X^2] &= \beta_X^2 + \sigma_X^2 \\ \mathbb{E}[Y^2] &= \alpha^2\mathbb{E}[X^2] + 2\alpha\beta_Y\beta_X + \beta_Y^2 + \sigma_Y^2 \end{aligned}$$

However, as we will see, this does not identify the parameters. In particular, there is a set of parameterizations which all give rise to the same observed distribution, and which imply different values of the signal-to-variance ratio $\rho_W := \beta_W^2/(\beta_W^2 + \sigma_W^2)$.

A class of observationally equivalent SCMs Let $\theta := (\beta_W, \beta_X, \beta_Y, \alpha, \sigma_W^2, \sigma_X^2, \sigma_Y^2) \in \mathbb{R}^7$ be the parameters of the SCM, and let $\Sigma = f(\theta)$ be the covariance matrix over (X, Y, W) implied by these parameters.

For any covariance matrix Σ, there exists a subset $C \subset [0, 1]$ such that for any $\rho_W \in C$, we can write the parameters as a function of ρ_W, such that $f(\theta(\rho_W)) = \Sigma$. The set C is constrained by the observed moments: In particular, as we show below, $\rho_W \geq \text{corr}(W, X)^2$ due to the constraint that $\sigma_X^2 \geq 0$, and the condition that $\sigma_Y^2 \geq 0$ also imposes a lower bound. In particular, for the covariance matrix below, we demonstrate numerically that $[0.06, 1] \subset C$.

$$\Sigma_{(X,Y,W)} := \begin{pmatrix} 9 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 9 \end{pmatrix}.$$



(a) (b)

Figure 10. (a) SCM parameters that all give rise to the same observational distribution, and observe that (b) the parameter $\gamma_{AR(A)}$ (as if A were observed) can diverge substantially from the solution $\gamma_{PAR(W)}$, when a single proxy is available. $\lambda = 5$ for this example.

We now give a strategy for constructing $\theta(\rho_W)$, given a desired ρ_W (including checking the constraint that this $\rho_W \in C$). Suppose that W and X are positively correlated, as in this example. Fixing some $\rho_W \in [0, 1]$, we start by writing β_W, σ_W as functions of ρ_W, where

$$\beta_W := \sqrt{\mathbb{E}[W^2]\rho_W}$$
$$\sigma_W^2 := \mathbb{E}[W^2](1 - \rho_W).$$

The first constraint, that $\sigma_X^2 \geq 0$, can be captured as follows. Let $\rho_X := \beta_X^2/\mathbb{E}[X^2]$. Observe that $\sqrt{\rho_X \rho_W} = \text{corr}(W, X)$. This implies a lower bound on ρ_W, given by $\rho_W \geq \text{corr}(W, X)^2$, since $\rho_X \leq 1$ due to $\sigma_X^2 \geq 0$. This also implies that ρ_X is determined uniquely by ρ_W, and is given by $\rho_X = \text{corr}(W, X)^2/\rho_W$. From this we can write

$$\beta_X := \sqrt{\mathbb{E}[X^2]\rho_X}$$
$$\sigma_X^2 := \mathbb{E}[X^2](1 - \rho_X).$$

These choices for $(\beta_W, \sigma_W^2, \beta_X, \sigma_X^2)$ match the observed moments $\mathbb{E}[X^2], \mathbb{E}[W^2], \mathbb{E}[WX]$. Then the rest of the parameters can be found as follows, where β_W, β_X are fixed as above

$$\beta_Y := \frac{1}{\beta_W(1 - \rho_X)}\left(\mathbb{E}[WY] - \frac{\mathbb{E}[XY]\mathbb{E}[WX]}{\mathbb{E}[X^2]}\right)$$
$$\alpha := \frac{\mathbb{E}[XY] - \beta_Y\beta_X}{\mathbb{E}[X^2]}$$
$$\sigma_Y^2 := \mathbb{E}[Y^2] - \beta_Y^2 - 2\alpha\beta_Y\beta_X - \alpha^2\mathbb{E}[X^2]$$

where all of these are functions of ρ_W, in that β_W, β_X are functions of ρ_W. It remains to verify that for a given choice of ρ_W, we satisfy the constraint that $\sigma_Y^2 \geq 0$. For simplicity, we check this constraint computationally in the context of Example 1, for a range of values of ρ_W, and we give the set of observationally-equivalent parameters in Figure 10a, where valid values of ρ_W range over $[0.06, 1]$.

Next we show that the Proxy Anchor Regression estimator, $\gamma_{PAR(W)}$, differs from the Anchor Regression estimator, $\gamma_{AR(A)}$, and more so when ρ_W becomes small. This is shown in Figure 10b, for $\lambda = 5$, and we give the relevant computations here.

Solution to PAR(W) If we have a single proxy, then we can write down the optimization problem Equation (10) as

$$\begin{aligned}&\min_\gamma \mathbb{E}[(Y - \gamma X)^2] + \lambda\mathbb{E}[(Y - \gamma X)W]^2\mathbb{E}[W^2]^{-1}\\ =&\min_\gamma \mathbb{E}[Y^2] - 2\gamma\mathbb{E}[YX] + \gamma^2 E[X^2]\\ &+ \lambda(\mathbb{E}[YW] - \gamma\mathbb{E}[XW])^2\mathbb{E}[W^2]^{-1},\end{aligned}$$

from which we obtain the optimal solution

$$\gamma_{PAR(W)} = \frac{\mathbb{E}[YX]\mathbb{E}[W^2] + \lambda\mathbb{E}[YW]}{\mathbb{E}[X^2]\mathbb{E}[W^2] + \lambda\mathbb{E}[XW]}.$$

Solution to AR(A) First, we can write the residual as

$$\begin{aligned}Y - \hat{Y} &= Y - \gamma X\\ &= \alpha X + \beta_Y A + \epsilon_Y - \gamma\beta_X A - \gamma\epsilon_X\\ &= \alpha(\beta_X A + \epsilon_X) + \beta_Y A + \epsilon_Y - \gamma\beta_X A - \gamma\epsilon_X\\ &= A((\alpha - \gamma)\beta_X + \beta_Y) + (\alpha - \gamma)\epsilon_X + \epsilon_Y,\end{aligned}$$

such that the expected squared error is given by

$$\begin{aligned}&\mathbb{E}_{do(A:=\nu)}(Y - \hat{Y})^2\\ &= ((\alpha - \gamma)\beta_X + \beta_Y)^2\mathbb{E}[\nu^2] + (\alpha - \gamma)^2\sigma_X^2 + \sigma_Y^2, \quad (21)\end{aligned}$$

and when $\nu \in \{\nu : \mathbb{E}[\nu^2] \leq (1 + \lambda)\}$, taking the supremum involves replacing $\mathbb{E}[\nu^2]$ with $(1 + \lambda)$. Optimizing Equation (21) with respect to γ, we obtain

$$\begin{aligned}&\frac{\partial}{\partial\gamma}\left[((\alpha - \gamma)\beta_X + \beta_Y)^2(1 + \lambda) + (\alpha - \gamma)^2\sigma_X^2 + \sigma_Y^2\right]\\ &= -2\beta_X(\alpha\beta_X - \gamma\beta_X + \beta_Y)(1 + \lambda) - 2(\alpha - \gamma)\sigma_X^2,\end{aligned}$$

which implies that

$$0 = \beta_X(\alpha\beta_X - \gamma\beta_X + \beta_Y)(1+\lambda) + (\alpha-\gamma)\sigma_X^2$$
$$= (\alpha\beta_X^2 + \beta_X\beta_Y)(1+\lambda) - \gamma\beta_X^2(1+\lambda) + \alpha\sigma_X^2 - \gamma\sigma_X^2,$$

so that the optimal choice of γ is given by

$$\gamma_{AR(A)} = \frac{(\alpha\beta_X^2 + \beta_X\beta_Y)(1+\lambda) + \alpha\sigma_X^2}{\beta_X^2(1+\lambda) + \sigma_X^2}.$$

If $\lambda = -1$, this recovers the causal effect of X on Y, and if $\lambda \to \infty$, this recovers a set of coefficients that are invariant to variation in A, as can be seen by plugging the resulting coefficient $\gamma = \alpha + \beta_Y/\beta_X$ into Equation (21).

C. Proofs

C.1. Auxiliary results

First, we show that the proof of Theorem 1 of Rothenhäusler et al. (2021) can be decomposed into two parts, and use this observation to simplify the proof of our Theorem 1. Proposition A1 establishes that ℓ_{PLS} can be written as a quadratic form in the structural parameters $w_\gamma^\top M_A$. Proposition A2 is a straightforward generalization of the techniques used in Rothenhäusler et al. (2021), and establishes that any regularization term that can be written in this way naturally implies a robustness guarantee.

By Assumption 1, our SCM can be written in the following form, where $\epsilon \perp\!\!\!\perp A$, and all variables are mean-zero and have bounded covariance.

