Estimating Treatment Effects with Observed Confounders and Mediators

Shantanu Gupta, Zachary C. Lipton, David Childers

Carnegie Mellon University
shantang@cs.cmu.edu, zlipton@cmu.edu, dchilder@andrew.cmu.edu

Abstract

Given a causal graph, the do-calculus can express treatment effects as functionals of the observational joint distribution that can be estimated empirically. Sometimes the do-calculus identifies multiple valid formulae, prompting us to compare the statistical properties of the corresponding estimators. For example, the backdoor formula applies when all confounders are observed and the frontdoor formula applies when an observed mediator transmits the causal effect. In this paper, we investigate the over-identified scenario where both confounders and mediators are observed, rendering both estimators valid. Addressing the linear Gaussian causal model, we show that either estimator can dominate the other by an unbounded constant factor depending on the model parameters. Next, we derive an optimal estimator, which leverages all observed variables to strictly outperform the backdoor and frontdoor estimators. We also present a procedure for combining two datasets, one with observed confounders and another with observed mediators. Finally, we evaluate our methods on both simulated data and the IHDP and JTPA datasets.

1 Introduction

Causal effects are not, in general, identifiable from observational data alone. The fundamental insight of causal inference is that given structural assumptions on the data generating process, causal effects may become expressible as functionals of the joint distribution over observed variables. The do-calculus, introduced by Pearl [1995], provides a set of three rules that can be used to convert causal quantities into such functionals. We are motivated by the observation that, for some causal graphs, treatment effects may be overidentified. Here, applications of the do-calculus produce distinct functionals, all of which, subject to positivity conditions, yield consistent estimators of the same causal effect. Consider a causal graph (see Figure 1) for which the treatment X, mediator M, confounder W, and outcome Y are all observable. Using the backdoor adjustment, we can expresses the average treatment effect of X on Y as a function of $P(X, W, Y)$, while the frontdoor adjustment expresses that same causal quantity via $P(X, M, Y)$ [Pearl, 1995]. Faced with the (fortunate) condition of overidentification, our focus shifts from identification—*is our effect estimable?*, to optimality—*which among multiple valid estimators dominates from a standpoint of statistical efficiency?*

In this paper, we address this very graph, focusing our analysis on the linear causal model [Wright, 1934], a central object of study in causal inference and econometrics and also explore the semiparametric setting. We focus on this graph because the frontdoor estimator is a canonical example of a novel identification result derived using graphical models. It is central in causality literature [Pearl and Mackenzie, 2018, Imbens, 2019] and is a natural first step in the study of overidentified causal models. Deriving the finite sample variance of the backdoor and frontdoor estimators, and precisely characterizing conditions under which each dominates, we find that either may outperform the other to an arbitrary degree depending on the underlying model parameters. These expressions can provide

Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020.

guidance to practitioners for assessing the suitability of each estimator. For example, one byproduct of our analysis is to characterize what properties make for the "ideal mediator". Moreover, in the data collection phase, if one has a choice between collecting data on the mediator or the confounder, these expressions, together with the practitioner's beliefs about likely ranges for model parameters, can be used to decide what data to collect.

Next, we propose techniques that leverage both observed confounders and mediators. For the setting where we *simultaneously* observe both the confounder and the mediator, we introduce an estimator that optimally combines all information. We prove theoretically that this method achieves lower mean squared error (MSE) than both the backdoor and frontdoor estimators, for all settings of the underlying model parameters. Moreover, the extent to which this estimator can dominate the better of the backdoor and frontdoor estimators is unbounded. Subsequently, we consider the partially-observed setting in which two datasets are available, one with observed confounders (but not mediators) $\{(X, W, Y)\}_{i=1}^{n}$, and another with observed mediators (but not confounders) $\{(X, M, Y)\}_{i=1}^{m}$. Interestingly, the likelihood is convex given simultaneous observations but non-convex under partially-observed data. We introduce an estimator that is guaranteed to achieve higher likelihood than either the backdoor or frontdoor estimators. Finally, we evaluate our methods on synthetic, semi-synthetic, and real datasets.

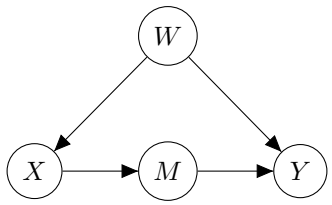

Figure 1: Causal graph with observed mediator and confounder. The backdoor and frontdoor estimators are both applicable.

Our principal contributions are the following:

1. Derivation of the parameter regimes where either of the frontdoor and backdoor estimators dominate vis-a-vis sample efficiency.
2. Demonstration of strict (and unbounded) improvements of the optimal (combined) estimator over *both* the frontdoor and backdoor estimators.
3. Adaptation of a semi-parametric estimator to our graph, showing the benefits of our approach in non-linear settings.
4. Analysis for the partially observed case, where mediators and confounders are observed separately (but never simultaneously).

2 Related Work

The backdoor adjustment formalizes the common practice of controlling for known confounders and is widely applied in statistics and econometrics [Pearl, 2009, 2010, Perković et al., 2015]. The frontdoor adjustment, which leverages observed mediators to identify causal effects even amid unobserved confounding, has seen increasing application in real-world datasets [Bellemare and Bloem, 2019, Glynn and Kashin, 2018, 2017, Chinco and Mayer, 2016, Cohen and Malloy, 2014].

