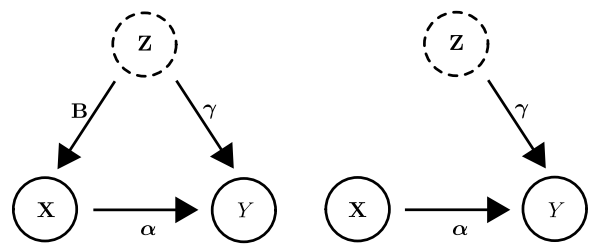
Causal Effect Estimation from Observational and Interventional Data Through Matrix Weighted Linear Estimators

Klaus-Rudolf Kladny[1,2] **Julius von Kügelgen**[2,3] **Bernhard Schölkopf**[1,2] **Michael Muehlebach**[2]

[1]Department of Computer Science, ETH Zürich, Switzerland
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]Department of Engineering, University of Cambridge, United Kingdom
{kkladny, jvk, bs, michaelm}@tue.mpg.de

Abstract

We study causal effect estimation from a mixture of observational and interventional data in a confounded linear regression model with multivariate treatments. We show that the statistical efficiency in terms of expected squared error can be improved by combining estimators arising from both the observational and interventional setting. To this end, we derive methods based on matrix weighted linear estimators and prove that our methods are asymptotically unbiased in the infinite sample limit. This is an important improvement compared to the pooled estimator using the union of interventional and observational data, for which the bias only vanishes if the ratio of observational to interventional data tends to zero. Studies on synthetic data confirm our theoretical findings. In settings where confounding is substantial and the ratio of observational to interventional data is large, our estimators outperform a Stein-type estimator and various other baselines.



(a) observational (b) interventional

Figure 1: **Overview.** We estimate the causal effect $\boldsymbol{\alpha}$ of a continuous multi-variate treatment $\mathbf{X}$ on a scalar outcome Y in a linear Gaussian structural equation model with unobserved confounders $\mathbf{Z}$. Given a large sample from the observational setting in (a) and a small sample from the interventional setting in (b), we establish an optimal, sample-size dependent matrix weighting scheme for combining the biased, low-variance observational estimator with the unbiased, high-variance interventional estimator.

1 INTRODUCTION

Estimating the causal effect of a treatment variable on an outcome of interest is a fundamental scientific problem that is central to disciplines such as econometrics, epidemiology, and social science (Angrist and Pischke, 2009; Morgan and Winship, 2014; Imbens and Rubin, 2015; Hernán and Robins, 2020). A fundamental obstacle to this task is the possibility of hidden confounding: unobserved variables that influence both the treatment and the outcome may introduce additional associations between them (Reichenbach, 1956). As a result, estimators purely based on observational (passively collected) data can be biased and typically do not recover the true causal effect.

This contrasts experimental studies such as randomized controlled trials (RCTs; Neyman, 1923; Fisher, 1936), where the treatment assignment mechanism is modified through an external intervention, thus breaking potential influences of confounders on the treatment. For this reason, RCTs have become the gold standard for causal effect estimation. However, obtaining such interventional data is difficult in practice because the necessary experiments are often infeasible, unethical, or very costly to perform.

In contrast, observational data is usually cheap and abundant, motivating the study of causal inference from observational data (Rubin, 1974; Pearl, 2009). In fact, in certain situations causal effects can be identified and estimated from purely observational data, even under hidden confounding, e.g., in the presence of natural experiments (instrumental variables; Angrist et al., 1996) or observed mediators (front-door adjustment; Pearl, 1995). However, this does not apply to the general case in which a treatment $\mathbf{X}$ and an outcome Y are confounded by an unobserved variable $\mathbf{Z}$ as shown in Fig. 1a.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

In the present work, we study treatment effect estimation in this general setting under the assumption that we have access to both observational and interventional data. The latter can be viewed as sampled from the setting shown in Fig. 1b, where the arrow $\mathbf{Z} \to \mathbf{X}$ has been removed as a result of the intervention on $\mathbf{X}$ (graph surgery; Spirtes et al., 2000), and is thus unbiased for our task. Due to small sample size, however, the estimator based only on interventional data may exhibit high variance. Our main idea is therefore to use the (potentially large amounts of) observational data for variance reduction—at the cost of introducing some bias. This is achieved by forming a combined estimator, which is superior to the purely interventional one in terms of mean squared error.

We make the key assumption that both the treatment $\mathbf{X} \to Y$ and confounding $\mathbf{Z} \to \{\mathbf{X}, Y\}$ effects are linear, but allow for treatment $\mathbf{X}$ and unobserved confounder $\mathbf{Z}$ to be continuous and multi-variate. We then consider a class of estimators of the causal effect parameter vector that combine the unbiased, but high-variance interventional estimator and the biased, but low-variance observational estimator through weight matrices—akin to a multi-variate convex combination. We study the statistical properties of these estimators, establish theoretical optimality results, and investigate their empirical behavior through simulations.

In summary, we highlight the following contributions:

- We introduce a new framework of weighing linear estimators using matrices and show that several existing approaches fall into this category (§ 4).
- We prove that, unlike pooling observational and interventional data (Prop. 4.1), our matrix weighting approaches achieve vanishing mean squared error in the interventional sample limit (Prop. 4.3 and Thm. 4.4) if the ratio between observational and interventional data is non-vanishing.
- We discuss two practical approaches for variance reduction in estimating optimal weight matrices (§ 4.4; Prop. 4.5), and demonstrate through simulations that our estimators outperform baselines and existing methods in situations where confounding is substantial (§ 5).

2 RELATED WORK

Causal reasoning, i.e., inferring a causal query such as a causal effect, can be split up into the tasks of (i) identification and (ii) estimation. Step (i) operates at the population level and seeks to answer whether a causal question can—at least in principle—be answered given infinite data. If the answer is positive and a valid estimand is provided, step (ii) then aims to construct a statistically efficient estimator.

A causal query is identified from a set of assumptions if it can be expressed in terms of the available distributions (e.g., a mixture of different observational and interventional distributions). To this end, Pearl's do-calculus (1995; 2009) provides an axiomatic set of rules for manipulating causal expressions based on graphical criteria. The identification task has been studied extensively (Tian and Pearl, 2002; Pearl and Bareinboim, 2014; Bareinboim and Pearl, 2016) and has by now been solved for many settings of interest: In these cases, the do-calculus—and its extensions (Correa and Bareinboim, 2020)—are sound and complete in that they provide a valid estimand if and only if one exists (Huang and Valtorta, 2006; Shpitser and Pearl, 2006; Bareinboim and Pearl, 2012; Lee et al., 2020).

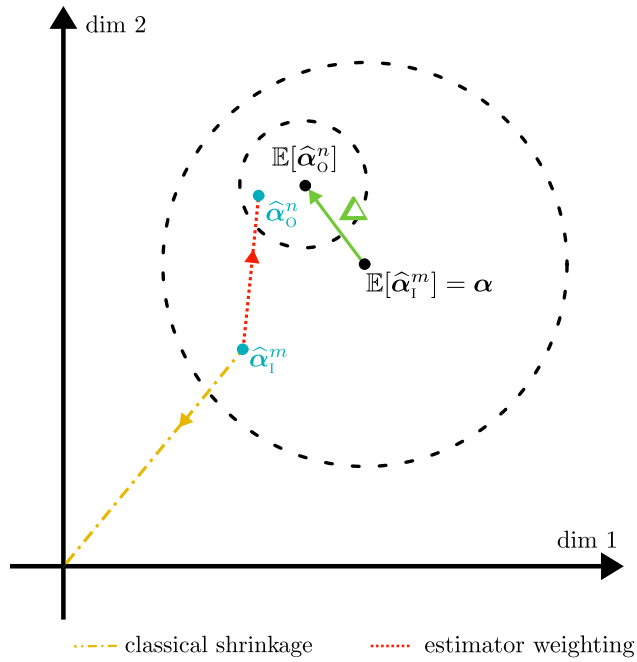


Figure 2: **Relation between shrinkage and estimator weighting in 2D.** Classical shrinkage methods like ridge regression (yellow) shrink the interventional estimator $\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m$ toward the origin. Scalar estimator weighting (red) instead shrinks toward the observational estimator $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n$. Dashed circles show the covariances of $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n$ and $\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m$, here assumed isotropic. $\boldsymbol{\Delta}$ (green) is the confounding-induced bias of $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n$.