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (Id - B)^{-1}(M_A A + \epsilon). \tag{22}$$

In this context, we use the following notational shorthand,

$$w_\gamma := \left((Id-B)^{-1}_{d_X+1,\cdot} - \gamma^\top (Id-B)^{-1}_{1:d_X,\cdot}\right)^\top, \tag{23}$$

such that we can write the residual as a function of both the exogenous noise ϵ and A as

$$R(\gamma) := Y - \gamma^\top X = w_\gamma^\top(\epsilon + M_A A), \tag{24}$$

under the training distribution. (This identity explains the valley in the loss landscape displayed in Figure 3: If $d_A \geq 2$, for any parameter γ, there exist an orthogonal intervention direction $\nu \in (w_\gamma^\top M_A)^\perp$, to which the loss is invariant.)

Proposition A1. *Under Assumption 1,*

$$\ell_{PLS}(X, Y, A; \gamma) = w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A^\top w_\gamma, \tag{25}$$

and

$$\ell_{PLS}(X, Y, W; \gamma) = w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]M_A^\top w_\gamma, \tag{26}$$

where w_γ is defined by Equation (23).

Proof. The first statement follows from Equation (6) and the observation that

$$\mathbb{E}[R(\gamma)A^\top] = \mathbb{E}[w_\gamma^\top(\epsilon + M_A A)A^\top] = w_\gamma^\top \mathbb{E}[\epsilon A^\top] + w_\gamma^\top M_A \mathbb{E}[AA^\top] = w_\gamma^\top M_A \mathbb{E}[AA^\top],$$

where we used $\epsilon \perp\!\!\!\perp A$. Similarly

$$\ell_{PLS}(X, Y, W; \gamma)$$
$$= \mathbb{E}[R(\gamma)W^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WR(\gamma)^\top]$$
$$= \mathbb{E}[w_\gamma^\top(\epsilon + MA)W^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WR(\gamma)^\top]$$
$$= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]M_A^\top w_\gamma,$$

where the first equality follows from Equation (6), and the final equality follows from the fact that $\epsilon \perp\!\!\!\perp W$. □

Proposition A2. *Under Assumption 1, for any λ and any real, symmetric Ω such that $0 \preceq \mathbb{E}[AA^\top] + \lambda\Omega$, any loss function of the form*

$$\ell(\gamma, \lambda) := \ell_{LS}(X, Y; \gamma) + \lambda w_\gamma^\top M_A \Omega M_A^\top w_\gamma, \tag{27}$$

where w_γ is defined by Equation (23), is equal to the following worst-case loss under bounded perturbations

$$\ell(\gamma, \lambda) = \sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2],$$

where

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

Proof. We have, making use of the fact that $\epsilon \perp\!\!\!\perp A$, and $\mathbb{E}[\epsilon] = 0$

$$\sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)}\left[(Y - \gamma^\top X)^2\right]$$
$$= \sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)}\left[(w_\gamma^\top(\epsilon + M_A\nu))^2\right]$$
$$= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + \sup_{\nu \in C(\lambda)} \mathbb{E}[(w_\gamma^\top M_A \nu)^2]$$
$$= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + \sup_{\nu \in C(\lambda)} w_\gamma^\top M_A \mathbb{E}[\nu\nu^\top] M_A{}^\top w_\gamma$$
$$= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + w_\gamma^\top M_A(\mathbb{E}[AA^\top] + \lambda\Omega)M_A{}^\top w_\gamma$$
$$= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A{}^\top w_\gamma + \lambda w_\gamma^\top M_A \Omega M_A{}^\top w_\gamma$$
$$= \mathbb{E}\left[(w_\gamma^\top(\epsilon + M_A A))^2\right] + \lambda w_\gamma^\top M_A \Omega M_A{}^\top w_\gamma$$
$$= \ell_{LS}(X, Y; \gamma) + \lambda w_\gamma^\top M_A \Omega M_A{}^\top w_\gamma$$
$$= \ell(\gamma, \lambda),$$

where in the fifth line we used the definition of $C(\lambda)$. The supremum is achievable even if ν is a deterministic vector, since we can take $\nu := \frac{Sb}{\sqrt{b^\top Sb}}$ where $S := \mathbb{E}[AA^\top] + \lambda\Omega$ and $b := {M_A}^\top w_\gamma$. Then the supremum value is achieved by ν, as $\nu\nu^\top = \frac{Sbb^\top S}{b^\top Sb}$ and $b^\top \nu\nu^\top b = \frac{b^\top Sbb^\top Sb}{b^\top Sb} = b^\top Sb$. To show that $\nu\nu^\top \preceq S$, such that $\nu \in C(\lambda)$, we can take any conformable vector x to see that

$$\begin{aligned} x^\top(S - \nu\nu^\top)x &= x^\top Sx - \frac{x^\top Sbb^\top Sx}{b^\top Sb} \\ &= \langle x, x\rangle - \frac{\langle x, b\rangle^2}{\langle b, b\rangle} \\ &\geq 0, \end{aligned}$$

where we use the fact that $\langle e, f\rangle := e^\top Sf$ defines an inner product, and we apply Cauchy-Schwarz: $\langle x, x\rangle\langle b, b\rangle \geq \langle x, b\rangle^2$. □

In the proofs for Section 3, we will occasionally make use of the following fact, which we prove here to simplify exposition later on.

Proposition A3. *In the setting of a single proxy (i.e., under Assumptions 1 and 2) let Ω_W be defined as follows*

$$\Omega_W := \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]. \tag{28}$$

Then $\Omega_W \preceq \mathbb{E}[AA^\top]$. Furthermore, if $\mathbb{E}[\epsilon_W\epsilon_W^\top]$ is positive definite, then this inequality is strict, that is, $\Omega_W \prec \mathbb{E}[AA^\top]$.

Proof. Recall that $\mathbb{E}[AA^\top]$ and $\mathbb{E}[WW^\top]$ are invertible (and hence positive definite) by assumption.

The inequality $\Omega_W \preceq \mathbb{E}[AA^\top]$ is equivalent to showing that $S := \mathbb{E}[AA^\top] - \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] \succeq 0$. Observe that S is the Schur complement of the matrix $K := \mathbb{E}\left[\begin{pmatrix} A \\ W \end{pmatrix}\begin{pmatrix} A \\ W \end{pmatrix}^\top\right]$. The matrix K is positive semi-definite (PSD) if and only if $\mathbb{E}[AA^\top]$ is positive definite (true by assumption) and S is PSD (see Zhang (2006, Theorem 1.12b)). Since K is PSD by construction, as the covariance matrix of A, W, this implies that $S \succeq 0$.

Similarly, K is positive definite (PD) if and only if $\mathbb{E}[AA^\top]$ and S are both PD (see Zhang (2006, Theorem 1.12a)). Under the condition that $\mathbb{E}[\epsilon_W\epsilon_W^\top]$ is full-rank, then K is PD, and the second inequality follows. □

C.2. Proof of additional results

Proof of Equation (9). It follows from Proposition A1 that

$$\begin{aligned} \ell_{PLS}(X, Y, A; \gamma) &= w_\gamma^\top M_A \Omega_A M_A^\top w_\gamma \\ \ell_{PLS}(X, Y, W; \gamma) &= w_\gamma^\top M_A \Omega_W M_A^\top w_\gamma, \end{aligned}$$

where $\Omega_W := \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]$ and $\Omega_A := \mathbb{E}[AA^\top]$ are both full rank because $\mathbb{E}[AW^\top] = \mathbb{E}[AA^\top]\beta_W$ and by assumptions that $\mathbb{E}[WW^\top], \mathbb{E}[AA^\top]$ and β_W are full rank. Hence both $\ell_{PLS}(X, Y, A; \gamma)$ and $\ell_{PLS}(X, Y, W; \gamma)$ are zero exactly when $w_\gamma^\top M_A = 0$. □

C.3. Proof of main results

C.3.1. Section 3

Proof of Theorem 1. We use the fact that ϵ is mean-zero and independent of both A and W. Recall that

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

where we suppress the dependence on X, Y in the notation. Letting w_γ be as defined in Equation (23), it follows from Equation (26) that

$$\begin{aligned} &\ell_{PLS}(X, Y, W; \gamma) \\ &= w_\gamma^\top M_A \underbrace{\mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]}_{\Omega_W} M_A^\top w_\gamma. \end{aligned}$$

The statement then follows from the application of Proposition A2, and the fact that $\Omega_W \preceq \mathbb{E}[AA^\top]$ (by Proposition A3), such that $\mathbb{E}[AA^\top] + \lambda\Omega_W \succeq 0$ for all $\lambda \geq -1$. □

Proof of Proposition 1. Recall that the guarantee regions are given by

$$\begin{aligned} C_A(\lambda) &= \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\mathbb{E}[AA^\top]\} \\ C_W(\lambda) &= \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\Omega_W\} \\ C_{OLS} &= \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top]\}, \end{aligned}$$

where

$$\Omega_W = \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top].$$

The fact that $\mathbb{E}[WW^\top]^{-1} \succ 0$ implies $\Omega_W \succeq 0$, and this implies that $C_{OLS} \subseteq C_W(\lambda)$ for $\lambda \geq 0$. Showing $C_W(\lambda) \subset C_A(\lambda)$ amounts to showing that $\Omega_W \prec \mathbb{E}[AA^\top]$, which holds by Proposition A3 when $\mathbb{E}[\epsilon_W\epsilon_W^\top] \succ 0$.