In the most similar work to ours, Glynn and Kashin [2018] compare the frontdoor and backdoor adjustments, computing bias (but not variance) formulas for each and performing sensitivity analysis. Exploring a real-world job training dataset, they demonstrate that the frontdoor estimator outperforms its backdoor counterpart (in terms of bias). The finite sample variance of the frontdoor estimator for the linear Gaussian case was previously derived by Kuroki [2000]. Henckel et al. [2019] introduce a graphical criterion for comparing the asymptotic variances of adjustment sets for the backdoor criterion in linear causal models. Rotnitzky and Smucler [2019] extend this work, showing that the same graphical criterion is valid for non-parametric causal models. They also present a semi-parametric efficient estimator that exploits the conditional independencies in a causal graph. Bhattacharya et al. [2020] provide semi-parametric influence functions for a large class of graphs with hidden variables and the front-door graph becomes a special case of their results.

Researchers have also worked to generalize the frontdoor criterion. Bareinboim et al. [2019] introduce the conditional frontdoor criterion, allowing for both treatment-mediator confounders and mediator-outcome confounders. Fulcher et al. [2020] propose a method for including observed confounders along with a mediator with discrete treatments.

The study of overidentified models dates at least back to Koopmans and Reiersøl [1950]. Sargan [1958], Hansen [1982] formalized the result that in the presence of overidentification, multiple estimators can be combined to improve efficiency. This was extended to the non-parametric setting by Chen and Santos [2018]. A related line of work considers methods for combining multiple datasets for causal inference. Bareinboim and Pearl [2016] study the problem of handling biases while combining heterogeneous datasets, while Jackson et al. [2009] present Bayesian methods for combining datasets with different covariates and some common covariates.

3 Preliminaries

In this work, we work within the structural causal model (SCM) framework due to Pearl [2009], formalizing causal relationships via directed acyclic graphs (DAGs). Each $X \to Y$ edge in this DAG indicates that the variable X is (potentially) a direct cause of variable Y. All measured variables are deterministic functions of their parents and a set of jointly independent per-variable noise terms.

Linear Gaussian SCM In linear Gaussian SCMs, each variable is assumed to be a linear function of its parents. The noise terms are assumed to be additive and Gaussian. In this paper, the finite sample results are derived for the linear Gaussian SCM for the overidentified confounder-mediator graph (Figure 1), where the structural equations can be written as

$$\begin{aligned} w_i &= u_i^w, & u_i^w &\sim \mathcal{N}(0, \sigma_{u_w}^2) \\ x_i &= dw_i + u_i^x, & u_i^x &\sim \mathcal{N}(0, \sigma_{u_x}^2) \\ m_i &= cx_i + u_i^m, & u_i^m &\sim \mathcal{N}(0, \sigma_{u_m}^2) \\ y_i &= am_i + bw_i + u_i^y, & u_i^y &\sim \mathcal{N}(0, \sigma_{u_y}^2). \end{aligned} \tag{1}$$

Here, w_i, x_i, m_i, and y_i are realized values of the random variables W, X, M, Y, respectively, and $u_i^w, u_i^x, u_i^m, u_i^y$ are realized values of the corresponding noise terms. The zero mean assumption in Eq. 1 simplifies analysis, but is not necessary for the results presented in this paper.

3.1 The Backdoor and Frontdoor Adjustments

The effect of a treatment X is expressible by reference to the post-intervention distributions of the outcome Y for different values of the treatment $X = x$. An intervention $do(X = x)$ in a causal graph can be expressed via the *mutilated graph* that results from deleting all incoming arrows to X, setting X's value to $X = x$ for all instances, while keeping the SCM otherwise identical. This distribution is denoted as $P(Y|do(X = x))$.

The backdoor and frontdoor adjustments [Pearl, 2009] express treatment effects as functionals of the observational distribution. Consider our running example of the causal model in Figure 1. We denote X as the treatment, Y as the outcome, W as a confounder, and M as a mediator. Our goal is to estimate the causal quantity $P(Y|do(X = x))$.

Backdoor Adjustment When all confounders of both X and Y are observed—in our example, W—then the causal effect of X on Y, i.e., $P(Y|do(X = x))$ can be written as

$$P(Y|do(X = x)) = \sum_w P(Y|X = x, W = w)P(W = w). \tag{2}$$

Frontdoor Adjustment This technique applies even when the confounder W is unobserved. Here we require access to a mediator M that (i) is observed; (ii) transmits the entire causal effect from X to Y; and (iii) is not influenced by the confounder W given X. The effect of X on Y is computed in two stages. We first find the effect of X on M, then the effect of M on Y as:

$$P(M = m|do(X = x)) = P(M = m|X = x) \tag{3}$$

$$P(Y|do(M = m)) = \sum_x P(Y|M = m, X = x)P(X = x). \tag{4}$$

We can then write the causal effect of X on Y as $P(Y|do(X = x)) = \sum_m P(M = m|do(X = x))P(Y|do(M = m))$.

4 Variance of Backdoor & Frontdoor Estimators

In this section, we analyze the backdoor and frontdoor estimators and characterize the regimes where each dominates. We work with the linear SCM described in Eq. 1. Throughout, our goal is to estimate the causal effect of X on Y. In terms of the underlying parameters of the linear SCM, the quantity that we wish to estimate is ac. Absent measurement error, both estimators are unbiased (see proof in Appendix C) and thus we focus our comparison on their respective variances.

Variance of the Backdoor Estimator The backdoor estimator requires only that we observe $\{X, Y, W\}$ (but not necessarily the mediator M). Say we observe the samples $\{x_i, y_i, w_i\}_{i=1}^n$. We can estimate the causal effect ac by taking the coefficient on X in an OLS regression of Y on $\{X, W\}$. This controls for the confounder W and corresponds naturally to the adjustment described in Eq. 2.