In our setting from Fig. 1, the causal effect $\boldsymbol{\alpha}$ is not identifiable from observational data, but is trivially identified by intervening on $\mathbf{X}$. Yet, this leaves open the question of *how to estimate $\boldsymbol{\alpha}$ from finite data in the best possible way*. In contrast to the plethora of works on identification, there is much less prior literature about statistical efficiency of causal parameter estimation, particularly for confounded settings.

A common source of inspiration for both prior work and our approach is that of shrinkage estimation. In light of the bias-variance decomposition of the mean squared error (e.g., Hastie et al., 2009, p. 24), shrinkage can yield a strictly better ("dominating") estimator by reducing variance, at the cost of introducing some bias. These ideas were first introduced in frequentist statistics by Stein (1956); James

and Stein (1961) who showed that the maximum likelihood estimate of a multivariate mean is dominated by shrinking towards a fixed point such as the origin. Similar ideas are also at the heart of empirical Bayes analysis (Robbins, 1964; Efron and Morris, 1973; Efron, 2012). For estimating a parameter vector $\boldsymbol{\alpha}$ in a linear model, as is the focus of the present work, a classical shrinkage method is ridge regression (Hoerl, 1970).

Instead of shrinking towards the origin, an intuitive idea for causal effect estimation is to *shrink towards the observational estimator*. The hope is that the latter constitutes a better attractor if the confounding bias is not too large—despite a slight increase in variance compared to shrinking toward a constant. We refer to this approach as scalar estimator weighting. Fig. 2 shows a visual comparison to classical shrinkage estimation. The most closely related work on estimator weighting is that of Green and Strawderman (1991); Green et al. (2005) and Rosenman et al. (2020). The former two consider general biased and unbiased estimators. The latter propose weighting schemes for estimating vectors of multiple *binary* treatment effects. These works are strongly inspired by James-Stein shrinkage estimators and minimize a generalized version of Stein's unbiased risk estimate (Wasserman, 2006, p. 150). Rosenman et al. (2020) show optimality among scalar weights with respect to minimizing the true risk as the dimensionality of the estimated treatment effects goes to infinity. However, these theoretical results rely on knowledge of the true covariance matrix of the interventional estimator (which is typically unknown in practice), and the behavior of their estimators in the infinite sample limit is not analyzed.

Other work that focuses on combining observational and interventional data to estimate causal effects of *binary* treatments includes, e.g., Kallus et al. (2018); Cheng and Cai (2021); Ilse et al. (2021); Rosenman et al. (2022); Hatt et al. (2022), see Colnet et al. (2020) for a comprehensive survey.

Yang and Ding (2020) also study combining estimators of binary treatment effects. However, in their framework an estimator with less bias in addition to a second error-prone estimator is computed from a second observational "validation set", in which all confounders are measured. Our framework, in contrast, does not require measurements of the confounders.

In the present work, we consider a general linear regression setting with continuous (rather than binary) multi-variate treatments. To combine observational and interventional data, we introduce a new class of matrix (rather than scalar) weighted estimators, of which ridge regression and data pooling are special cases. Instead of employing Stein's unbiased risk estimate, we develop and analyze estimates for the theoretically optimal weight matrix, without making assumptions about the covariance structure of estimators.

Most approaches to causal estimation, including the present work, assume that the causal structure among variables is known and takes the general form of the directed acyclic graph in Fig. 1. For prior work on leveraging observational and interventional data for causal discovery, or structure learning, see, e.g., Wang et al. (2017).

3 SETTING & PRELIMINARIES

Notation. Upper case Y denotes a scalar random variable, lower-case y a scalar, bold lower-case $\mathbf{x}$ a vector, and bold upper-case $\mathbf{X}$ either a matrix or random vector. The spectral norm of a matrix $\mathbf{X}$ is denoted by $\|\mathbf{X}\|_2$.

Causal Model. To formalize our problem setting, we adopt the structural causal model framework of Pearl (2009). Specifically, we assume that the causal relationships between the d-dimensional confounder $\mathbf{Z}$, the p-dimensional treatment $\mathbf{X}$, and the scalar outcome Y are captured by the following linear Gaussian structural equation model (SEM):

$$\mathbf{Z} \leftarrow \mathbf{N}_{\mathbf{Z}}, \qquad \mathbf{N}_{\mathbf{Z}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{N}_{\mathbf{Z}}}, \boldsymbol{\Sigma}_{\mathbf{N}_{\mathbf{Z}}}) \tag{1}$$

$$\mathbf{X} \leftarrow \mathbf{B}\mathbf{Z} + \mathbf{N}_{\mathbf{X}}, \qquad \mathbf{N}_{\mathbf{X}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{N}_{\mathbf{X}}}, \boldsymbol{\Sigma}_{\mathbf{N}_{\mathbf{X}}}) \tag{2}$$

$$Y \leftarrow \mathbf{Z}^\top \boldsymbol{\gamma} + \mathbf{X}^\top \boldsymbol{\alpha} + N_Y, \quad N_Y \sim \mathcal{N}(\mu_{N_Y}, \sigma^2_{N_Y}) \tag{3}$$

with $\mathbf{B} \in \mathbb{R}^{p \times d}$, $\boldsymbol{\gamma} \in \mathbb{R}^d$, $\boldsymbol{\alpha} \in \mathbb{R}^p$, and $(\mathbf{N}_{\mathbf{Z}}, \mathbf{N}_{\mathbf{X}}, N_Y)$ mutually independent exogenous noise variables. The SEM in (1)–(3) induces an observational distribution over $(\mathbf{Z}, \mathbf{X}, Y)$ which is referred to as $\mathbb{P}_{\text{obs}}$, see Fig. 1a.

To model the interventional setting, we consider a soft intervention (Eberhardt and Scheines, 2007), which randomizes the treatment $\mathbf{X}$ by replacing the assignment in (2) with

$$\mathbf{X} \leftarrow \widetilde{\mathbf{N}}_{\mathbf{X}}, \qquad \widetilde{\mathbf{N}}_{\mathbf{X}} \sim \mathbb{P}_{\widetilde{\mathbf{N}}_{\mathbf{X}}}, \tag{4}$$

where $\widetilde{\mathbf{N}}_{\mathbf{X}}$ is mutually independent of $\mathbf{N}_{\mathbf{Z}}$ and N_Y. We note that $\widetilde{\mathbf{N}}_{\mathbf{X}}$ may be non-Gaussian. The modified interventional SEM consisting of (1), (3) and (4) induces a different, interventional distribution over $(\mathbf{Z}, \mathbf{X}, Y)$, which we refer to as $\mathbb{P}_{\text{int}}$, see Fig. 1b.

For ease of notation and for the remainder of this work, we assume without loss of generality that all noise variables are zero-mean. Details on how to extend our method to non zero-mean noise variables are provided in App. D.