Next, we prove that C_W is monotonically decreasing in the noise $\mathbb{E}[\epsilon_W\epsilon_W^\top]$, in the sense that if $\mathbb{E}[\epsilon_W\epsilon_W^\top] \preceq \mathbb{E}[\eta_W\eta_W^\top]$ then

$$\begin{aligned} &\mathbb{E}_\eta[AW^\top]\mathbb{E}_\eta[WW^\top]^{-1}\mathbb{E}_\eta[WA^\top] \\ &\preceq \mathbb{E}_\epsilon[AW^\top]\mathbb{E}_\epsilon[WW^\top]^{-1}\mathbb{E}_\epsilon[WA^\top], \end{aligned}$$

where $\mathbb{E}_\eta$ is the expectation in the SCM where $W := \beta_W^\top A + \eta_W$ (and similar for $\mathbb{E}_\epsilon$).

Suppose that $\mathbb{E}[\epsilon_W\epsilon_W^\top] \preceq \mathbb{E}[\eta_W\eta_W^\top]$. Then $\mathbb{E}_\eta[WW^\top]^{-1} \preceq \mathbb{E}_\epsilon[WW^\top]^{-1}$, and since $\mathbb{E}_\eta[AW^\top] =$

$\mathbb{E}_\epsilon[AW^\top]$, for any vector $x \in \mathbb{R}^{d_A}$ it holds that,

$$(\mathbb{E}_\eta[WA^\top]x)^\top \mathbb{E}_\eta[WW^\top]^{-1}(\mathbb{E}_\eta[WA^\top]x) \leq (\mathbb{E}_\epsilon[WA^\top]x)^\top \mathbb{E}_\epsilon[WW^\top]^{-1}(\mathbb{E}_\epsilon[WA^\top]x).$$

This establishes the matrix inequality.

To conclude the proof, suppose that $\mathbb{E}[\epsilon_W \epsilon_W^\top] = 0$, $d_A = d_W$ and that β_W has full rank. It then follows that

$$\begin{aligned}\Omega_W &= \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] \\ &= \mathbb{E}[AA^\top]\beta_W(\beta_W^\top \mathbb{E}[AA^\top]\beta_W)^{-1}\beta_W^\top \mathbb{E}[AA^\top] \\ &= \mathbb{E}[AA^\top]\beta_W\beta_W^{-1}\mathbb{E}[AA^\top]^{-1}{\beta_W^\top}^{-1}\beta_W^\top \mathbb{E}[AA^\top] \\ &= \mathbb{E}[AA^\top],\end{aligned}$$

such that $C_W(\lambda) = \mathbb{E}[AA^\top] + \lambda\Omega_W = (1+\lambda)\mathbb{E}[AA^\top] = C_A(\lambda)$. □

Proof of Theorem 2. Let w_γ be defined as in Equation (23). We can write the population quantity as follows, making use of the fact that ϵ, ϵ_Z, and ϵ_W are jointly independent, and that all errors have zero mean.

$$\begin{aligned}&\ell_\times(W, Z; \gamma) \\ &= \mathbb{E}[(Y - \gamma^\top X)W^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[Z(Y - \gamma^\top X)^\top] \\ &= \mathbb{E}[w_\gamma^\top(M_A A + \epsilon)W^\top]\mathbb{E}[ZW^\top]^{-1} \\ &\quad \cdot \mathbb{E}[Z(A^\top M_A^\top + \epsilon^\top)w_\gamma] \\ &= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZA^\top]M_A^\top w_\gamma \\ &= w_\gamma^\top M_A \mathbb{E}[A(A^\top\beta_W + \epsilon_W^\top)] \\ &\quad \mathbb{E}[(\beta_Z^\top A + \epsilon_Z)(A^\top\beta_W + \epsilon_W^\top)]^{-1} \\ &\quad \mathbb{E}[(\beta_Z^\top A + \epsilon_Z)A^\top]M_A^\top w_\gamma \\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top]\beta_W\left(\beta_Z^\top \mathbb{E}[AA^\top]\beta_W\right)^{-1} \\ &\quad \beta_Z^\top \mathbb{E}[AA^\top]M_A^\top w_\gamma \\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top]\beta_W\beta_W^{-1}\mathbb{E}[AA^\top]^{-1}(\beta_Z^\top)^{-1} \\ &\quad \beta_Z^\top \mathbb{E}[AA^\top]M_A^\top w_\gamma \\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top]\mathbb{E}[AA^\top]^{-1}\mathbb{E}[AA^\top]M_A^\top w_\gamma \\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top]M_A^\top w_\gamma\end{aligned}$$

The result follows from Proposition A1. □

In the main text, we state that the xPAR(W, Z) objective is convex in γ and has a closed form solution. We give the proof here:

Proposition A4. *Under Assumptions 1, 3 and 4, the loss in Equation (14) is convex in γ, and its minimizer is given by*

$$\gamma^*_{\times PAR} := \left(2\mathbb{E}[XX^\top] + \lambda(L + L^\top)\right)^{-1} \left(2\mathbb{E}[XY^\top] + \lambda(K_1 + K_2)\right),$$

where we define

$$\begin{aligned}L &:= \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZX^\top], \\ K_1 &:= \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZY^\top] \\ K_2 &:= \mathbb{E}[XZ^\top]\mathbb{E}[WZ^\top]^{-1}\mathbb{E}[WY^\top].\end{aligned}$$

Proof. By Theorem 2 and Equation (7), $\ell_{\times PAR}(W, Z; \gamma, \lambda) = \ell_{AR}(X, Y, A; \gamma, \lambda)$, and the latter is convex in γ, since it is the sum ℓ_{LS}, which is convex, and $\lambda\ell_{PLS}(X, Y, A; \gamma)$, which is a quadratic form by Proposition A1 and hence convex.

Consequently optimal solution can be found by taking the gradient of $\ell_{\times PAR}(W, Z; \gamma, \lambda) = \ell_{LS} + \lambda\ell_\times$ with respect to γ and equating it to 0. Letting $D := \mathbb{E}[ZW^\top]^{-1}$, we can differentiate $\ell_{\times PAR}$ term wise, using Equation (13) to rewrite $\ell_\times$:

$$\begin{aligned}0 = 2\gamma^\top\mathbb{E}[XX^\top] &- 2\mathbb{E}[YX^\top] \\ &- \lambda\mathbb{E}[YW^\top]D\mathbb{E}[ZX^\top] \\ &- \lambda\mathbb{E}[YZ^\top]D^\top\mathbb{E}[WX^\top] \\ &+ \lambda\gamma^\top(L + L^\top),\end{aligned}$$

where $L := \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZX^\top]$. Defining $K_1 := \mathbb{E}[XW^\top]D\mathbb{E}[ZY^\top]$ and $K_2 := \mathbb{E}[XZ^\top]D^\top\mathbb{E}[WY^\top]$, and rearranging, we obtain:

$$\begin{aligned}&\gamma^\top(2\mathbb{E}[XX^\top] + \lambda(L + L^\top)) \\ &= 2\mathbb{E}[YX^\top] + \lambda(K_1^\top + K_2^\top),\end{aligned}$$

so by transposing and solving for γ, we get the expression from the statement. □

C.3.2. Section 4

Proof of Proposition 2. Let w_γ be defined by (23) and for any γ let $b_\gamma^\top := w_\gamma^\top M_A$. We can write the loss as follows

$$\begin{aligned}&\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2] \\ &= \mathbb{E}[(w_\gamma^\top(\epsilon + M_A\nu) - \alpha)^2] \\ &= \mathbb{E}[(w_\gamma^\top\epsilon + w_\gamma^\top M_A\nu - \alpha)^2] \\ &\overset{\epsilon \perp\!\!\!\perp \nu}{=} \mathbb{E}[(w_\gamma^\top\epsilon)^2] + \mathbb{E}[(w_\gamma^\top M_A\nu - \alpha)^2] \\ &= \mathbb{E}[(w_\gamma^\top\epsilon)^2] + \mathbb{E}[(w_\gamma^\top M_A A)^2] \\ &\quad - \mathbb{E}[(w_\gamma^\top M_A A)^2] + \mathbb{E}[(w_\gamma^\top M_A\nu - \alpha)^2] \\ &= \ell_{LS}(\gamma) - \mathbb{E}[(b_\gamma^\top A)^2] + \mathbb{E}[(b_\gamma^\top\nu - \alpha)^2] \\ &= \ell_{LS}(\gamma) - b_\gamma^\top\mathbb{E}[AA^\top]b_\gamma^\top \\ &\quad + b_\gamma^\top\mathbb{E}[\nu\nu^\top]b_\gamma - 2\mathbb{E}[b_\gamma^\top\nu]\alpha + \alpha^2 \\ &= \ell_{LS}(\gamma) + b_\gamma^\top\left(\mathbb{E}[\nu\nu^\top] - \mathbb{E}[AA^\top]\right)b_\gamma \\ &\quad - 2\mathbb{E}[b_\gamma^\top\nu]\alpha + \alpha^2\end{aligned}$$

$$
\begin{aligned}
&= \ell_{LS}(\gamma) \\
&\quad + b_\gamma^\top \left(\mathbb{E}[\nu\nu^\top] - \mathbb{E}[AA^\top]\right) b_\gamma - \left(b_\gamma^\top \mathbb{E}[\nu]\right)^2 \\
&\quad + \left(b_\gamma^\top \mathbb{E}[\nu]\right)^2 - 2\mathbb{E}[b_\gamma^\top \nu]\alpha + \alpha^2 \\
&= \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma + \left(b_\gamma^\top \mathbb{E}[\nu] - \alpha\right)^2,
\end{aligned}
$$

where for any value of γ, that minimizing with respect to α yields $\alpha^* = b_\gamma^\top \mathbb{E}[\nu]$, where $b_\gamma^\top = w_\gamma^\top M_A$. Given that we can write the structural relationship $Y - \gamma^\top X = b_\gamma^\top A + w_\gamma^\top \epsilon$, and knowing that $\mathbb{E}[\epsilon] = 0$ and that $\epsilon \perp\!\!\!\perp A$, we know that $b_\gamma^\top A$ is the conditional expectation of $R(\gamma)$ given A. □

In the main text, we note that Equation (16) (the objective function ℓ_{TAR}) is convex in γ, α, and has a closed form solution. We prove that result here.