The finite sample and asymptotic variances of the backdoor estimator are (see proof in Appendix D.1)

$$\text{Var}(\widehat{ac})_{\text{backdoor}} = \frac{a^2\sigma_{u_m}^2 + \sigma_{u_y}^2}{(n-3)\sigma_{u_x}^2}, \lim_{n\to\infty} \text{Var}(\sqrt{n}(\widehat{ac} - ac))_{\text{backdoor}} = \frac{a^2\sigma_{u_m}^2 + \sigma_{u_y}^2}{\sigma_{u_x}^2}. \tag{5}$$

Variance of the Frontdoor Estimator The frontdoor estimator is used when $\{X, Y, M\}$ samples are observed. Say we observe the samples $\{x_i, y_i, m_i\}_{i=1}^n$. First, we estimate c by taking the coefficient on X in an OLS regression of M on X. Let the estimate be $\widehat{c}$. This corresponds to the adjustment in Eq. 3. Then, we estimate a by taking the coefficient on M in an OLS regression of Y on $\{M, X\}$. Let the estimate be $\widehat{a}_f$. This corresponds to the adjustment in Eq. 4.

The finite sample variances of $\widehat{c}$ and $\widehat{a}_f$ are (see proof in Appendix D.2)

$$\text{Var}(\widehat{c}) = \frac{\sigma_{u_m}^2}{(n-2)(d^2\sigma_{u_w}^2 + \sigma_{u_x}^2)}, \ \text{Var}(\widehat{a}_f) = \frac{b^2\sigma_{u_w}^2\sigma_{u_x}^2 + \sigma_{u_y}^2(d^2\sigma_{u_w}^2 + \sigma_{u_x}^2)}{(n-3)(d^2\sigma_{u_w}^2 + \sigma_{u_x}^2)\sigma_{u_m}^2}. \tag{6}$$

Using the facts that $\text{Cov}(\widehat{a}_f, \widehat{c}) = 0$ and $\text{Cov}(\widehat{a}_f^2, \widehat{c}^2) = \text{Var}(\widehat{a}_f)\text{Var}(\widehat{c})$, the finite sample variance of the frontdoor estimator is (see proof in Appendix D.2.4)

$$\text{Var}(\widehat{a}_f\widehat{c}) = c^2\text{Var}(\widehat{a}_f) + a^2\text{Var}(\widehat{c}) + 2\text{Var}(\widehat{a}_f)\text{Var}(\widehat{c}). \tag{7}$$

And the asymptotic variance, which does not require Gaussianity, is (see proof in Appendix D.2.5)

$$\lim_{n\to\infty} \text{Var}(\sqrt{n}\widehat{a}_f\widehat{c}) = \frac{c^2(b^2\sigma_{u_w}^2\sigma_{u_x}^2 + \sigma_{u_y}^2(d^2\sigma_{u_w}^2 + \sigma_{u_x}^2))}{(d^2\sigma_{u_w}^2 + \sigma_{u_x}^2)\sigma_{u_m}^2} + \frac{a^2\sigma_{u_m}^2}{d^2\sigma_{u_w}^2 + \sigma_{u_x}^2}. \tag{8}$$

The Ideal Frontdoor Mediator A natural question then arises: what properties of a mediator make the frontdoor estimator most precise? We can see that $\text{Var}(\widehat{a}_f\widehat{c})$ is non-monotonic in the mediator noise σ_{u_m}. Eq. 7 provides us with guidance. $\text{Var}(\widehat{a}_f\widehat{c})$ is a convex function of $\sigma_{u_m}^2$. The *ideal mediator* will have noise variance $\sigma_{u_m}^{2*}$ which minimizes Eq. 7. That is,

$$\sigma_{u_m}^{2*} = \arg\min_{\sigma_{u_m}^2} [\text{Var}(\widehat{a}_f\widehat{c})] \implies \sigma_{u_m}^{2*} = \frac{|c|\sqrt{b^2\sigma_{u_w}^2\sigma_{u_x}^2 + \sigma_{u_y}^2(d^2\sigma_{u_w}^2 + \sigma_{u_x}^2)}}{|a|}\sqrt{\frac{n-2}{n-3}}.$$

Comparison of Backdoor and Frontdoor Estimators The relative performance of the backdoor and frontdoor estimators depend on the underlying SCM's parameters. Using Eqs. 5 and 7, the ratio of the backdoor to frontdoor variance is

$$R_{\text{Var}} = \frac{\text{Var}(\widehat{ac})_{\text{backdoor}}}{\text{Var}(\widehat{a}_f\widehat{c})} = \frac{(n-2)\sigma_{u_m}^2 D^2(a^2\sigma_{u_m}^2 + \sigma_{u_y}^2)}{\sigma_{u_x}^2((n-3)a^2\sigma_{u_m}^4 D + (2\sigma_{u_m}^2 + c^2(n-2)D)E)}, \tag{9}$$

where $D = (d^2\sigma_{u_w}^2 + \sigma_{u_x}^2)$ and $E = (b^2\sigma_{u_w}^2\sigma_{u_x}^2 + \sigma_{u_y}^2 D)$. The backdoor estimator dominates when $R_{\text{Var}} < 1$ and vice versa when $R_{\text{Var}} > 1$. Note that there exist parameters that cause any value of $R_{\text{Var}} > 0$. In particular, as $\sigma_{u_x}^2 \to 0$, $R_{\text{Var}} \to \infty$ and as $\sigma_{u_x}^2 \to \infty$, $R_{\text{Var}} \to 0$, regardless of the sample size n. Thus, either estimator can dominate the other by any arbitrary constant factor.