Data. We assume access to two separate datasets of observations of $(\mathbf{X}, Y)$ of size n and m, each sampled independently from the observational and interventional distributions (i.i.d.), respectively:

$$(\mathbf{x}_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{obs}}, \quad i = 1, ..., n,$$

$$(\mathbf{x}_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{int}}, \quad i = n+1, ..., n+m,$$

where $\mathbb{P}_{\text{obs}}$ and $\mathbb{P}_{\text{int}}$ denote the distributions of $(\mathbf{X}, Y)$ in the observational and interventional settings, respectively. We note that the confounder $\mathbf{Z}$ remains unobserved. We concatenate the observational sample in a treatment matrix $\mathbf{X}_{\text{O}} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$ and outcome vector $\mathbf{y}_{\text{O}} = (y_1, ..., y_n)^\top \in \mathbb{R}^n$, and similarly with $\mathbf{X}_{\text{I}}, \mathbf{y}_{\text{I}}$ for the interventional sample. Finally, we denote the pooled data by $\mathbf{X}_{\text{P}} = (\mathbf{X}_{\text{O}}, \mathbf{X}_{\text{I}}) \in \mathbb{R}^{(n+m)\times p}$ and $\mathbf{y}_{\text{P}} = (\mathbf{y}_{\text{O}}, \mathbf{y}_{\text{I}}) \in \mathbb{R}^{n+m}$.

Goal. Our objective is to obtain an accurate estimate of the parameter vector $\boldsymbol{\alpha}$, which characterizes the linear causal effect of $\mathbf{X}$ on Y in (3). Formally, it is given by

$$\boldsymbol{\alpha} = \nabla_{\mathbf{x}} \mathbb{E}[Y|\text{do}(\mathbf{X} \leftarrow \mathbf{x})],$$

where the $\text{do}(\cdot)$ operator denotes a manipulation of the treatment assignment akin to (4), and the expectation is taken with respect to the corresponding conditional distribution.

Confounding Issues. In the general case with non-zero $\mathbf{B}$ and $\boldsymbol{\gamma}$, the observational setting is confounded, meaning

$$\mathbb{P}_{\text{obs}}(Y|\mathbf{X} = \mathbf{x}) \neq \mathbb{P}(Y|\text{do}(\mathbf{X} \leftarrow \mathbf{x})) = \mathbb{P}_{\text{int}}(Y|\mathbf{X} = \mathbf{x}),$$

which complicates the use of observational data. Specifically, for our assumed model (1)–(3) the conditional expectation of Y under $\mathbb{P}_{\text{obs}}$ is given by the following perturbed linear model (Ćevid et al., 2020):

$$\mathbb{E}_{\text{obs}}[Y|\mathbf{X} = \mathbf{x}] = (\boldsymbol{\alpha} + \boldsymbol{\Delta})^\top \mathbf{x}, \tag{5}$$

where $\boldsymbol{\Delta} \in \mathbb{R}^p$ denotes the *confounding bias*, which is given explicitly in terms of the model parameters as

$$\boldsymbol{\Delta} = (\boldsymbol{\Sigma}_{\mathbf{N_X}} + \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{N_Z}}\mathbf{B}^\top)^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{N_Z}}\boldsymbol{\gamma}. \tag{6}$$

It can be seen from (6) that the confounding bias $\boldsymbol{\Delta}$ is zero if $\mathbf{B}$ or $\boldsymbol{\gamma}$ are zero (i.e., $\mathbf{Z}$ only affects either $\mathbf{X}$ or Y). Furthermore, we have that, in general,

$$\text{Var}_{\text{obs}}(Y|\mathbf{X}) = \sigma^2_{Y|\mathbf{X}} \neq \sigma^2_{Y|\text{do}(\mathbf{X})} = \text{Var}_{\text{int}}(Y|\mathbf{X}). \tag{7}$$

Assessing Estimator Quality. We rely on mean squared error with respect to the true parameter $\boldsymbol{\alpha}$ as a measure for comparing different estimators.

Definition 3.1 (MSE)**.** *Let $\widehat{\boldsymbol{\alpha}}$ be any function of the pooled data $(\mathbf{X}_{\text{P}}, \mathbf{y}_{\text{P}})$ taking values in $\mathbb{R}^p$. Then*

$$\text{MSE}(\widehat{\boldsymbol{\alpha}}) := \mathbb{E}\left[\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2\right],$$

where the expectation is taken over $\mathbf{X}_{\text{P}}, \mathbf{y}_{\text{P}}$.

We note that the mean squared error can also be written as follows:

$$\text{MSE}(\widehat{\boldsymbol{\alpha}}) = \|\mathbf{Bias}(\widehat{\boldsymbol{\alpha}})\|_2^2 + \text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}})), \tag{8}$$

where

$$\mathbf{Bias}(\widehat{\boldsymbol{\alpha}}) = \mathbb{E}[\widehat{\boldsymbol{\alpha}}] - \boldsymbol{\alpha},$$
$$\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}) = \mathbb{E}[(\widehat{\boldsymbol{\alpha}} - \mathbb{E}[\widehat{\boldsymbol{\alpha}}])(\widehat{\boldsymbol{\alpha}} - \mathbb{E}[\widehat{\boldsymbol{\alpha}}])^\top].$$

This decomposition highlights that biased estimators can dominate unbiased ones through variance reduction.

Pure Estimators. We study estimators for $\boldsymbol{\alpha}$ that are linear combinations of the following ordinary least squares estimators obtained on the two data sets individually.

Definition 3.2 (Pure Estimators)**.** *For non-singular moment matrices $\mathbf{X}_{\text{O}}^\top\mathbf{X}_{\text{O}}$ and $\mathbf{X}_{\text{I}}^\top\mathbf{X}_{\text{I}}$, the* pure estimators *based only on the observational/interventional sample are given by:*

$$\widehat{\boldsymbol{\alpha}}_{\text{O}}^n := (\mathbf{X}_{\text{O}}^\top\mathbf{X}_{\text{O}})^{-1}\mathbf{X}_{\text{O}}^\top\mathbf{y}_{\text{O}},$$
$$\widehat{\boldsymbol{\alpha}}_{\text{I}}^m := (\mathbf{X}_{\text{I}}^\top\mathbf{X}_{\text{I}})^{-1}\mathbf{X}_{\text{I}}^\top\mathbf{y}_{\text{I}}.$$

Recall that $\widehat{\boldsymbol{\alpha}}_{\text{I}}^m$ is unbiased while $\widehat{\boldsymbol{\alpha}}_{\text{O}}^n$ has bias $\boldsymbol{\Delta}$. Their covariances conditionally on $\mathbf{X}_{\text{O}}$ and $\mathbf{X}_{\text{I}}$ are given by

$$\begin{aligned}\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) &= (\mathbf{X}_{\text{O}}^\top\mathbf{X}_{\text{O}})^{-1}\sigma^2_{Y|\mathbf{X}},\\ \mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m) &= (\mathbf{X}_{\text{I}}^\top\mathbf{X}_{\text{I}})^{-1}\sigma^2_{Y|\text{do}(\mathbf{X})}.\end{aligned} \tag{9}$$

Unlike previous work (see § 2), we do not make assumptions about the covariance structure of either estimator.

Almost sure convergence. To analyze the behavior of estimators in the infinite sample limit, we will employ the following characterization known as *almost sure convergence*.

Definition 3.3 (Almost Sure Convergence)**.** *Let $\mathbf{M}$ be a random matrix with realizations in $\mathbb{R}^{p\times p}$. We say a sequence of random matrices $\widehat{\mathbf{M}}_m$ indexed by $m \in \mathbb{N}$* converges almost surely *to $\mathbf{M}$, denoted $\widehat{\mathbf{M}}_m \xrightarrow{a.s.} \mathbf{M}$, if and only if*

$$\lim_{m\to\infty} P\left(\widehat{\mathbf{M}}_m = \mathbf{M}\right) = 1,$$

where P denotes probability.

4 MATRIX WEIGHTED LINEAR ESTIMATORS

We now introduce our class of matrix weighted linear estimators, which combine the two pure estimators from Def. 3.2 using a weight matrix $\mathbf{W}$ to obtain a new (better) estimator.