Proposition A5. *Under Assumption 1, the minimizer $\gamma^*_{TAR}, \alpha^*_{TAR}$ of Equation (16) is given by*

$$
\begin{aligned}
\gamma^* &= \left(\mathbb{E}[XX^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AX^\top]\right)^{-1} \\
&\quad \left(\mathbb{E}[XY^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AY^\top]\right) \\
\alpha^* &= b_{\gamma^*}^\top \mu_\nu,
\end{aligned}
$$

where $\Omega = \mathbb{E}[AA^\top]^{-1}(\Sigma_\nu - \Sigma_A)\mathbb{E}[AA^\top]^{-1}$, and $b_\gamma^\top$ is defined in Equation (15).

Proof of Proposition A5. Let w_γ be as defined in Equation (23) and let $b_\gamma^\top := w_\gamma^\top M_A$. Since $\mathbb{E}[(Y - \gamma^\top X) \mid A] = \mathbb{E}[w_\gamma^\top [M_A A + \epsilon] \mid A] = b_\gamma^\top A$, for any γ, $b_\gamma^\top$ is the linear regression coefficient of $(Y - \gamma^\top X)$ onto A, so we may write $b_\gamma^\top = \mathbb{E}[(Y - \gamma^\top X)A^\top]\mathbb{E}[AA^\top]^{-1}$. Plugging in the optimal value $\alpha(\gamma) := b_\gamma^\top \mu_\nu$, we obtain

$$
\begin{aligned}
&\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha(\gamma)) \\
&= \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma \\
&= \ell_{LS}(\gamma) + \mathbb{E}[(Y - \gamma^\top X)A^\top]\Omega\mathbb{E}[A(Y - \gamma^\top X)^\top]
\end{aligned}
$$

This objective is convex in γ. The derivative of the loss with respect to γ is

$$
-2(\mathbb{E}[(Y - \gamma^\top X)X^\top] + \mathbb{E}[(Y - \gamma^\top X)A^\top]\Omega\mathbb{E}[AX^\top]),
$$

and equating to 0 and solving for γ yields

$$
\begin{aligned}
\gamma^* &= \left(\mathbb{E}[XX^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AX^\top]\right)^{-1} \\
&\quad \left(\mathbb{E}[XY^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AY^\top]\right).
\end{aligned}
$$

□

We also claim in the main text that if ν is a constant, then the minimizer of Equation (16) can be found by performing OLS using both X, A as predictors, and then plugging in the known value ν for A in prediction. We prove that result here.

Proof. If ν is a constant, then we can write the first two terms as follows, where w_γ is defined in Equation (23).

$$
\begin{aligned}
&\ell_{LS} - b_\gamma^\top \Sigma_A b_\gamma \\
&= \mathbb{E}[(w_\gamma^\top (M_A A + \epsilon))^2] - w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A^\top b_\gamma \\
&= \mathbb{E}[(w_\gamma^\top (M_A A + \epsilon))^2] - \mathbb{E}[(w_\gamma^\top M_A A)^2] \\
&= \mathbb{E}[(w_\gamma^\top \epsilon)^2]
\end{aligned}
$$

which is equivalent to the objective for the loss when Y, X are residualized with respect to A (see Section 8.6 of Rothenhäusler et al. (2021)). By the Frish-Waugh-Lovell theorem (Lovell, 1963; 2008), this yields the same coefficients γ for X as if we had performed regression on X, A together. For this value of γ, $b_\gamma^\top$ is the coefficient that we would obtain for A in the joint regression, because it equals the regression coefficients for $Y - \gamma^\top X$ on A. □

Proof of Proposition 3. We use ν to denote the random shift. Let $\nu \in T(\mu_\nu, \Sigma_\nu)$, or equivalently, let $\nu := \mu_\nu + \delta$, where μ_ν is fixed and δ satisfies the constraint that $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$, where Σ_ν is a symmetric positive definite matrix. Let w_γ be defined by (23) and for any γ let $b_\gamma^\top := w_\gamma^\top M_A$. We can write the loss as follows

$$
\begin{aligned}
&\sup_{\nu \in T} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2] \\
&= \sup_{\nu \in T} \mathbb{E}[(w_\gamma^\top (\epsilon + M_A \nu) - \alpha)^2] \\
&= \sup_{\nu \in T} \mathbb{E}[(w_\gamma^\top \epsilon + w_\gamma^\top M_A \nu - \alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top \epsilon)^2] + \sup_{\nu \in T} \mathbb{E}[(w_\gamma^\top M_A \nu - \alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top \epsilon)^2] + \mathbb{E}[(w_\gamma^\top M_A A)^2] \\
&\qquad - \mathbb{E}[(w_\gamma^\top M_A A)^2] + \sup_{\nu \in T} \mathbb{E}[(w_\gamma^\top M_A \nu - \alpha)^2] \\
&= \ell_{LS}(\gamma) - \mathbb{E}[(b_\gamma^\top A)^2] + \sup_{\nu \in T} \mathbb{E}[(b_\gamma^\top \nu - \alpha)^2],
\end{aligned}
$$

where on the fourth line we used the fact that $\mathbb{E}[\epsilon\nu] = 0$ by the fact that $\nu = \mu_v + \delta$, and δ is independent of ϵ. In the last line we replaced $w_\gamma^\top M_A$ by $b_\gamma^\top$. We can re-write the last term as follows, where the supremum with respect to δ is constrained in the set $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$

$$
\begin{aligned}
&\sup_{\nu \in T} \mathbb{E}[(b_\gamma^\top \nu - \alpha)^2] \\
&= \sup_{\delta : \mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu} \mathbb{E}[(b_\gamma^\top (\delta + \mu_\nu) - \alpha)^2] \\
&= \sup_{\delta} \mathbb{E}[(b_\gamma^\top \delta + b_\gamma^\top \mu_\nu - \alpha)^2] \\
&= \sup_{\delta} \mathbb{E}[(b_\gamma^\top \delta)^2] + 2\mathbb{E}[(b_\gamma^\top \delta)](b_\gamma^\top \mu_\nu - \alpha) + \mathbb{E}[(b_\gamma^\top \mu_\nu - \alpha)^2] \\
&= b_\gamma^\top \Sigma_\nu b_\gamma + 2 \left\| b_\gamma \right\|_{\Sigma_\nu} \cdot \left| b_\gamma^\top \mu_\nu - \alpha \right| + (b_\gamma^\top \mu_\nu - \alpha)^2,
\end{aligned}
$$

where $\|b_\gamma\|_{\Sigma_\nu} := \sqrt{b_\gamma^\top \Sigma_\nu b_\gamma}$ is the norm induced by the inner product defined with respect to Σ_ν. In the last line, we have used the fact that the expression is maximized (subject to the constraint) by the deterministic distribution $\delta_* = \pm \frac{\Sigma_\nu b_\gamma}{\sqrt{b_\gamma^\top \Sigma_\nu b_\gamma}}$ where the sign depends on the sign of $(b_\gamma^\top \mu_\nu - \alpha)$: δ_* satisfies $b_\gamma^\top \delta_* \delta_*^\top b_\gamma = b_\gamma^\top \Sigma_\nu b_\gamma$, maximizing the first term. Further, the second term is also maximized by δ_*, because if any other random or deterministic δ satisfies $|\mathbb{E} b_\gamma^\top \delta| > |b_\gamma^\top \delta_*|$, it follows by Jensens inequality that $\mathbb{E}[(b_\gamma^\top \delta)^2] \geq (\mathbb{E}[(b_\gamma^\top \delta)])^2 > (b_\gamma^\top \delta_*)^2 = b_\gamma^\top \Sigma_\nu b_\gamma$, such that $\mathbb{E}[\delta\delta^\top] \succ \Sigma_\nu$, so δ is not in the set over which the supremum is taken. Consequently, the supremum is attained at δ_*, because δ_* maximizes both terms.