5 Combining Mediators & Confounders

We now consider optimal strategies for estimating treatment effects in the overidentified regime when both the confounder and mediator are observed simultaneously. Say we observe n samples $\{x_i, y_i, w_i, m_i\}_{i=1}^n$. We show that the maximum likelihood estimator (MLE) is strictly better than the backdoor and frontdoor estimators. The MLE will be optimal (in terms of asymptotic variance) since our model satisfies the necessary regularity conditions (by virtue of being linear and Gaussian) [Le Cam, 1990].

Let the vector $\mathbf{s_i} = [x_i, y_i, w_i, m_i]$ denote the i^{th} sample. Since the data is multivariate Gaussian, the log-likelihood of the data is $\mathcal{LL} = -\frac{n}{2}\left[\log\left(\det \mathbf{\Sigma}\right) + \text{Tr}\left(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\right)\right]$, where $\mathbf{\Sigma} = \text{Cov}([X, Y, W, M])$ and $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{s_i}\mathbf{s_i}^\top$. The MLE for a Gaussian graphical model is $\mathbf{\Sigma}^{\text{MLE}} = \widehat{\mathbf{\Sigma}}$ [Uhler, 2019]. Let the MLE estimates for parameters c and a be $\widehat{c}$ and $\widehat{a}_c$, respectively. Then

$$\widehat{c} = \frac{\widehat{\mathbf{\Sigma}}_{1,4}}{\widehat{\mathbf{\Sigma}}_{1,1}}, \;\; \widehat{a}_c = \frac{\widehat{\mathbf{\Sigma}}_{1,4}\widehat{\mathbf{\Sigma}}_{3,3} - \widehat{\mathbf{\Sigma}}_{1,3}\widehat{\mathbf{\Sigma}}_{3,4}}{\widehat{\mathbf{\Sigma}}_{3,3}\widehat{\mathbf{\Sigma}}_{4,4} - \widehat{\mathbf{\Sigma}}_{3,4}^2}. \tag{10}$$

The MLE estimate for c in Eq. 10 is the same as for the frontdoor—the coefficient of X in an OLS regression of M on X. The MLE estimate for a in Eq. 10 is the coefficient of M in an OLS regression of Y on $\{M, W\}$. The finite sample variance of $\widehat{a}_c$ is (see proof in Appendix D.3.1)

$$\text{Var}(\widehat{a}_c) = \frac{\sigma_{u_y}^2}{(n-3)(c^2\sigma_{u_x}^2 + \sigma_{u_m}^2)}. \tag{11}$$

The variance of $\widehat{c}$ is the same as the frontdoor case as in Eq. 6. Let $r_1 = \sqrt{\frac{n-3}{n-5}}$, $r_2 = \sqrt{\frac{3(n-2)}{n-4}}$, and $L = \left(\frac{c^2\sigma_{u_y}^2}{c^2\sigma_{u_x}^2+\sigma_{u_m}^2} + \frac{a^2\sigma_{u_m}^2}{d^2\sigma_{u_w}^2+\sigma_{u_x}^2}\right)$. We can bound the finite sample variance of the combined estimator as (see proof in Appendix D.3.2)

$$\frac{L}{n} \leq \text{Var}(\widehat{a}_c\widehat{c}) \leq c^2\text{Var}(\widehat{a}_c) + a^2\text{Var}(\widehat{c}) + r_1\left(2|c|\text{Var}(\widehat{a}_c)\sqrt{\text{Var}(\widehat{c})} + r_2\text{Var}(\widehat{a}_c)\text{Var}(\widehat{c})\right). \tag{12}$$

And the asymptotic variance, which does not require Gaussianity, is (see proof in Appendix D.3.3)

$$\lim_{n\to\infty} \text{Var}(\sqrt{n}\widehat{a}_c\widehat{c}) = L. \tag{13}$$

The Ideal Mediator Just as with the frontdoor estimator, we can ask what makes for an *ideal mediator* in this case. Eq. 13 shows that $\lim_{n\to\infty}\text{Var}(\sqrt{n}\widehat{a}_c\widehat{c})$ is a convex function of $\sigma_{u_m}^2$. The ideal mediator will have noise variance $\sigma_{u_m}^{2*}$ which minimizes the variance in Eq. 13. This means that

$$\sigma_{u_m}^{2*} = \arg\min_{\sigma_{u_m}^2}\left[\lim_{n\to\infty}\text{Var}(\sqrt{n}\widehat{a}_c\widehat{c})\right] \implies \sigma_{u_m}^{2*} = \max\left\{0, \frac{|c|\sigma_{u_y}\sqrt{d^2\sigma_{u_w}^2+\sigma_{u_x}^2}}{|a|} - c^2\sigma_{u_x}^2\right\}.$$

5.1 Comparison with Backdoor and Frontdoor Estimators

We can compare Eqs. 5 and 13 to see that, asymptotically, the combined estimator has lower variance than the backdoor estimator for all values of model parameters. That is, as $n \to \infty$, $\text{Var}(\sqrt{n}\widehat{a}_c\widehat{c}) \leq \text{Var}(\sqrt{n}\widehat{ac})_{\text{backdoor}}$. Similarly, we can compare Eqs. 8 and 13 to see that, asymptotically, the combined estimator is always better than the frontdoor estimator for all values of model parameters. That is, as $n \to \infty$, $\text{Var}(\sqrt{n}\widehat{a}_c\widehat{c}) \leq \text{Var}(\sqrt{n}\widehat{a}_f\widehat{c})$.