Definition 4.1 ($\mathbf{W}$-weighted Linear Estimator)**.** *Let $\mathbf{W} \in \mathbb{R}^{p\times p}$ (possibly random). The* $\mathbf{W}$-weighted linear estimator *for $\boldsymbol{\alpha}$ is given by*

$$\widehat{\boldsymbol{\alpha}}_{\mathbf{W}}^m := \mathbf{W}\widehat{\boldsymbol{\alpha}}_{\text{I}}^m + (\mathbf{I}_p - \mathbf{W})\widehat{\boldsymbol{\alpha}}_{\text{O}}^n.$$

We furthermore refer to $\mathbf{W}$ as a weight matrix.

We will generally think of n as a function of m, where we sometimes even explicitly write $n(m)$. However, to simplify notation we index estimators by m only, omitting the dependence $n(m)$.

Note that the purely interventional estimator is a special case of a $\mathbf{W}$-weighted estimator with $\mathbf{W} = \mathbf{I}_p$. However,

while unbiased, it may be subject to high variance if m is very small.[1] Hence, we generally prefer to employ the observational data as well and choose $\mathbf{W} \neq \mathbf{I}_p$.

4.1 EXISTING METHODS AS SPECIAL CASES

First, we show that several standard approaches can be viewed as special cases of matrix-weighted estimators.

Data Pooling. A straightforward approach for combining both data sets is to compute an estimator on the pooled data. The resulting least-squares estimator $\widehat{\boldsymbol{\alpha}}_{\text{P}}^m$ is:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\text{P}}^m &:= (\mathbf{X}_{\text{P}}^\top \mathbf{X}_{\text{P}})^{-1} \mathbf{X}_{\text{P}}^\top \mathbf{y}_{\text{P}} \\ &= (\mathbf{X}_{\text{O}}^\top \mathbf{X}_{\text{O}} + \mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}})^{-1} (\mathbf{X}_{\text{O}}^\top \mathbf{y}_{\text{O}} + \mathbf{X}_{\text{I}}^\top \mathbf{y}_{\text{I}}) \\ &= \mathbf{W}_{\text{P}}^m \widehat{\boldsymbol{\alpha}}_{\text{I}}^m + (\mathbf{I} - \mathbf{W}_{\text{P}}^m) \widehat{\boldsymbol{\alpha}}_{\text{O}}^n,\end{aligned} \tag{10}$$

where

$$\mathbf{W}_{\text{P}}^m := (\mathbf{X}_{\text{O}}^\top \mathbf{X}_{\text{O}} + \mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}})^{-1} \mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}}. \tag{11}$$

We see that $\widehat{\boldsymbol{\alpha}}_{\text{P}}^m$ indeed qualifies as a valid matrix weighted estimator in the sense of Def. 4.1.

However, data pooling can lead to highly undesirable limiting behavior in cases where the amount of observational data $n(m)$ does not vanish in the limit of infinite interventional data $m \to \infty$. An example for this is given in the following proposition.

Proposition 4.1. *Let* $\lim_{m\to\infty} \frac{n(m)}{m} = c$ *for some* $c > 0$ *and* $\boldsymbol{\Delta} \neq \mathbf{0}$. *Then, it holds that*

$$\lim_{m\to\infty} \text{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\text{P}}^m\right) > 0.$$

The proof of Prop. 4.1 is provided in App. A.1. We note, however, that this does not happen for a vanishing amount of observational data, that is $\lim_{m\to\infty} \frac{n(m)}{m} = 0$ (see Prop. 4.2 in App. A.2).

Ridge Regression. The ridge regression estimator on the interventional data, which shrinks $\widehat{\boldsymbol{\alpha}}_{\text{I}}^m$ towards the origin (see § 2 and Fig. 2), is given by

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\text{ridge}}^m &= (\mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_{\text{I}}^\top \mathbf{y}_{\text{I}} \\ &= \widehat{\mathbf{W}}_{\text{ridge}}^m \widehat{\boldsymbol{\alpha}}_{\text{I}}^m + (\mathbf{I}_p - \widehat{\mathbf{W}}_{\text{ridge}}^m)\mathbf{0},\end{aligned}$$

where

$$\widehat{\mathbf{W}}_{\text{ridge}}^m := (\mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}}. \tag{12}$$

Hence, $\widehat{\boldsymbol{\alpha}}_{\text{ridge}}^m$ can also be seen as a special case of a matrix weighted estimator with no observational data and $\widehat{\boldsymbol{\alpha}}_{\text{O}}^n = \mathbf{0}$. Further, comparing (11) and (12) suggest an interpretation of ridge regression as a *poor man's* data pooling since access to observational data is replaced by a positive definite data matrix $\lambda \mathbf{I}_p$. However, λ is a constant, and therefore $\lim_{m\to\infty} \text{MSE}(\widehat{\boldsymbol{\alpha}}_{\text{ridge}}^m) = 0$ even in the setting of Prop. 4.1, which contrasts data pooling.

[1] E.g., consider a one-dimensional setting with $x_i = 1$ if i is even and -1 otherwise. Then, for odd m, $\text{Var}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m | \mathbf{X}_{\text{I}}) \propto (\sum_i x_i^2)^{-1} = \frac{1}{m}$.

4.2 OPTIMAL WEIGHTING SCHEMES

We now establish theoretically optimal weighting schemes that minimize the mean squared error of $\mathbf{W}$-weighted linear estimators $\widehat{\boldsymbol{\alpha}}_{\mathbf{W}}^m$ for different classes of weight matrices $\mathbf{W}$ by exploiting the specific structure of our problem setting (§ 3).

Optimal Scalar Weight. First, we consider the special case of scalar estimator weighting by considering weight matrices of the form $\mathbf{W} = w\mathbf{I}_p$ with weight $w \in [0, 1]$. The optimal scalar weight is then derived as follows:

$$\begin{aligned}&\frac{\partial}{\partial w} \text{MSE}\left(\widehat{\boldsymbol{\alpha}}_{w\mathbf{I}_p}^m\right) \overset{!}{=} 0 \\ \overset{(8)}{\iff}\ &\frac{\partial}{\partial w}\left(\left\|\mathbb{E}[\widehat{\boldsymbol{\alpha}}_{w\mathbf{I}_p}^m - \boldsymbol{\alpha}]\right\|_2^2 + \text{Tr}\left(\mathbf{Cov}\left(\widehat{\boldsymbol{\alpha}}_{w\mathbf{I}_p}^m\right)\right)\right) \overset{!}{=} 0 \\ \implies\ & w_*^m = \frac{\text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n)) + \|\boldsymbol{\Delta}\|_2^2}{\text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m)) + \text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n)) + \|\boldsymbol{\Delta}\|_2^2}.\end{aligned}$$

Optimal Diagonal Weight Matrix. A more general case is to weigh each dimension individually by different scalars $w^{(k)} \in [0, 1]$, $k = 1, ..., p$, corresponding to a weight matrix of the form $\mathbf{W} = \text{diag}(\mathbf{w})$. The optimal diagonal weighting $\text{diag}(\mathbf{w}_*^m)$ is then given by

$$w_*^{m(k)} = \frac{\text{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) + \Delta^{(k)\,2}}{\text{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m) + \text{Cov}^{(k,k)}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) + \Delta^{(k)\,2}},$$

for $k = 1, ..., p$. The derivation is analogous to that for the optimal scalar weight above, with the only difference being that we optimize over each dimension separately.

Optimal Weight Matrix. Finally, we can also determine the optimum weighting as follows:

$$\begin{aligned}\mathbf{W}_*^m = &\left(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right) \\ &\left(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m) + \mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right)^{-1}.\end{aligned} \tag{13}$$

A thorough derivation of the proposed weighting schemes can be found in App. C. In addition, we elaborate on how this weighting scheme handles sample imbalance in App. E.