Using this expression for the supremum, we can write the objective as

$$\begin{aligned}
&\sup_{\nu \in T} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2] \\
&= \ell_{LS}(\gamma) + b_\gamma^\top (\Sigma_\nu - \Sigma_A) b_\gamma \\
&\qquad + 2\,\|b_\gamma\|_\Sigma \cdot \left| b_\gamma^\top \mu_\nu - \alpha \right| + (b_\gamma^\top \mu_\nu - \alpha)^2,
\end{aligned}$$

for which the optimal choice of α^* is given by $b_\gamma^\top \mu_\nu$, for any γ, and for this choice of α, we can see that $\gamma^* = \arg\min_\gamma \ell_{LS}(\gamma) + b_\gamma^\top (\Sigma_\nu - \Sigma_A) b_\gamma$. □

D. Targeting with proxies

Definition 6 (Proxy Targeted Anchor Regression)**.** Let $\tilde{\mu} := \mathbb{E}_{do(A:=\nu)}[W]$ denote the mean of W under intervention, and let $\tilde{\Sigma}_W := \mathrm{Cov}_{do(A:=\nu)}(W)$ denote the covariance. We define

$$\begin{aligned}
&\ell_{PTAR}(W; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha) \qquad (29)\\
&= \ell_{LS}(\gamma) + c_\gamma^\top \left(\tilde{\Sigma}_W - \Sigma_W \right) c_\gamma + (c_\gamma^\top \tilde{\mu} - \alpha)^2,
\end{aligned}$$

where $c_\gamma^\top := \mathbb{E}[R(\gamma) W^\top] {\Sigma_W}^{-1}$.

As mentioned in the main text, Equation (29) is not generally equal to Equation (16), and does not generally yield the optimal predictor under the targeted loss. A simple example is given in Proposition A6.

Proposition A6. *Assume Assumptions 1, 2, and that $\mathbb{E}[\epsilon_W \epsilon_W^\top]$ is full rank. Let $\nu \overset{(d)}{=} A + \eta$ for the deterministic vector $\eta^T = \mathbb{E}[R(\gamma^*_{OLS}) A^\top]$, where $\overset{(d)}{=}$ indicates equality of distribution, and assume $\eta \neq 0$. Then, the minimizers of Equations (16) and (29) differ, in that*

$$\alpha^*_{PTAR} < \alpha^*_{TAR}$$

and if $d_W = d_A = 1$, and A has unit variance, then $\frac{\alpha^*_{PTAR}}{\alpha^*_{TAR}} = \rho_W$, where $\rho_W := \beta_W^2 / (\beta_W^2 + \mathbb{E}[\epsilon_W^2])$.

Proof. The assumption that $\nu = A + \eta$ implies that $\Sigma_\nu - \Sigma_A = 0$, and $\mathbb{E}[\nu] = \eta$. That is, we have changed the mean of the distribution, but not the covariance. This implies

$$\begin{aligned}
\mathbb{E}[\tilde{W}] &= \beta_W^\top \mathbb{E}[\nu] = \beta_W^\top \eta \\
\Sigma_{\tilde{W}} - \Sigma_W &= \beta_W^\top (\Sigma_\nu - \Sigma_A) \beta_W = 0,
\end{aligned}$$

where in the second equation we use the fact that $\Sigma_W = \beta_W^\top \mathbb{E}[AA^\top] \beta_W + \mathbb{E}[\epsilon_W \epsilon_W^\top]$ (and similarly for $\Sigma_{\tilde{W}}$), and the ϵ_W terms cancel in the subtraction. We can then write both objectives as follows

$$\begin{aligned}
&\ell_{PTAR}(W, \tilde{W}; \gamma, \alpha) \\
&\qquad = \ell_{LS}(\gamma) + \left(c_\gamma^\top \beta_W^\top \eta - \alpha \right)^2 \\
&\qquad = \ell_{LS}(\gamma) + \left(\mathbb{E}[R(\gamma) A^T] \beta_W \Sigma_W^{-1} \beta_W^\top \eta - \alpha \right)^2 \\
&\ell_{TAR}(A, \nu; \gamma, \alpha) \\
&\qquad = \ell_{LS}(\gamma) + \left(b_\gamma^\top \eta - \alpha \right)^2 \\
&\qquad = \ell_{LS}(\gamma) + \left(\mathbb{E}[R(\gamma) A^T] \Sigma_A^{-1} \eta - \alpha \right)^2
\end{aligned}$$

This gives the optimal value of α in both cases as the value that minimizes the second term

$$\begin{aligned}
\alpha^*_{PTAR} &= \mathbb{E}[R(\gamma^*_{PTAR}) A^T] (\beta_W \Sigma_W^{-1} \beta_W^\top) \eta \\
\alpha^*_{TAR} &= \mathbb{E}[R(\gamma^*_{TAR}) A^T] \Sigma_A^{-1} \eta,
\end{aligned}$$

and since the second term can be made equal to zero by these choices of α, the optimal γ in both cases is identically $\gamma^*_{PTAR} = \gamma^*_{TAR} = \gamma^*_{OLS}$, the value of γ that minimizes the first term $\ell_{LS}(\gamma)$. Hence, we can write the difference between these terms as

$$\begin{aligned}
&\alpha^*_{TAR} - \alpha^*_{PTAR} \\
&= \mathbb{E}[R(\gamma^*_{OLS}) A^T] (\Sigma_A^{-1} - \beta_W \Sigma_W^{-1} \beta_W^\top) \mathbb{E}[AR(\gamma^*_{OLS})],
\end{aligned}$$

where we have replaced η with the assumed value of $\mathbb{E}[AR(\gamma^*_{OLS})]$. By assumption, Σ_A is full-rank, so that matrix $\Omega := (\Sigma_A^{-1} - \beta_W \Sigma_W^{-1} \beta_W^\top)$ is positive definite if and only if $\Sigma_A \Omega \Sigma_A$ is positive definite. Working with this representation, we can see that

$$\begin{aligned}
\Sigma_A \Omega \Sigma_A &= \Sigma_A - \Sigma_A \beta_W \Sigma_W^{-1} \beta_W^T \Sigma_A \\
&= \mathbb{E}[AA^\top] - \mathbb{E}[AW^\top] \mathbb{E}[WW^\top]^{-1} \mathbb{E}[WA^\top] \\
&\succ 0,
\end{aligned}$$

where the last line follows from Proposition A3. In the case where $d_W = d_A = 1$, and A has unit variance, then let $\rho_W = \beta_W^2 / (\beta_W^2 + \mathbb{E}[\epsilon_W^2])$, and observe that

$$\alpha^*_{PTAR} = \eta^2 \rho_W \qquad \alpha^*_{TAR} = \eta^2.$$

□

Proposition A6 describes a worst-case mean-shift in A, where η is taken in the direction that maximizes the loss of the OLS solution γ^*_{OLS}. This is also a particularly simple case to analyze for building intuition, because the optimal solution to both Equations (16) and (29) is to take $\gamma = \gamma^*_{OLS}$ and to estimate an intercept term α equal to the bias incurred by the shift in the mean of A. However, the noise in W results in under-estimating the impact of the shift, and the gap to the optimal solution depends on the signal-to-variance relationship in W, which (as discussed in Section 3) is not generally identified.

We also prove that the Cross-Proxy Targeted Anchor Regression objective is equal to that of Targeted Anchor Regression.

Theorem 3. *Under Assumptions 1, 3, and 4, for all* $\gamma \in \mathbb{R}^{d_X}, \alpha \in \mathbb{R}$,

$$\ell_{\times TAR}(W, Z; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha) = \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2]$$

where $\tilde{\mu} := \mathbb{E}_{do(A:=\nu)}[W]$ *is the mean of* W *under intervention, and* $\tilde{\Sigma}_W$ *is the covariance* $\tilde{\Sigma}_W := Cov_{do(A:=\nu)}(W)$.

Proof of Theorem 3. We have

$$\begin{aligned} a_\gamma^\top &= \mathbb{E}[R(\gamma)Z^\top](\mathbb{E}[WZ^\top])^{-1} \\ &= \mathbb{E}[R(\gamma)(A^\top\beta_Z + \epsilon_Z^\top)] \\ &\qquad \mathbb{E}[(\beta_W^\top A + \epsilon_W)(\beta_Z^\top A + \epsilon_Z)^\top]^{-1} \\ &= \mathbb{E}[R(\gamma)A^\top]\beta_Z(\beta_W^\top\mathbb{E}[AA^\top]\beta_Z)^{-1} \\ &= \mathbb{E}[R(\gamma)A^\top](\mathbb{E}[AA^\top])^{-1}(\beta_W^\top)^{-1}, \end{aligned}$$

while

$$\begin{aligned} \tilde{\mu} &= \beta_W^\top\mathbb{E}[\nu] \\ \tilde{\Sigma}_W - \Sigma_W &= \beta_W^\top(\Sigma_\nu - \Sigma_A)\beta_W. \end{aligned}$$

With $b_\gamma^\top := w_\gamma^\top M_A$ and w_γ defined by (23), we have that

$$\begin{aligned} a_\gamma^\top\tilde{\mu} &= b_\gamma^\top\mathbb{E}[\nu] \\ a_\gamma^\top(\tilde{\Sigma}_W - \Sigma_W)a_\gamma &= b_\gamma^\top(\Sigma_\nu - \Sigma_A)b_\gamma, \end{aligned}$$

which is equivalent to $\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha)$ (Definition 4, Equation (16)). The proof is complete by Proposition 2. □

Note that the argument is symmetric for using an observed shift in either Z or W, so it suffices to know the anticipated shift with respect to one proxy.