In the finite sample case, using Eqs. 5 and 12, we can see that for a large enough n, the combined estimator will dominate the backdoor estimator for all model parameters. That is, $\exists N$, s.t., $\forall n > N, \text{Var}(\widehat{a}_c\widehat{c}) \leq \text{Var}(\widehat{ac})_{\text{backdoor}}$, where the dependence of N on the model parameters is stated in Appendix E.1. As an example, for parameter values $a = b = c = d = \sigma_{u_w} = \sigma_{u_x} = \sigma_{u_m} = \sigma_{u_y} = 1$, we have $N = 10$. We can make a similar argument for the dominance of the combined estimator over the frontdoor estimator. Using Eqs 7 and 12, it can be shown that $\exists N$, s.t., $\forall n > N, \text{Var}(\widehat{a}_c\widehat{c}) \leq \text{Var}(\widehat{a}_f\widehat{c})$, where the dependence of N on the model parameters is stated in Appendix E.2.

Next, we show that the combined estimator can dominate the better of the backdoor and frontdoor estimators by an arbitrary amount. That is, we show that the quantity $R = \frac{\min\{\text{Var}(\widehat{ac})_{\text{backdoor}}, \text{Var}(\widehat{a}_f\widehat{c})\}}{\text{Var}(\widehat{a}_c\widehat{c})}$ is unbounded. Consider the case when $\text{Var}(\widehat{ac})_{\text{backdoor}} = \text{Var}(\widehat{a}_f\widehat{c})$. This condition holds for certain settings of the model parameters (see Appendix E.3 for an example). In this case,

$$R = \frac{\text{Var}(\widehat{ac})_{\text{backdoor}}}{\text{Var}(\widehat{a}_c\widehat{c})} \geq \frac{(n-2)DE(a^2\sigma^2_{u_m} + \sigma^2_{u_y})}{\sigma^2_{u_x}\left(F + \sigma^2_{u_y}\left(\sigma^2_{u_m} + \sqrt{3}\sigma^2_{u_m}\left(r_1 r_2 + |c|(n-2)D\left(|c| + G\right)\right)\right)\right)}, \quad (14)$$

where $D = d^2\sigma^2_{u_w} + \sigma^2_{u_x}$, $E = c^2\sigma^2_{u_x} + \sigma^2_{u_m}$, $r_1 = \sqrt{\frac{n-3}{n-5}}$, $r_2 = \sqrt{\frac{n-2}{n-4}}$, $F = (n-3)a^2\sigma^2_{u_m}E$, $G = r_1\frac{\sigma_{u_m}}{\sqrt{(n-2)D}}$ and, in Eq. 14, we used Eq. 12. As $\sigma_{u_x} \to 0$, $R \to \infty$ and thus R is unbounded. This shows that, even in finite samples, combining confounders and mediators can lead to an arbitrarily better estimator than the better of the backdoor and frontdoor estimators.

5.2 Semi-Parametric Estimators

Fulcher et al. [2020] derive the efficient influence function and semi-parametric efficiency bound for a generalized model with discrete treatment and non-linear relationships between the variables. While they allow for confounding of the treatment-mediator link and the mediator-outcome link, the graph in Figure 1 has additional restrictions. As per Chen and Santos [2018], this graph is *locally overidentified*. This suggests that it is possible to improve the estimator by Fulcher et al. [2020, Eq. (6)] (which we refer to as IF-Unrestricted). In our model, we have $Y \perp\!\!\!\perp X|(M, W)$, and $M \perp\!\!\!\perp W|X$. So we incorporate these conditional independences in IF-Unrestricted by using $\mathbb{E}[Y|M, W, X] = \mathbb{E}[Y|M, W]$, and $f(M|X, W) = f(M|X)$ to create an estimator we refer to as IF-Restricted:

$$\widehat{\Psi} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mathbb{E}}[Y|M_i, W_i])\frac{\widehat{f}(M|x^*)}{\widehat{f}(M|X_i)} + \frac{1\{X_i = x^*\}}{\widehat{P}(X_i = x^*|W_i)}$$
$$\times\left\{\widehat{\mathbb{E}}[Y|M_i, W_i] - \sum_m \widehat{\mathbb{E}}[Y|m, W_i]f(m|X_i)\right\} + \sum_m \widehat{\mathbb{E}}[Y|m, W_i]\widehat{f}(m|x^*),$$

where, if $\widehat{f}$, $\widehat{P}$, and $\widehat{\mathbb{E}}$ are consistent estimators, then $\widehat{\Psi} \xrightarrow{p} \mathbb{E}[Y|do(X = x^*)]$. By double robustness of the given estimator, if $\widehat{f}$, $\widehat{P}$, and $\widehat{\mathbb{E}}$ are correctly specified, then IF-Restricted has identical asymptotic distribution as IF-Unrestricted. But using the additional restrictions improves estimation of nuisance functions. Rotnitzky and Smucler [2019], in contemporaneous work, analyzed the same graph and showed that, in addition, the efficient influence function is also changed when imposing these conditional independences (see Example 10 in their paper) (we refer to the estimator for this influence function as IF-Efficient). IF-Efficient can also be obtained as a special case of Bhattacharya et al. [2020, Theorem 9] For our experiments with binary treatments, we use linear regression for $\widehat{f}$, $\widehat{\mathbb{E}}$ and logistic regression for $\widehat{P}$. Another way to adapt IF-Unrestricted is for the case when we do not observe the confounders (as in the frontdoor adjustment). In this case, we can set $W_i = \varnothing$ and apply $\widehat{\Psi}$. We call this special case IF-Frontdoor.

6 Combining Revealed-confounder and Revealed-mediator Datasets

We now consider a situation in which the practitioner has access to two datasets. In the first one, the confounders are observed but the mediators are unobserved. In the second one, the mediators are observed but the confounders are unobserved. Given the two datasets, we wish to optimally leverage all available data to estimate the effect of X on Y.

A naive approach would be to apply the backdoor and frontdoor estimator to the first and second dataset, respectively, and take a weighted average of the two estimates. However, in this case, the variance will be between that of the frontdoor and backdoor estimator. We analyze the MLE, showing that it has lower variance than both the backdoor and frontdoor estimators.