Remark 4.2. *If (i)* $\boldsymbol{\Delta} = \mathbf{0}$ *and (ii)* $\sigma_{Y|\mathbf{X}}^2 = \sigma_{Y|do(\mathbf{X})}^2$, *then* $\mathbf{W}_*^m = \mathbf{W}_{\text{P}}^m$, *i.e., data pooling corresponds to weighing with the optimal weight matrix under these two assumptions.*

Remark 4.2 can be verified by simplifying (13) with assumptions (i) and (ii) and comparing to (11). It agrees

with our intuition: Ordinary least squares relies on the assumption that $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}_i] = \boldsymbol{\alpha}^\top \mathbf{x}_i$ with equal variance, for all i. Thus, data pooling recovers the optimal estimator if these assumptions are true, i.e., the two conditional distributions $\mathbb{P}_{\text{obs}}(Y|X)$ and $\mathbb{P}_{\text{int}}(Y|\text{do}(X))$ are identical. However, in general, they will not be identical and data pooling then amounts to model misspecification. This is likely to result in a non-vanishing mean squared error for $m \to \infty$ as highlighted in Prop. 4.1.

4.3 PRACTICAL ESTIMATORS

Unfortunately, the optimal weighting derived in (13) cannot be implemented directly, since the quantities $\boldsymbol{\Delta}$, $\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n)$, and $\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m)$ are unknown in practice. To construct practical estimators informed by our theoretical insights, one option is thus to rely on plug-in estimates of these unknown quantities. For $\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m)$ and $\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n)$, we use the standard estimators

$$\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m) = (\mathbf{X}_{\text{I}}^\top \mathbf{X}_{\text{I}})^{-1} \widehat{\sigma}^2_{Y|\text{do}(\mathbf{X})},$$
$$\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) = (\mathbf{X}_{\text{O}}^\top \mathbf{X}_{\text{O}})^{-1} \widehat{\sigma}^2_{Y|\mathbf{X}},$$

which replace the conditional variances in (9) by

$$\widehat{\sigma}^2_{Y|\text{do}(\mathbf{X})} = \frac{1}{m-1} \|\mathbf{y}_{\text{I}} - \mathbf{X}_{\text{I}} \widehat{\boldsymbol{\alpha}}_{\text{I}}^m\|_2^2,$$
$$\widehat{\sigma}^2_{Y|\mathbf{X}} = \frac{1}{n-1} \|\mathbf{y}_{\text{O}} - \mathbf{X}_{\text{O}} \widehat{\boldsymbol{\alpha}}_{\text{O}}^n\|_2^2.$$

For $\boldsymbol{\Delta}$, one may consider using the unbiased estimator

$$\widehat{\boldsymbol{\Delta}}_m = \widehat{\boldsymbol{\alpha}}_{\text{O}}^n - \widehat{\boldsymbol{\alpha}}_{\text{I}}^m. \tag{14}$$

Substituting these into (13) then yields:

$$\widehat{\mathbf{W}}_*^m = \left(\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) + \widehat{\boldsymbol{\Delta}}_m \widehat{\boldsymbol{\Delta}}_m^\top + \epsilon \mathbf{I}_p\right) \left(\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m) + \widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n) + \widehat{\boldsymbol{\Delta}}_m \widehat{\boldsymbol{\Delta}}_m^\top + \epsilon \mathbf{I}_p\right)^{-1}. \tag{15}$$

The regularization with $\epsilon > 0$ ensures that the inverse remains stable even in the large sample limit where $\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n)$ and $\widehat{\mathbf{Cov}}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m)$ tend to zero. The reason for instability without such regularization is that $\mathbf{W}_*^m$ is not uniquely defined in the infinite sample limit. With regularization, however, we can guarantee that $\widehat{\mathbf{W}}_*^m$ converges to $\mathbf{I}_p$ almost surely.

Proposition 4.3 (Weight Matrix Convergence)**.** *Let* $\lim_{m\to\infty} \frac{n(m)}{m} = c$*, for some constant* $c > 0$*. Then,* $\widehat{\mathbf{W}}_*^m$ *from* (15) *converges almost surely to* $\mathbf{I}_p$*, i.e.,* $\widehat{\mathbf{W}}_*^m \xrightarrow{a.s.} \mathbf{I}_p$*.*

The proof for Prop. 4.3 is included in App. A.3. We can show that this convergence implies that the mean squared error vanishes asymptotically.

Theorem 4.4 (Zero Mean Squared Error in the Sample Limit)**.** *Let* $\widehat{\mathbf{W}}^m$ *be any sequence of random weight matrices such that* $\widehat{\mathbf{W}}^m \xrightarrow{a.s.} \mathbf{I}_p$ *and* $\lim_{m\to\infty} \frac{n(m)}{m} = c$ *for some constant* $c > 0$*. Then,*

$$\lim_{m\to\infty} \text{MSE}\left(\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m}\right) = 0,$$

where $\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m}$ *denotes the matrix-weighted linear estimator with weight matrix* $\widehat{\mathbf{W}}^m$*, as defined in Def. 4.1.*

The proof of Thm. 4.4 is included in App. A.4.

Thm. 4.4 has the following relevant implication: we can incorporate an arbitrarily large amount of biased observational data and are still guaranteed that the bias (and also variance) of $\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m_*}$ will vanish in the infinite sample limit. Moreover, this guarantee is independent of $\boldsymbol{\Delta}$ and $|\sigma^2_{Y|\mathbf{X}} - \sigma^2_{Y|\text{do}(\mathbf{X})}|$.

We also note that Thm. 4.4 does not imply unbiasedness of $\widehat{\boldsymbol{\alpha}}^m_{\widehat{\mathbf{W}}^m_*}$ for any finite sample size.

Further, we note that almost sure convergence of $\widehat{\mathbf{W}}^m$ to $\mathbf{I}_p$ may generally not be the only option to achieve vanishing mean squared error. For example, if $\boldsymbol{\Delta} = \mathbf{0}$ such that $\widehat{\boldsymbol{\alpha}}_{\text{O}}^n$ is unbiased, we also obtain vanishing mean squared error for almost sure convergence of $\widehat{\mathbf{W}}^m$ to $\mathbf{0}$.

4.4 SUITABLE INDUCTIVE BIASES

Despite the desirable performance established in Thm. 4.4, the plug-in estimates from § 4.3 will often not perform very well in finite sample settings. The main issue is the estimation of $\boldsymbol{\Delta}$, which has a large variance when done according to (14). To see this, we first note that

$$\text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\Delta}}_m)) = \text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m)) + \text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{O}}^n)), \tag{16}$$

since the observational and interventional data are independent. Now, if we only have a small interventional sample (as is typically the case), $\text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\alpha}}_{\text{I}}^m))$ and hence according to (16) also $\text{Tr}(\mathbf{Cov}(\widehat{\boldsymbol{\Delta}}_m))$ will be large.

We therefore explore possible inductive biases in the form of additional assumptions on the type of confounding that lead to reduced variance when estimating $\widehat{\boldsymbol{\Delta}}_m$. These inductive biases can be motivated from domain knowledge and validation techniques such as cross-validation (Schaffer, 1993). Specifically, the application itself may provide some prior knowledge about the nature of confounding, which can then be confirmed by a better validation score compared to the other inductive biases/methods proposed here.