E. Details for experiments

E.1. Details of Section 5.1

We outline the details of the simulation experiment in Section 5.1.

Summary We simulate a training data set $\mathcal{D}_{\text{train}}$ from a SCM that induces the structure in Figure 2, fix $\lambda := 5$ and fit estimators PAR(W) and xPAR(W, Z). We consider the intervention $\mathbb{P}_{do(A:=\nu)}$ with $\nu = (-2.83, 0.35, 0.71)^\top$, and simulate a test data set $\mathcal{D}_{\text{test}}$ from that distribution. We then compute the intervention mean squared prediction error (MSPE) $\hat{\mathbb{E}}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$ both for PAR(W) and xPAR(W, Z). We repeat this procedure $m = 10^5$ times for several signal-to-variance ratios x (not including 0), and display the quantiles of the losses in Figure 5. We also plot the population losses $\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$ for PAR(W) and xPAR(W, Z), as well as AR(A) and OLS.

Technical details We let $\mathbb{E}[AA^\top] = \beta = \text{Id}$ and $\mathbb{E}[\epsilon_W\epsilon_W^\top] = s^2\,\text{Id}$, such that $W = \beta^\top A + s \cdot \epsilon_W$. Then Ω_W as defined in Equation (11) simplifies to

$$\begin{aligned} \Omega_W &= \mathbb{E}[AA^\top]\beta(\beta^\top\mathbb{E}[AA^\top]\beta + \mathbb{E}[\epsilon_W\epsilon_W^\top])^{-1}\beta^\top\mathbb{E}[AA^\top] \\ &= \frac{1}{1+s^2}\,\text{Id}\,. \end{aligned}$$

We call $x = (1+s^2)^{-1}$ the signal-to-variance ratio, and we can obtain a given signal-to-variance ratio x, by setting $s = \sqrt{(1-x)/x}$.

For each $n \in \{150, 500\}$ and signal-to-variance ratio $x \in \{1/20, 2/20, \ldots, 20/20\}$, we set $s = \sqrt{(1-x)/x}$ and sample a data set $\mathcal{D}^i_{n,s}$ for $i = 1, \ldots, 5000$, each with sample size n, from the structural equations:

$$\begin{aligned} A &:= \epsilon_A \\ W &:= A + s \cdot \epsilon_W \\ Z &:= A + s \cdot \epsilon_Z \\ (Y, X, H) &:= (\text{Id} - B)^{-1}(MA + \epsilon), \end{aligned} \tag{30}$$

where $d_A = d_W = d_Z = d_X = 3$, $d_Y = d_H = 1$. M and B are given by

$$M = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 2 & 1 \\ -1 & 3 & 0 \\ 2 & 2 & -3 \\ 0 & -2 & 2 \end{pmatrix}, B = \begin{pmatrix} 0 & -2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and all noise variables are i.i.d., $\epsilon_A, \epsilon_W, \epsilon_Z, \epsilon \sim \mathcal{N}(0, \text{Id})$. For every combination (n, s) we have 5000 data sets $\mathcal{D}^i_{n,s}$, $i = 1, \ldots, 5000$. For each data set, we compute the proxy estimators $\gamma^i_{n,s,W}$ and $\gamma^i_{n,s,W;Z}$, using one or two proxies respectively, and we simulate 5000 corresponding test data sets of size n from $\mathbb{P}_{do(A:=\nu)}$ (using the structural equations above, except for changing the assignment for A to $A := \nu$). The prediction MSE for the i'th test data set is then $\frac{1}{n}\sum_{j=1}^n (Y_j - \gamma^\top X_j)^2$, resulting in 5000 values of the MSE for each combination of (n, s).

At each combination of (n, s) we plot the median by a line of the estimated worst case losses, and by a shaded region

indicate the interval between the 25% and 75% quantiles of the observed distribution. We plot the median instead of the mean since for small x, $s^2 = \frac{1-x}{x}$ is large, and especially for $\text{WCL}^i_{n,s}(W, Z)$ and $n = 150$, the mean will be driven very much by outliers for small x.

The population versions of losses for any s is computed first by computing the population estimators γ from the parameter matrices M, B, and then computing the loss at ν by $\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2] = w_\gamma^\top M \nu\nu^\top M^\top w_\gamma + w_\gamma^\top \mathbb{E}[\epsilon\epsilon^\top] w_\gamma$.

E.2. Details of Section 5.2

We outline the details of the simulation experiment in Section 5.2.

Summary We analyze the effect of applying anchor regression with one proxy, PAR(W), when the signal-to-variance ratio is potentially misspecified. To do so, we simulate data from the same SCM as in Section 5.1 ($n = 10^4$), and in particular from a range of true (unknown) signal-to-variance ratios $x \in (0, 1]$. To each data set, we apply anchor regression with one proxy, PAR(W), and with $\lambda := 5$. We further assume the signal-to-variance ratio to be 40% – independently of its true value. This means, by Proposition 1, that we assume that PAR(W) minimizes the worst case mean squared prediction error (MSPE) over the region $C := \{\nu\nu^\top \preceq (1 + 0.4 \cdot \lambda)\mathbb{E}[AA^\top]\}$, with the worst case MSPE for being equal to the optimal value of the PAR(W) objective function. If $x = 0.4$, then PAR(W) indeed minimizes the worst case MSPE over C and the estimated worst case MSPE over C is close to the actual worst case MSPE over C. But if $x \neq 0.4$, the estimator minimizes the worst case MSPE over a different set, and then expect that the true worst case MSPE over C differs from its estimate. Figure 6 shows that this is indeed the case: We observe that if the true signal-to-variance ratio is larger than the assumed 40%, our estimate of the MSPE is too conservative. On the contrary, if the true signal-to-variance ratio is smaller than assumed, our estimates of the MSPE over C are too small, meaning that we underestimate the worst case MSPE in the region C.

Technical details For a fixed signal-to-variance ratio x, we simulate a training data set $\mathcal{D}_{\text{train}}$ ($n = 10^4$) from the same procedure as in Section E.1, i.e. using the structural equations in Equation (30), and with the same parameters M and B. We fit the PAR(W) estimator to the data using $\lambda := 5$, and the estimated worst case mean squared prediction error (MSPE) over C is then the value of the objective function in the estimated parameter (by Theorem 1).

To find the actual worst case MSPE over C for a given estimator λ, we use the fact from Equation (24) that

$$\mathbb{E}_{do(A:=v)}[(R - \gamma^\top X)^2] = (b_\gamma^\top v)^2 + w_\gamma^\top w_\gamma, \qquad (31)$$

where we use that $\mathbb{E}[\epsilon\epsilon^\top] = \text{Id}$, w_γ is given by Equation (23) and $b_\gamma^\top = w_\gamma^\top M_A$. The second term doesn't depend on v, and since C is spherical, the worst case MSPE over C is attained in the direction $v \propto b_\gamma$, with v normalized such that $\|v\|^2 = (1 + 0.4 \cdot \lambda)$ (that is v lies on the boundary of C). Using the known M and B, we compute w_γ, b_γ, and the actual worst case MSPE over C is given by Equation (31) plugging in $v = b_\gamma \cdot \sqrt{(1 + 0.4 \cdot \lambda)}/\|b_\gamma\|$.

We compute also the worst case MSPE over C when using an OLS estimator for the prediction. We fit $\hat{\gamma}_{OLS}$ from $\mathcal{D}_{train}$, and, as for the actual MSPE of PAR(W), the worst case MSPE over C using OLS can be computed, by computing vectors $b_{\hat{\gamma}_{OLS}}, w_{\hat{\gamma}_{OLS}}$. Again the worst case MSPE over C using $\hat{\gamma}_{OLS}$ is attained by setting $v = b_{\hat{\gamma}_{OLS}} \cdot \sqrt{(1 + 0.4 \cdot \lambda)}/\|b_{\hat{\gamma}_{OLS}}\|$ and plugging v, $b_{\hat{\gamma}_{OLS}}$ and $w_{\hat{\gamma}_{OLS}}$ into Equation (31).

For every signal-to-variance ratio $x \in \{1/20, \ldots, 20/20\}$, we repeat the procedure $m = 1000$ times, for each computing the estimated and actual MSPEs. In Figure 6 we plot the median MSPE as well as the interval from the 25% quantile to the 75% quantile.

E.3. Details of Section 5.3

We outline the details of the simulation experiment in Section 5.3.

Summary We demonstrate the ability of Proxy Anchor Regression to select invariant predictors, in a synthetic setting where predictors X may contain both causal and anti-causal predictors. We simulate data sets ($n = 10^5$) from a SCM with the structure shown in 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, in this scenario, is that A_2 is measured with significantly more noise than A_1, across both proxies. As a consequence, proxy anchor regression with one proxy, PAR(W), puts more weight on anti-causal features: the noise in W is mistaken for fluctuations in A_2, resulting in $X_{\text{anti-causal}}$ mistakenly appearing invariant to shifts in A_2. In contrast, when two proxies W, Z are available, the estimator xPAR(W, Z) asymptotically equals that of anchor regression with observed anchors, and its regression coefficients puts more weight on the causal predictors; see Figure 7 (bottom).