Combined Log-Likelihood under Partial Observability Say we have P samples of $\{x_i, y_i, w_i\}_{i=1}^P$. Let each such sample be denoted by the vector $\mathbf{p}_i = [x_i, y_i, w_i]$. Moreover, say we have Q samples of $\{x_i, y_i, m_i\}_{i=1}^Q$. Let each such sample be denoted using

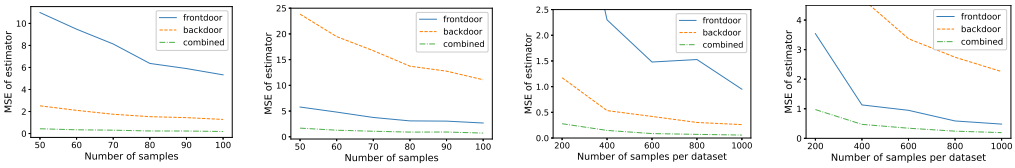

Figure 2: Comparison of MSE. In 2a and 2b, confounders and mediators are observed simulataneously. In 2c and 2d, they are observed in separate datasets. The combined estimator always outperforms the others.

the vector $\mathbf{q}_j = [x_j, y_j, m_j]$. Let the observed data be represented as D. That is, $D = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_P, \mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_Q\}$. Let $N = P + Q$ and let $k = \frac{P}{N}$. Since the data is multivariate Gaussian, the conditional log-likelihood given k can be written as

$$\mathcal{LL}(D|k) = -\frac{N}{2}[k\left(\log\det\boldsymbol{\Sigma}_p + \mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_p\boldsymbol{\Sigma}_p^{-1}\right)\right) + (1-k)\left(\log\det\boldsymbol{\Sigma}_q + \mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_q\boldsymbol{\Sigma}_q^{-1}\right)\right)], \quad (15)$$

where $\boldsymbol{\Sigma}_p = \mathrm{Cov}([X, Y, W])$, $\boldsymbol{\Sigma}_q = \mathrm{Cov}([X, Y, M])$, $\widehat{\boldsymbol{\Sigma}}_p = \frac{\sum_{i=1}^{P}\mathbf{p}_i\mathbf{p}_i^\top}{P}$ and $\widehat{\boldsymbol{\Sigma}}_q = \frac{\sum_{i=1}^{Q}\mathbf{q}_i\mathbf{q}_i^\top}{Q}$.

Cramer-Rao Lower Bound In order to compute the variance of the estimate of parameter $e = ac$, we compute the Cramer-Rao variance lower bound. We first compute the Fisher information matrix $\mathbf{I}$ as $\mathbf{I} = -\mathbb{E}\left[\nabla_\theta^2\mathcal{LL}\right]$, where θ represents the eight model parameters. Let $\widehat{e}$ be the MLE. Since regularity holds for our model (due to linearity and Gaussianity), the MLE is asymptotically normal. Using the Cramer-Rao theorem, for constant k, as $N \to \infty$, we have $\sqrt{N}(\widehat{e} - e) \xrightarrow{d} \mathcal{N}(0, V_e)$, where V_e is a function of $\mathbf{I}^{-1}$. The closed form expression for V_e is given in Appendix F.1.

Note that, for a fixed k, $\lim_{P\to\infty,Q\to\infty}(V_e - \mathrm{Var}(\sqrt{P}\widehat{a}\widehat{c})_{\text{backdoor}}) < 0$ and $\lim_{P\to\infty,Q\to\infty}(V_e - \mathrm{Var}(\sqrt{Q}\widehat{a}_f\widehat{c})) < 0$. This shows that the combined estimator always has lower variance than that of the backdoor and frontdoor estimators on the individual datasets. Moreover, we also find cases where the combined estimator outperforms both the backdoor and frontdoor estimators even when the total number of samples are the same. That is, there exist model parameters such that $\lim_{N\to\infty}(V_e - \mathrm{Var}(\sqrt{N}\widehat{a}\widehat{c})_{\text{backdoor}}) < 0$ and $\lim_{N\to\infty}(V_e - \mathrm{Var}(\sqrt{N}\widehat{a}_f\widehat{c})) < 0$ for some $k \in (0, 1)$. This means that is these cases, it is better to collect a mix of confounders and mediators rather than only collecting mediators or confounders. Despite having access to the same number of samples, a mix of confounders and mediators can lead to lower variance. This happens when the variances of the backdoor and frontdoor estimators are close to each other (see Appendix F.2 for more details).

The Maximum Likelihood Estimator Computing an analytical solution for the model parameters that maximizes the log-likelihood turns out to be intractable. As a result, we update our estimated parameters to maximize the likelihood numerically. The likelihood in Eq. 15 is non-convex. So we intialize the parameters using the two datasets (see Appendix F.3 for details) and run the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [Fletcher, 2013] to maximize the likelihood. In our experiments, for the sample sizes considered and given our initialization procedure, the non-convexity of the likelihood never proved a practical problem. When we find the global minimum, this estimator is optimal and dominates both the backdoor and frontdoor estimators.