To this end, we observe that (14) can be written as the solution of the following two-step ordinary least squares

procedure:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n} &\leftarrow \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\left\{\|\mathbf{y}_{\mathrm{O}} - \mathbf{X}_{\mathrm{O}}\boldsymbol{\alpha}\|_2^2\right\}\\ \mathbf{r} &\leftarrow \mathbf{y}_{\mathrm{I}} - \mathbf{X}_{\mathrm{I}}\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}\\ \widehat{\boldsymbol{\Delta}}_m &\leftarrow \arg\min_{\boldsymbol{\Delta}\in\mathbb{R}^p}\left\{\|\mathbf{r} + \mathbf{X}_{\mathrm{I}}\boldsymbol{\Delta}\|_2^2\right\}. \qquad (17)\end{aligned}$$

Small $\|\boldsymbol{\Delta}\|_2$. In some settings, we may be willing to assume that, despite the existence of unobserved confounders, the resulting confounding bias is rather weak, i.e., that its Euclidean norm $\|\boldsymbol{\Delta}\|_2$ is small. Since this is precisely the assumption underlying ridge regression, we reformulate (17) using a regularizer $\lambda_{\ell^2} > 0$ as

$$\widehat{\boldsymbol{\Delta}}_m^{\ell^2} \leftarrow \arg\min_{\boldsymbol{\Delta}\in\mathbb{R}^p}\left\{\|\mathbf{r} + \mathbf{X}_{\mathrm{I}}\boldsymbol{\Delta}\|_2^2 + \lambda_{\ell^2}\|\boldsymbol{\Delta}\|_2^2\right\},$$

for which a closed-form solution of the same computational complexity as least squares exists. We refer to the weight matrix estimate obtained by using $\widehat{\boldsymbol{\Delta}}_m^{\ell^2}$ in place of $\widehat{\boldsymbol{\Delta}}_m$ in (15) as $\widehat{\mathbf{W}}_{\ell^2}^m$. By Prop. 4.5, we still obtain the same limiting guarantees of Thm. 4.4 for $\widehat{\mathbf{W}}_{\ell^2}^m$, as long as λ_{ℓ^2} is fixed (λ_{ℓ^2} is independent of m, $\mathbf{X}_{\mathrm{P}}$, $\mathbf{y}_{\mathrm{P}}$).

Proposition 4.5. *Let* $\lim_{m\to\infty}\frac{n(m)}{m} = c$ *and* $\lambda_{\ell^2} > 0$ *be fixed. Then,*

$$\lim_{m\to\infty} \mathrm{MSE}\left(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathbf{W}}_{\ell^2}^m}^m\right) = 0.$$

The proof for Proposition 4.5 is given in App. A.5.

Small $\|\boldsymbol{\Delta}\|_0$. In other settings, we may have prior beliefs that only some treatment variables X_i are confounded, i.e., that the number of nonzero elements of $\boldsymbol{\Delta}$, denoted by $\|\boldsymbol{\Delta}\|_0$, is small. If we are unaware of which treatments are confounded, but p is small, we can simply fit all 2^p possible models or use best subset selection (James et al., 2013, p. 205). For larger p, a more efficient technique known as the LASSO employs ℓ^1-regularization and has become a standard tool (Tibshirani, 1996). For the LASSO, approximate optimization techniques exist that have a computational complexity of $\mathcal{O}(p^2 n)$ (Efron et al., 2004), which is of the same order as ordinary least squares. In this case, we reformulate (17) as

$$\widehat{\boldsymbol{\Delta}}_m^{\ell^1} \leftarrow \arg\min_{\boldsymbol{\Delta}\in\mathbb{R}^p}\left\{\|\mathbf{r} + \mathbf{X}_{\mathrm{I}}\boldsymbol{\Delta}\|_2^2 + \lambda_{\ell^1}\|\boldsymbol{\Delta}\|_1\right\},$$

for some $\lambda_{\ell^1} > 0$, and where $\|\cdot\|_1$ denotes the ℓ^1-norm. We refer to the weight matrix obtained by using $\widehat{\boldsymbol{\Delta}}_m^{\ell^1}$ in place of $\widehat{\boldsymbol{\Delta}}_m$ in (15) as $\widehat{\mathbf{W}}_{\ell^1}^m$.

5 EXPERIMENTS

We investigate the empirical behavior of our proposed matrix weighted estimators in a finite sample setting and compare them with baselines and existing methods through simulations on synthetic data.[2] To this end, we consider different experimental settings in which we vary the strength and sparsity of confounding, as well as the ratio and absolute quantity of observational and interventional data.

Compared Methods. We report the mean squared error attained by the theoretically optimal weight matrix $\mathbf{W}_*^m$ from (13) as an oracle, as well as the plug-in estimator $\widehat{\mathbf{W}}_*^m$ thereof from (15), and the regularized regression-based $\widehat{\mathbf{W}}_{\ell^2}^m$ and $\widehat{\mathbf{W}}_{\ell^1}^m$ from § 4.4. For the latter two, we choose the regularization hyperparameters λ_{ℓ^2} and λ_{ℓ^1} by cross-validation on the interventional data. As baselines, we consider only using interventional data ($\mathbf{W}_{\mathrm{I}}^m = \mathbf{I}_p$) and data pooling according to $\mathbf{W}_{\mathrm{P}}^m$ from (11). We also compare to the Rosenman et al. (2020) scalar weighting scheme which was proposed for vectors of binary treatment effects and is given by $\mathbf{W} = \widehat{w}_{\mathrm{rm}}^m \mathbf{I}_p$ with

$$\widehat{w}_{\mathrm{rm}}^m := \max\left\{1 - \frac{\mathrm{Tr}\left(\widehat{\mathbf{Cov}}\left(\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m\right)\right)}{\|\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^m - \widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^n\|_2^2},\ 0\right\}.$$

We emphasize that other commonly used methods for causal effect estimation from observational data such as propensity score matching (Imai and Dyk, 2004) are not applicable, because they require the relevant confounders to be observed, which is not the case in our setting.

General Setup. In all experiments, we use $p = 30$ treatments, a one-dimensional ($d = 1$) confounder Z, and unit/isotropic (co)variances: $\sigma_{N_Y}^2 = \sigma_{N_Z}^2 = 1$, $\boldsymbol{\Sigma}_{\mathbf{N_X}} = \mathbf{I}_p$. We sample $\tilde{\mathbf{N}}_{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{Cov}(\mathbf{X}_{\mathrm{O}}))$, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I}_p)$, and choose $\mathbf{b}$ and γ depending on the settings described below. Unless otherwise specified, we then draw $m = 300$ interventional and $n = 600$ observational examples from $\mathbb{P}_{\mathrm{int}}$ and $\mathbb{P}_{\mathrm{obs}}$, respectively, and compute estimates of $\boldsymbol{\alpha}$ using the different weighting approaches. We repeat this procedure 1000 times and report the resulting mean and standard deviation of the mean squared error.

Different Types of Confounding. In our main experiment, we investigate how estimators perform under different types of confounding encoded by (2) and (3), specifically by the parameters $\mathbf{b} \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}$ (for a scalar confounder Z). For *spread* confounding, we sample $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ such that the confounder affects all treatment variables almost surely. For *sparse* confounding, we sample $b^{(k)} \sim \mathcal{N}(0, 1)$ for $k = 1, ..., 5$, and $b^{(k)} = 0$ otherwise, such that only the first five treatments are confounded. In both cases, we investigate $\gamma \in \{1, 5\}$ which controls the strength of $Z \to Y$ and thus the extent to which $\boldsymbol{\Delta} = 0$ is violated.

[2] The source code for all experiments is available at: https://github.com/rudolfwilliam/matrix_weighted_linear_estimators

Table 1: **Mean squared error for the causal effect parameter α using various weighting schemes for different types of confounding.** The standard plug-in optimal weight matrix estimator $\widehat{\mathbf{W}}_*^m$ generally does not perform well, while $\widehat{\mathbf{W}}_{\ell^2}^m$ and $\widehat{\mathbf{W}}_{\ell^1}^m$, which benefit from prior knowledge, outperform prior work. Note that $\mathbf{W}_*^m$ is an oracle that is generally not computable in practice. Numbers correspond to mean $\pm$ std. dev. over 1000 runs; the best method is highlighted in bold.