Technical details With $d_{A_1} = d_{A_2} = d_W = d_W = 6$, $d_{X_{\text{causal}}} = d_{X_{\text{anti-causal}}} = 3$ and $d_Y = 1$, we simulate data from

the SCM in Figure 7 (top) which amounts to simulating from the following structural equations:

$$\begin{aligned} A_1 &:= \epsilon_{A_1} \\ A_2 &:= \epsilon_{A_2} \\ W &:= (A_1, A_2)^\top + (\epsilon_{W,1}, \epsilon_{W,2})^\top \\ Z &:= (A_1, A_2)^\top + (\epsilon_{Z,1}, \epsilon_{Z,2})^\top \\ X_{\text{causal}} &:= M_1 A_1 + \epsilon_{X_{\text{causal}}} \\ Y &:= \gamma_{\text{causal}}^\top X_{\text{causal}} + \epsilon_Y \\ X_2 &:= M_2 A_2 + \gamma_{\text{anti-causal}} Y + \epsilon_{X_{\text{anti-causal}}}. \end{aligned}$$

Here $M_1 \in \mathbb{R}^{d_{X_{\text{causal}}} \times d_{A_1}}$ and $M_2 \in \mathbb{R}^{d_{X_{\text{anti-causal}}} \times d_{A_2}}$ are matrices with 1 in every entry, $\gamma_{\text{causal}} = (1/4, 1/4, 1/4)^\top$ and $\gamma_{\text{anti-causal}} = (4, 4, 4)^\top$ (such that the regression coefficients of Y onto $X_{\text{causal}}, X_{\text{anti-causal}}$ are of similar magnitudes). All noise terms are independent and $\epsilon_{A_1}, \epsilon_{A_2}, \epsilon_{X_{\text{causal}}}, \epsilon_{X_{\text{anti-causal}}}, \epsilon_Y \sim \mathcal{N}(0, \text{Id})$, and $\epsilon_{W,1}, \epsilon_{Z,1} \sim \mathcal{N}(0, \text{Id})$, $\epsilon_{W,2}, \epsilon_{Z,2} \sim \mathcal{N}(0, 3^2 \cdot \text{Id})$.

We simulate a data set $\mathcal{D}$ ($n = 10^5$) from these structural equations, and fit the proxy anchor regression estimators $\gamma(W)$ and $\gamma(W, Z)$ from Section 3. We repeat this $m = 10^4$ times, and display the mean absolute value of the regression coefficients (that is the entries of the vectors $\gamma(W)$ and $\gamma(W, Z)$) in Figure 7 (bottom), as well as the standard deviation of the absolute value of the regression coefficients as error bars.

E.4. Details of Section 5.4

Summary 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}, where γ_{TAR} is targeted to a particular mean and covariance of a random intervention $do(A := \nu)$, 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.

Technical details The entire procedure below produces a prediction MSE for each of three methods and two settings, and we repeat this $m = 10^5$ times, to produce the histograms of MSEs shown in Figure 8.

We simulate a training data set $\mathcal{D}_{\text{train}}$ ($n_{\text{train}} = 10^5$) from the structural equations

$$\begin{aligned} A &:= \epsilon_A \\ (Y, X, H) &:= (\text{Id} - B)^{-1}(MA + \epsilon), \end{aligned}$$

where $d_A = d_X = 2$ and $d_Y = d_H = 1$, $\epsilon_A, \epsilon \sim \mathcal{N}(0, \text{Id})$ and M and B were selected by a simulation resulting in:

$$M = \begin{pmatrix} 2 & 1 \\ 0 & 1 \\ 2 & 2 \\ 0 & 3 \end{pmatrix}, B = \begin{pmatrix} 0 & -0.06 & 0.07 & 0.04 \\ 0.05 & 0 & 0.19 & 0.03 \\ 0.11 & -0.11 & 0 & 0.1 \\ -0.02 & 0.02 & 0.09 & 0 \end{pmatrix}.$$

We consider the target distribution $do(A := \kappa^\top A + \eta)$ where

$$\kappa = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{pmatrix}, \eta = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

and so we fit the targeted AR estimator $(\gamma_{\text{targeted-AR}}, \alpha_{\text{targeted-AR}})$ from Equation (16), where the covariance of the anticipated shift is given by $\Sigma_\nu := \kappa^\top \mathbb{E}[AA^\top]\kappa$, and the mean shift is simply η. We also fit OLS estimates $\gamma_{\text{OLS}}(X, Y)$ and $\gamma_{\text{AR}}(X, Y, A)$ where for AR we select λ such that $(1 + \lambda)$ equals the largest eigenvalue of $\kappa^\top \mathbb{E}[AA^\top]\kappa + \eta\eta^\top$, such that $\mathbb{E}[(\kappa^\top A + \eta)(\kappa^\top A + \eta)^\top] \preceq (1 + \lambda)\mathbb{E}[AA^\top]$.

We then simulate a test data set ($n_{\text{test}} = 10^5$) both from 1) the training distribution (i.e. same simulation procedure as for the training set) or 2) by changing the structural equation for A to $A := \kappa^\top \epsilon_A + \eta$, and keeping all other quantities as for the simulation of training data (i.e. the test distribution is the anticipated distribution). We evaluate the prediction MSE on each of the data sets by $\frac{1}{n_{\text{test}}} \sum_j (Y_j - \gamma^\top X_j)^2$ (including the term $\alpha_{\text{targeted-AR}}$ for the targeted AR).

E.5. Details of Section 6

Features The dataset contains time-stamps as well as season indicators, which we do not use anywhere as features. The remaining features are Dew Point (Celsius Degree), Temperature (Celsius Degree), Humidity (%), Pressure (hPa), Combined wind direction (NE, NW, SE, SW, or CV, indicating calm and variable), Cumulated wind speed (m/s), Hourly precipitation (mm), and Cumulated precipitation (mm).

Data Processing Each city has PM2.5 readings from multiple sites, which we average to get a single reading, and we take a log transformation. For Precipitation (Cumulative) we subtract off the (current hour) precipitation to avoid colinearity. We take a log transformation of the variable for Wind Speed, Precipitation (Hourly) and Precipitation (Cumulative), due to skewness. We drop all rows that contain any missing data.

Proxies (Temperature) We use temperature as our proxy variable, and treat it as unavailable at test time. We construct two synthetic proxies of temperature to serve as W, Z, adding independent Gaussian noise while controlling the signal-to-variance ratio (in the training distribution) at $\text{Var}(A)/\text{Var}(W) = 0.9$. This results in different standard deviations of the Gaussian noise across different environments, because of differences in the training distributions across training seasons and cities. The standard error of the noise varies between 2 and 5 degrees, to maintain the same signal-to-variance ratio.

Training Details (PAR, xPAR) For the distributional robustness approaches described in 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 Proxy Anchor Regression using Temperature directly, there is heterogeneity in the cross-validated choice of λ: In 9 out of 20 scenarios, $\lambda = 40$ is chosen, but in the remaining 11, $\lambda = 0$ is chosen, which is equivalent to OLS. We saw a similar result when the maximum value of λ was 20, and increased the maximum limit to 40 without seeing much difference, so we did not increase it further. Concretely, with λ in [0, 20], there are some scenarios where PAR (TempC) has slightly worse or slightly better MSE (vs. λ in [0, 40]), but the differences are all less than 0.001. The only observable difference in Table 1 when running with λ in [0, 20] is that the "best" performance is -0.040 ($\lambda = 20$), as opposed to -0.041 ($\lambda = 40$) [where lower is better, rounded to nearest 0.001]. For Proxy Anchor Regression using W and for Cross-Proxy Anchor Regression (xPAR) using W, Z together, we use the same values of λ as above, for comparability.

Training Details (PTAR, xPTAR) For the targeted approaches described in Section 4, we use the mean and variance of the temperature in the test distribution to target our predictors, and similarly use the distribution of the proxies when using Proxy Targeted Anchor Regression (PTAR) with W and Cross-Proxy TAR (xPTAR) with W, Z. Note that xPTAR (unlike xPAR) is asymmetric in the proxies, but in this case the proxies are distributed identically.

Benchmarks As described in the main text, our primary benchmark is OLS, trained on the three training seasons, evaluated on the held-out season. We also include two other baselines: First, OLS that has access to temperature during both train and test, which we denote OLS (TempC), and OLS that includes temperature during training, and attempts to estimate a bias term by plugging in the mean (test) value for temperature during prediction.

In Table 2 we give the full results over all 20 scenarios, which includes the 11 scenarios where $\lambda = 0$ is chosen by cross-validation, rendering the PAR and xPAR solutions equivalent to OLS.

Table 2. MSE (lower is better) over 20 scenarios consisting of five cities and four held-out seasons. Average difference to OLS estimator (lower is better) given in the second column, and minimum / maximum difference in remaining columns.