7 Experiments

Synthetic Data We first show that the empirical variance of the various estimators is close to the theoretical variance (Table 1). We randomly initialize parameters and for each instance, we compute the Mean and Standard Deviation of Absolute Percentage Error of theoretical variance as a predictor of empirical variance (see Appendix G.1 for more details). Next, we compare the estimators under different settings of the model parameters. Unless stated otherwise, the model parameter values we use for experiments are $a = 10, b = 4, c = 5, d = 5, \sigma_{u_w}^2 = 1, \sigma_{u_x}^2 = 1, \sigma_{u_m}^2 = 1, \sigma_{u_y}^2 = 1$. The quantity of interest is the causal effect $ac = 50$. For Figure 2a, we set $\{\sigma_{u_x}^2 = 0.05, \sigma_{u_m}^2 = 0.05\}$, which makes the backdoor estimator better as predicted by Eq. 9. For Figure 2b, we set $\{\sigma_{u_w}^2 = 2, \sigma_{u_x}^2 = 0.01, \sigma_{u_m}^2 = 0.1\}$ which makes the frontdoor estimator better as predicted by Eq. 9. The plots in Figure 2a and 2b corroborate these predictions at different sample sizes. Furthermore, the optimal combined estimator always outperforms both the backdoor and frontdoor estimators.

Table 1: Mean Absolute Percentage Error of the theoretical variance as a predictor of the empirical variance. The values are reported as mean $\pm$ std. The % error is small even for small sample sizes.

ESTIMATOR	$n = 50$	$n = 100$	$n = 200$
BACKDOOR	$0.36 \pm .3$	$0.32 \pm .2$	$0.34 \pm .1$
FRONTDOOR	$0.33 \pm .2$	$0.30 \pm .2$	$0.23 \pm .1$
COMBINED	1.20 ± 1.1	$0.97 \pm .6$	$0.58 \pm .2$

Table 2: Results on the IHDP and JTPA data. The *complete* (C) data setting is when $\{W, X, M, Y\}$ are observed and *partial* (P) is with $\{X, M, Y\}$ and $\{W, X, Y\}$ in two separate datasets.

		IHDP MSE		JTPA	
ESTIMATOR	DATA	S1	S2	VAR	MSE
BACKDOOR	C	2.14	1.07	NA	NA
FRONTDOOR	C	1.97	2.81	40.9K	75.3K
COMBINED	C	**1.78**	**0.93**	**33.1K**	**70.1K**
IF-FRONTDOOR	C	4.24	2.07	46.6K	77.9K
IF-RESTR	C	**3.49**	**1.48**	**40.4K**	**42.1K**
IF-UNRESTR	C	3.82	1.87	45.1K	46.2K
IF-EFFICIENT	C	3.58	2.01	NA	NA
BACKDOOR	P	5.44	2.43	NA	NA
FRONTDOOR	P	3.92	4.94	**74.8K**	**115.1K**
COMBINED	P	**2.97**	**1.62**	79.5K	123.1K

Next, we evaluate the procedure for combining datasets described in Section 6, generating two datasets with equal numbers of samples. In the first, only $\{X, Y, W\}$ are observed. In the second, only $\{X, Y, M\}$ are observed. We set $\{\sigma^2_{u_x} = 0.05, \sigma^2_{u_m} = 0.05\}$, which makes the backdoor estimator better (Figure 2c), and then set $\{\sigma^2_{u_w} = 2, \sigma^2_{u_x} = 0.01, \sigma^2_{u_m} = 0.1\}$, which makes the frondoor estimator better (Figure 2d). The plots show that the combined estimator has lower MSE than either for various sample sizes (Figure 2), supporting our theoretical claims.

IHDP Dataset Hill [2011] constructed a dataset from the Infant Health and Development Program (IHDP). This semi-synthetic dataset, which has been used for benchmarking causal inference algorithms [Shi et al., 2019, Shalit et al., 2017], is based on a randomized experiment to measure the effect of home visits from a specialist on future test scores of children. We use samples from the NPCI package [Dorie, 2016]. The randomized data is converted to an observational study by removing a biased subset of the treated group. This set contains 747 samples with 25 covariates.

We use the covariates and the treatment assignment from the real study. We use a procedure similar to Hill [2011] to simulate the mediator and the outcome. The mediator M takes the form $M \sim \mathcal{N}(cX, \sigma^2_{u_m})$, where X is the treatment. The response Y takes the form $Y \sim \mathcal{N}(aM + W\mathbf{b}, 1)$ where W is the matrix of standardized (zero mean and unit variance) covariates and values in the vector $\mathbf{b}$ are randomly sampled (0, 1, 2, 3, 4) with probabilities (0.5, 0.2, 0.15, 0.1, 0.05). The ground truth causal effect is $c \times a$.

We evaluate our estimators and the four IF estimators — IF-Unrestricted, IF-Restricted, IF-Frontdoor, and IF-Efficient (Section 5.2). We test the estimators on two settings of the model parameters (Table 2, the *Complete* dataset setting). The MSE values are computed across 1000 instantiations of the dataset created by simulating the mediators and outcomes. We first evaluate the estimators on the complete dataset of 747 samples. We see that for Setting 1 (S1): $a = 10, c = 5, \sigma_{u_m} = 1$, the backdoor estimator dominates the frontdoor estimator whereas for Setting 2 (S2): $a = 10, c = 1, \sigma_{u_m} = 2$, the frontdoor estimator is better. In both cases, the combined estimator (Section 5) outperforms both estimators. Furthermore, we see that IF-Restricted outperforms IF-Frontdoor, showing the value of leveraging the covariates. Moreover, IF-Restricted also outperforms IF-Unrestricted, suggesting that incorporating model restrictions improves performance. Next, we randomly split the data into two sets, one with the confounder observed and the other with the mediator observed, finding that the estimator that combines the datasets (Section 6) outperforms the frontdoor and backdoor estimator (Table 2, the *Partial* dataset setting). We compute the MSE across 1000 instantiations of the dataset.