		$\widehat{w}_{\text{rm}}^m$	$\mathbf{W}_{\text{I}}^m$	$\mathbf{W}_{\text{P}}^m$	$\widehat{\mathbf{W}}_*^m$	$\widehat{\mathbf{W}}_{\ell^1}^m$	$\widehat{\mathbf{W}}_{\ell^2}^m$	$\mathbf{W}_*^m$
spread conf.	$\gamma = 1$	$\mathbf{0.07} \pm 0.02$	0.21 ± 0.06	$\mathbf{0.07} \pm 0.01$	0.21 ± 0.06	0.10 ± 0.04	0.08 ± 0.03	0.04 ± 0.01
	$\gamma = 5$	0.89 ± 0.20	2.79 ± 0.78	0.92 ± 0.14	2.77 ± 0.77	1.11 ± 0.42	$\mathbf{0.76} \pm 0.29$	0.10 ± 0.03
sparse conf.	$\gamma = 1$	0.12 ± 0.02	0.21 ± 0.06	0.13 ± 0.02	0.21 ± 0.06	$\mathbf{0.10} \pm 0.04$	0.16 ± 0.05	0.05 ± 0.01
	$\gamma = 5$	1.80 ± 0.37	2.79 ± 0.78	2.42 ± 0.24	2.77 ± 0.77	$\mathbf{0.95} \pm 0.48$	2.28 ± 0.63	0.30 ± 0.08

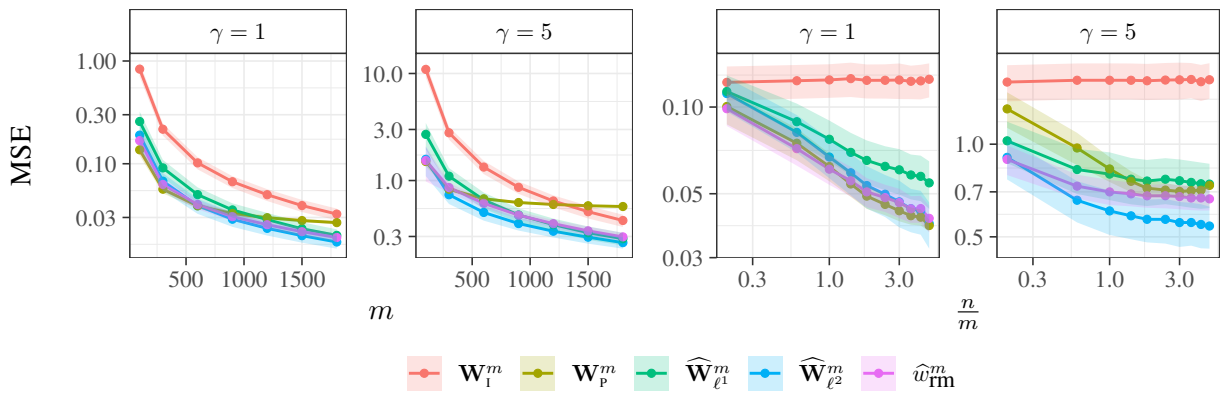


Figure 3: **Performance for varying dataset sizes and ratios.** *(Left)* All methods improve as the amount of data is increased. More sophisticated weighting schemes outperform the purely interventional ($\mathbf{W}_{\text{I}}^m$) and plug in estimators ($\widehat{\mathbf{W}}_*^m$) (not depicted due to close performance overlap with pure interventional), whereas data pooling ($\mathbf{W}_{\text{P}}^m$) works well only for small m and γ. *(Right)* When keeping m fixed and adding more observational data, $\widehat{\mathbf{W}}_{\ell^2}^m$ clearly works best in strongly confounded ($\gamma = 5$) settings. MSE and $\frac{n}{m}$ are plotted on a $\log_{10}$ scale. Shaded areas indicate ± 0.5 standard deviations.

Main Results. The results are presented in Tab. 1. We find that our regularized estimators generally perform well, particularly when the underlying assumptions are satisfied: under sparse confounding $\widehat{\mathbf{W}}_{\ell^1}^m$ works best, and in the spread confounding case $\widehat{\mathbf{W}}_{\ell^2}^m$ is only narrowly outperformed by $\widehat{w}_{\text{rm}}^m$ and $\mathbf{W}_{\text{P}}^m$ when $\gamma = 1$. Data pooling works relatively well when $\gamma = 1$ (compared to $\gamma = 5$) where the violation of the identically distributed assumption is weak and the variance from estimating unknown quantities is not compensated by the bias reduction. In contrast, both the purely interventional approach $\mathbf{W}_{\text{I}}^m$ and the plug-in estimator $\widehat{\mathbf{W}}_*^m$ do not perform very well in this finite sample setting due to high variance, as explained in § 4.4.

Varying Data Set Sizes and Ratios. In Fig. 3, we investigate how the different estimators behave across different data set sizes and ratios for the spread confounding setting. In the left two plots, we vary the amount of interventional data m while fixing the amount of observational data to $n = 3m$. The results confirm our theoretical results: For small data set sizes, data pooling is a worthwhile alternative to more sophisticated weights, in particular if the violation against the assumption of identical distribution is minor ($\gamma = 1$). However, for large enough data set sizes, the approaches from both previous work and ours achieve a better score. Particularly, we see that $\widehat{\mathbf{W}}_{\ell^2}^m$ outperforms all other weights in both scenarios for large enough data sets.

In the right two plots, we keep $m = 500$ fixed and change n and thus the ratio of interventional to observational data. Unsurprisingly, we find that the mean squared error of $\mathbf{W}_{\text{I}}^m$ remains constant. For strong confounding ($\gamma = 5$), we see that $\widehat{\mathbf{W}}_{\ell^2}^m$ adapts best with a considerable margin: Unlike $\widehat{w}_{\text{rm}}^m$, it explicitly takes into account (an estimate of) the covariance structure of $\widehat{\boldsymbol{\alpha}}_{\text{O}}^n$ in constructing the weight matrix.

6 DISCUSSION

Connection to Transfer Learning. Our setting bears resemblance to transfer and multi-task learning (Thrun, 1995; Caruana, 1997), specifically to supervised domain adaptation, which aims to leverage knowledge from a source do-

main to improve a model in a target domain, for which typically much less data is available. In our case, we aim to use the source model $\widehat{\boldsymbol{\alpha}}_{\mathrm{O}}^{n}$, learned by estimating $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$ in the observational setting, to improve our (high-variance) target model $\widehat{\boldsymbol{\alpha}}_{\mathrm{I}}^{m}$ of $\mathbb{E}[Y|\mathrm{do}(\mathbf{X}\leftarrow\mathbf{x})]$. Transfer learning can only work if the domains are sufficiently similar, resulting in numerous approaches leveraging different assumptions about shared components (Quiñonero-Candela et al., 2008). These assumptions are often phrased in causal terms (Schölkopf et al., 2012; Zhang et al., 2013; Gong et al., 2016; Rojas-Carulla et al., 2018). Similarly, our observational (source) and interventional (target) domains share the same causal model and only differ in the treatment assignment mechanisms (2) and (4). Still, the bias in (6) can in theory be arbitrary large, and our methods from § 4.4 implicitly rely on it being small or sparse.

Beyond Linear Regression. Some of our derivations and theoretical results rely on the fact that the confounding bias in (5) is linear in $\mathbf{x}$. For the class of *linear* SCMs (1)–(3), Gaussianity is necessary and sufficient[3] for this condition to hold, but it may also hold for more general classes of SCMs. For binary treatments $\mathbf{X}\in\{0,1\}^{p}$, in particular, it is always possible to write the difference between the biased and unbiased average treatment effect estimates using a constant offset $\boldsymbol{\Delta}$ akin to (14), irrespective of the confounding relationship.[4] Future work may thus investigate nonlinear extensions, e.g., by drawing inspiration from semiparametrics (Robins and Rotnitzky, 1995), doubly robust estimation (Bang and Robins, 2005), and debiased machine learning (Chernozhukov et al., 2018).