Estimator	Mean	Diff	Min	Max
OLS	0.457			
OLS (TempC)	0.455	-0.002	-0.028	0.026
OLS + Est. Bias	0.474	0.018	-0.072	0.150
PAR (TempC)	0.454	-0.003	-0.041	0.006
PAR (W)	0.454	-0.002	-0.037	0.006
xPAR (W, Z)	0.454	-0.003	-0.039	0.007
PTAR	0.450	-0.007	-0.061	0.002
PTAR (W)	0.452	-0.005	-0.038	0.001
xPTAR (W, Z)	0.450	-0.007	-0.059	0.003

Regularization paths In Figure 12 we have shown how the solution in the "best" scenario differs for Proxy Anchor Regression (PAR) with $\lambda = 40$ versus OLS (i.e., $\lambda = 0$). In Figure 13, we show how the coefficients change in-between these two extremes: for every integer value of λ in [0, 40] we show the difference in the PAR vs. OLS coefficients for each feature. Increasing λ further does not make a significant difference for this particular example.

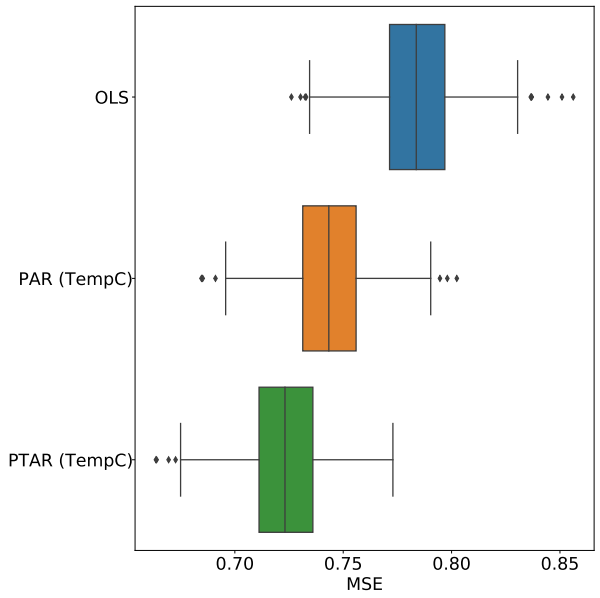


Figure 11. Best performance for Proxy Anchor Regression (PAR) and Proxy Targeted AR (PTAR), corresponding to Summer in Beijing. Variance estimates generated by bootstrapping the test residuals of the fitted models.

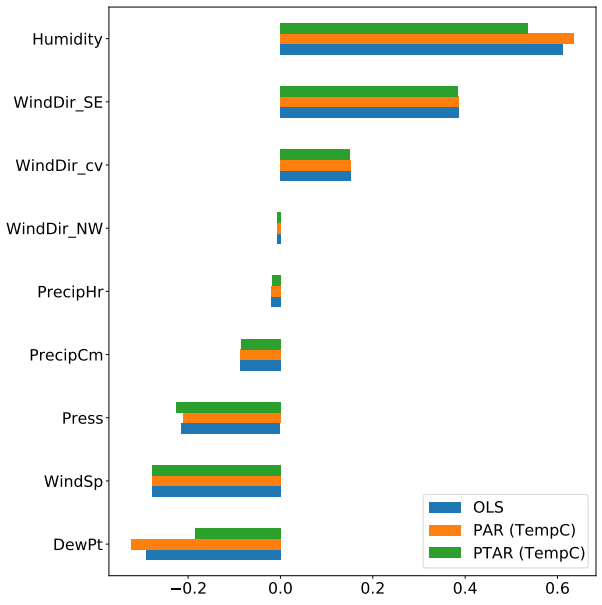


Figure 12. Comparison of learned coefficients. All variables were standardized to unit variance. The intercept for OLS and AR is the same (by construction) at $\alpha = 4.087$ while the intercept for TAR is lower at $\alpha = 3.885$.

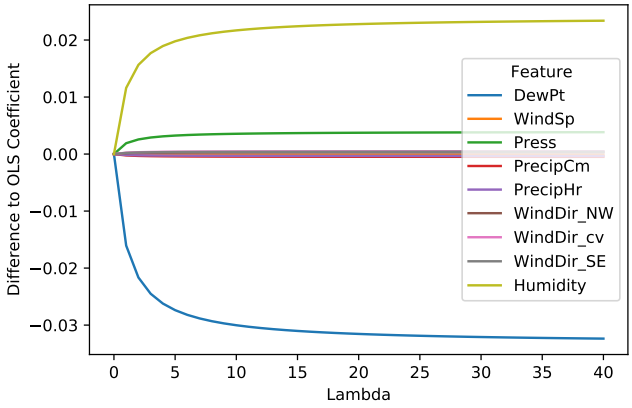


Figure 13. Coefficient path, showing the difference between the PAR and OLS coefficients in Figure 12 for different values of λ.

F. Additional experiment: Signal-to-variance ratio

To examine the effect of the signal strengths β_W and β_Z, we scale the signals $\beta_{W,s} = \beta_{Z,s} = s\,\mathrm{Id}$ for $s \in \{0, \sqrt{2/3}, 0.8\}$, which for the single proxy estimator $\hat{\gamma}_{\text{PAR}}$ amounts to optimizing over worst case loss in the robustness regions $C(\lambda) = \{vv^\top \preceq (1 + \lambda \frac{s^2}{1+s^2})\,\mathrm{Id}\}$.

For $s \in \{1, 3\}$, such that the signal-to-variance ratio $\frac{s^2}{1+s^2}$ equals either 10% or 50%, we simulate a training data set $\mathcal{D}_{\text{train}}$ with two proxies W and Z from the structural equations $A := \epsilon_A, (X^\top, Y^\top, H^\top)^\top := (1 - B)^{-1}(M_a A + \epsilon), W := \beta_{W,s}^\top A + \epsilon_W$ and $Z := \beta_{Z,s}^\top A + \epsilon_Z$ where all noise terms are i.i.d with unit covariance and M_A, B are given by:

$$M := \begin{pmatrix} 2 & 1 \\ 0 & 1 \\ 2 & 2 \\ 0 & 3 \end{pmatrix}, B := \begin{pmatrix} 0 & -0.57 & 0.73 & 0.37 \\ 0.53 & 0 & 1.91 & 0.33 \\ 1.14 & -1.13 & 0 & 0.96 \\ -0.22 & 0.16 & 0.87 & 0 \end{pmatrix}.$$

Since for this experiment we are not interested in finite sample properties of the estimators, we use sample size $n = 10^7$.

For each data set we fit estimators $\hat{\gamma}_{\text{PAR}(W)}$ (using only one proxy), $\hat{\gamma}_{\text{xPAR(W, Z)}}$ (using both proxies), $\hat{\gamma}_{\text{AR}(A)}$, and $\hat{\gamma}_{\text{OLS}}$, and evaluate the estimators at data sampled from interventional distributions $\mathbb{P}_{do(A:=v)}$ for several interventions v of increasing strength (i.e. increasing distance from $\mathbb{E}[A] = 0$).

As the signal to variance ratio increases, the PAR(W) loss approaches the AR(A). Further we observe that xPAR(W, Z) coincides with the $AR(A)$ estimator for both signal-to-variance levels. This is illustrated in Figure 14.

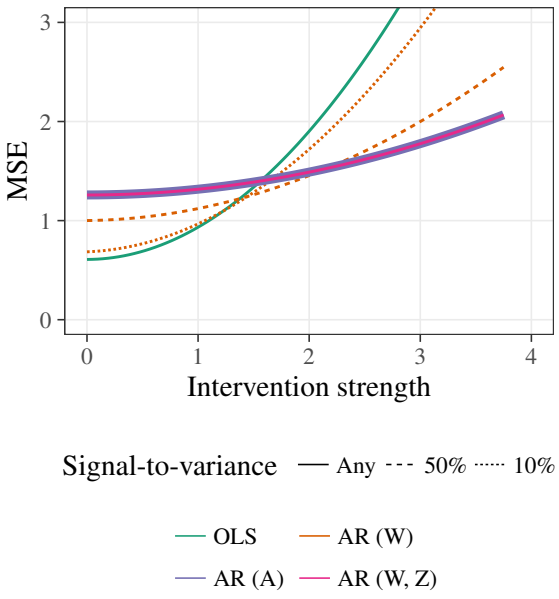


Figure 14. Anchor and proxy estimators for different levels of signal-to-variance ratio $\beta(\mathbb{E}[WW^\top])^{-1}\beta^\top$. A training data set ($n = 10^7$) with two proxies W, Z is simulated and the estimators $\hat{\gamma}_{\text{PAR}(A)}$, $\hat{\gamma}_{\text{xPAR}(W,Z)}$, $\hat{\gamma}_{\text{AR}(A)}$, and $\hat{\gamma}_{\text{OLS}}$ are fitted using a fixed λ. Interventions v of increasing strength is sampled, and for each a new data set ($n = 10^5$) is sampled from $\mathbb{P}^{do(A:=v)}$, and for each estimator $\hat{\gamma}$, the prediction mean squared error $\mathbb{E}_{do(A:=(v_1,v_2))}[(Y - \hat{\gamma}^\top X)^2]$ is computed. This procedure is repeated for signal-to-variance ratios 10% and 50%.