National JTPA Study The National Job Training Partnership Act (JTPA) Study evaluates the effect of a job training program on future earnings. The treatment X represents if a participant signed up for the program. The outcome Y represents future earnings. The collected covariates (like race, study location, age) are the confounders W. There was non-compliance among the treated units. The mediator M represents compliance, that is, whether the participant make use of JTPA services after signing up. The ground truth treatment effect, computed from the randomized component of this study, is 862.74. Glynn and Kashin [2018] showed that the backdoor estimator has high bias, suggesting that there was unmeasured confounding, so we omit that in our results. They also justify the assumptions required for the frontdoor estimator and show that it works well for this study.

A comparison of the frontdoor estimator, the combined estimator (Section 5), IF-Restricted, IF-Frontdoor and IF-Unrestricted (Section 5.2) is shown in Table 2 (the "C" data setting). For IF-Restricted, we only use the $f(M|X,W) = f(M|X)$ restriction and do not use the $\mathbb{E}[Y|M,W,X] = \mathbb{E}[Y|M,W]$ restriction since it is not valid. We compute the variance and MSE using 1000 bootstrap iterations. The combined estimator has lower variance and MSE than the frontdoor estimator. IF-Restricted outperforms IF-Frontdoor, reinforcing the utility of combined estimators. Furthermore, IF-Restricted outperforms IF-Unrestricted, showing that using model restrictions is valuable. Next, we evaluate our procedure for the partially-observed setting (Section 6). We compute variance and MSE across 1000 bootstrap iterations. At each iteration, we randomly split our dataset into two datasets of equal size, one with revealed confounders, one with revealed mediators. The combined estimator does not outperform the frontdoor estimator (Table 2, the "P" data setting). This is expected since the backdoor adjustment works poorly and the revealed-confounder data is unlikely to help. Despite this, the combined estimator does not suffer too badly.

8 Discussion

In this paper, we studied over-identified graphs with confounders and mediators, showing that the two identification strategies can lead to estimators with arbitrarily different variances. We show that having access to both confounders and mediators (either simultaneously or in separate datasets) can give (unbounded) performance gains. We also show that our results qualitatively apply to general non-linear settings.

Future Work We see several promising lines for future work, including (i) extensions to more general graphs; (ii) online data collection subject to some cost structure over the observations; and (iii) leveraging overidentification to mitigate errors due to measurement and confounding. Our experiments show the applicability of our methods in the frontdoor-backdoor graph, with combined estimators yielding gains in both linear and non-linear settings. We hope next to extend the results to more general over-identified causal graphs. Additionally, we plan to analyze the online data collection setting. Here, subject to budget constraints, a practitioner must choose which variables to observe at each time step. This direction seems especially important in medical applications (where each test may be costly) and survey studies (with a cap on the number of questions). Our current results suggest that the optimal strategy must depend on the model parameters. At each step, the revealed data will improve our estimates of the parameters, in turn impacting what we collect in the future.

One potential limitation of the method is that situations where there exist multiple valid identification formulas may be uncommon in practice, when finding a single source of identification can already be difficult. However, we believe that in reality, many identification approaches are often available, but members of the community find flaws in each of the proposed estimators. In these cases, with multiple imperfect estimators of the same causal effect, we believe that overidentification might be leveraged to create robust combined estimators.

References

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

E. Bareinboim et al. Causal inference and data-fusion in econometrics. Technical report, arXiv. org, 2019.

M. F. Bellemare and J. R. Bloem. The paper of how: Estimating treatment effects using the front-door criterion. *Working Paper*, 2019.

R. Bhattacharya, R. Nabi, and I. Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

X. Chen and A. Santos. Overidentification in regular models. *Econometrica*, 86(5):1771–1817, 2018.

A. Chinco and C. Mayer. Misinformed speculators and mispricing in the housing market. *The Review of Financial Studies*, 29(2):486–522, 2016.

L. Cohen and C. J. Malloy. Friends in high places. *American Economic Journal: Economic Policy*, 6 (3):63–91, 2014.

V. Dorie. Non-parametrics for causal inference. `https://github.com/vdorie/npci`, 2016.

M. L. Eaton. *Chapter 8: The Wishart Distribution*, volume Volume 53 of *Lecture Notes–Monograph Series*, pages 302–333. Institute of Mathematical Statistics, 2007. doi: 10.1214/lnms/1196285114. URL `https://doi.org/10.1214/lnms/1196285114`.

R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

I. R. Fulcher, I. Shpitser, S. Marealle, and E. J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020.

A. N. Glynn and K. Kashin. Front-door difference-in-differences estimators. *American Journal of Political Science*, 61(4):989–1002, 2017.

A. N. Glynn and K. Kashin. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association*, 113(523):1040–1049, 2018.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

G. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Technical report, National Bureau of Economic Research, 2019.

C. Jackson, N. Best, and S. Richardson. Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics*, 10(2):335–351, 2009.

T. C. Koopmans and O. Reiersøl. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

M. Kuroki. Selection of post-treatment variables for estimating total effect from empirical research. *Journal of the Japan Statistical Society*, 2000.

L. Le Cam. Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique*, 1990.

T. J. Page Jr. Multivariate statistics: A vector space approach. *JMR, Journal of Marketing Research (pre-1986)*, 1984.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl. The foundations of causal inference. *Sociological Methodology*, 40(1):75–149, 2010.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. A complete generalized adjustment criterion. In *Uncertainty in Artificial Intelligence (UAI)*, 2015.

A. Rotnitzky and E. Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306*, 2019.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning (ICML)*, 2017.

C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2503–2513, 2019.

C. Uhler. Gaussian graphical models: An algebraic and geometric perspective. *Chapter in Handbook of Graphical Models*, 2019.

S. Wright. The method of path coefficients. *Ann. Math. Statist.*, 5(3):161–215, 09 1934. doi: 10.1214/aoms/1177732676.