Incorporating Covariates. Our current formulation does not *explicitly* account for observed confounders, or pretreatment covariates, which need to be adjusted for in the observational setting to avoid introducing further bias. In principle, such covariates can simply be included in $\mathbf{X}$, as different treatment components X_i are allowed to be dependent. However, this may result in high-dimensional treatments and thus render full randomization in (4) unrealistic. Other covariates, while unproblematic with regard to bias, may help further reduce variance (Henckel et al., 2022). Extending our framework to incorporate different types of covariates is thus a worthwhile future direction.

7 CONCLUSION

In the present work, we have introduced a new class of matrix weighted linear estimators for learning causal effects of continuous treatments from finite observational and interventional data. Here, our focus has been on optimizing statistical efficiency, which complements the vast causal inference literature on identification from heterogeneous data. Our estimators are connected to classical ideas from shrinkage estimation applied to causal learning and provide a unifying account of data pooling and ridge regression, which emerge as special cases. We show that our estimators are theoretically grounded and compare favorably to baselines and prior work in simulations. While we restricted our analysis to linear models for now, we hope that the insights and methods developed here will also be useful for a broader class of causal models and transfer learning problems.

[3]Note $\mathbb{E}[Y|\mathbf{X}]=\boldsymbol{\gamma}^{\top}\mathbb{E}[\mathbf{Z}|\mathbf{X}]+\boldsymbol{\alpha}^{\top}\mathbf{X}$ and $\mathbb{E}[\mathbf{Z}|\mathbf{X}]$ is linear in $\mathbf{X}$ only in the Gaussian case (Peters et al., 2017, Thm. 4.2).

[4]Specifically, we have $\boldsymbol{\Delta}=\mathbb{E}[Y|\mathbf{X}=\mathbf{1}]-\mathbb{E}[Y|\mathbf{X}=\mathbf{0}]-(\mathbb{E}[Y|\mathrm{do}(\mathbf{X}\leftarrow\mathbf{1})]-\mathbb{E}[Y|\mathrm{do}(\mathbf{X}\leftarrow\mathbf{0})])$.

Acknowledgements

We thank the anonymous reviewers for useful comments and suggestions that helped improve the manuscript.

The authors thank the Branco Weiss Fellowship, administered by ETH Zurich, for the support. This work was further supported by the Tübingen AI Center and by the German Research Foundation (DFG) under Germany's excellence strategy – EXC number 2064/1 – project number 390727645.

References

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2009. [Cited on page 1.]

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. [Cited on page 1.]

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005. [Cited on page 9.]

E. Bareinboim and J. Pearl. Causal Inference by Surrogate Experiments: z-Identifiability. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 113–120, 2012. [Cited on page 2.]

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. [Cited on page 2.]

R. Caruana. Multitask Learning. *Machine Learning*, 28(1): 41–75, 1997. [Cited on page 8.]

D. Ćevid, P. Bühlmann, and N. Meinshausen. Spectral Deconfounding via Perturbed Sparse Linear Models. *The Journal of Machine Learning Research*, 21(1):9442–9482, 2020. [Cited on page 4.]

D. Cheng and T. Cai. Adaptive Combination of Randomized and Observational Data. *arXiv:2111.15012*, 2021. [Cited on page 3.]

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1), 2018. [Cited on page 9.]

B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv:2011.08047*, 2020. [Cited on page 3.]

J. Correa and E. Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020. [Cited on page 2.]

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007. [Cited on page 3.]

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012. [Cited on page 3.]

B. Efron and C. Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973. [Cited on page 3.]

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2), 2004. [Cited on page 7.]

R. A. Fisher. Design of experiments. *British Medical Journal*, 1(3923):554, 1936. [Cited on page 1.]

M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848, 2016. [Cited on page 9.]

E. J. Green and W. E. Strawderman. A James-Stein Type Estimator for Combining Unbiased and Possibly Biased Estimators. *Journal of the American Statistical Association*, 86(416):1001–1006, 1991. [Cited on page 3.]

E. J. Green, W. E. Strawderman, R. L. Amateis, and G. A. Reams. Improved Estimation for Multiple Means with Heterogeneous Variances. *Forest Science*, 51(1):1–6, 2005. [Cited on page 3.]

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. [Cited on page 2.]

T. Hatt, J. Berrevoets, A. Curth, S. Feuerriegel, and M. van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv:2202.12891*, 2022. [Cited on page 3.]

L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B*, 84(2):579–599, 2022. [Cited on page 9.]

M. A. Hernán and J. M. Robins. *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC, 2020. [Cited on page 1.]

A. E. Hoerl. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970. [Cited on page 3.]

Y. Huang and M. Valtorta. Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, pages 1149–1154, 2006. [Cited on page 2.]

M. Ilse, P. Forré, M. Welling, and J. M. Mooij. Combining Interventional and Observational Data Using Causal Reductions. *arXiv:2103.04786*, pages 1–42, 2021. [Cited on page 3.]

K. Imai and D. A. V. Dyk. Causal Inference with General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association*, 99(467): 854–866, 2004. [Cited on page 7.]

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015. [Cited on page 1.]

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013. [Cited on page 7.]

W. James and C. Stein. Estimation with Quadratic Loss. In *Proceedings of the 4th Berkeley Symposium on Probability and Statistics*. Berkeley, CA: University of California Press, 1961. [Cited on page 2.]

N. Kallus, A. M. Puli, and U. Shalit. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 31, 2018. [Cited on page 3.]

S. Lee, J. D. Correa, and E. Bareinboim. General Identifiability with Arbitrary Surrogate Experiments. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, pages 389–398, 2020. [Cited on page 2.]

S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2014. [Cited on page 1.]

J. Neyman. On the application of probability theory to agricultural experiments: essay on principles. *Statistical Science*, 5:465–480, 1923. [Cited on page 1.]

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. [Cited on pages 1 and 2.]

J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2nd edition, 2009. [Cited on pages 1, 2, and 3.]

J. Pearl and E. Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595, 2014. [Cited on page 2.]

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017. [Cited on page 9.]

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2008. [Cited on page 9.]

H. Reichenbach. *The Direction of Time*, volume 65. University of California Press, 1956. [Cited on page 1.]

H. Robbins. The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35(1):1–20, 1964. [Cited on page 3.]

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. [Cited on page 9.]

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant Models for Causal Transfer Learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018. [Cited on page 9.]

E. Rosenman, G. Basse, A. Owen, and M. Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 2020. [Cited on pages 3 and 7.]

E. T. Rosenman, A. B. Owen, M. Baiocchi, and H. R. Banack. Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine*, 41(1):65–86, 2022. [Cited on page 3.]

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688–701, 1974. [Cited on page 1.]

C. Schaffer. Selecting a Classification Method by Cross-Validation. *Machine Learning*, 13:135–143, 1993. [Cited on page 6.]

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On Causal and Anticausal Learning. In *International Conference on Machine Learning*, 2012. [Cited on page 9.]

I. Shpitser and J. Pearl. Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, pages 1219–1226, 2006. [Cited on page 2.]

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000. [Cited on page 2.]

C. Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 197–207. University of California Press, 1956. [Cited on page 2.]

S. Thrun. Is Learning The n-th Thing Any Easier Than Learning The First? *Advances in Neural Information Processing Systems*, 8, 1995. [Cited on page 8.]

J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 567–573, 2002. [Cited on page 2.]

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. [Cited on page 7.]

Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation-based Causal Inference Algorithms with Interventions. *Advances in Neural Information Processing Systems*, 30, 2017. [Cited on page 3.]

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006. [Cited on page 3.]

S. Yang and P. Ding. Combining Multiple Observational Data Sources to Estimate Causal Effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020. [Cited on page 3.]

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain Adaptation under Target and Conditional Shift. In *International Conference on Machine Learning*, pages 819–827, 2013. [Cited on page 9.]