Variance Estimation and Selecting Estimands for Causal Effect Queries

Anna K Raichev[1] **Rina Dechter**[1] **Jin Tian**[2] **Alexander Ihler**[1]

[1]Computer Science Dept., University of California, Irvine, Irvine, California, USA
[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates,

Abstract

This paper investigates the statistical efficiency of algebraic expressions ("estimands") used to answer causal queries. We first examine structural rules, developing a partial dominance relationship for front-door estimands that extends recent results on back-door estimands. In many models, however, structural rules alone may be insufficient to select an estimand. For such cases, we propose empirical techniques for estimating and comparing the variance of different estimands using both the graph and observational data: 1) a bootstrap-based method, and 2) a computationally simpler yet practically effective method for discrete models which estimates the Fisher information matrices. We illustrate both methods' effectiveness on a variety of causal diagrams and estimand forms.

1 INTRODUCTION

Causal models are at the heart of many fields of research [Pearl, 2009, Shpitser and Pearl, 2006a, Spirtes et al., 2000], yet tools for answering causal queries remain limited. Structural Causal Models (SCMs) provide a popular framework for causal inference [Pearl, 2009].

When the full SCM is available, namely functions specifying the actual causal relationships are known, we can use standard probabilistic inference [Darwiche, 2009, Dechter, 2013] to directly answer causal queries evaluating how forcing a variable's assignment $X = x$ affects another variable Y, written as $P(Y|do(X = x))$. In practice however we rarely have access to the full model. Instead, we may have access to partial information in the form of the causal diagram – a directed graph capturing the causal relationships of the underlying SCM.

Causal diagrams typically include both *observable variables*, which can be measured from data, and *latent variables*, which are unobservable and for which data cannot be collected. When only the diagram is available, some queries can still be answered using observable data only [Tian and Pearl, 2002, Shpitser and Pearl, 2006a,b, Huang and Valtorta, 2008]. In such cases, the query is said to be *identifiable*. This also means that for the query, $P(Y|do(X))$ there exists an expression in terms of only observational (conditional) probabilities. Importantly, identifiability can be determined efficiently based on the causal diagram alone [Tian, 2002, Shpitser and Pearl, 2006a]. Indeed, the standard approach for answering identifiable queries from the observational distribution is to construct an algebraic expression, called an *estimand*, that expresses the query symbolically using probabilistic quantities over the observed variables only.

However, there may be more than one estimand for a given query. In particular, expressions fall into well-studied special families known as *back-door* (and inverse weighting methods) or *front-door* expressions. The do-calculus [Pearl, 2009] provides a complete approach for deriving an estimand and can produce more complex expressions. A popular and efficient algorithm called the *ID algorithm* can determine a query's identifiability and, if the query is identifiable, provides a single estimand [Tian and Pearl, 2002, Tian, 2002]. Methods for enumerating adjustment sets, including back-door and front-door expressions are available [Wienöbst et al., 2024, van der Zander et al., 2014, Jeong et al., 2022].

Given an estimand for a query, *query evaluation* can be accomplished by estimating the probabilistic quantities in the estimand from the observational data, and then evaluating the estimand using the estimated probabilistic quantities. Over the past few years, a variety of strategies have been developed for query evaluation using statistical estimation methods [Jung et al., 2021a, Bhattacharya et al., 2022, Jung et al., 2020a,b, 2021b].

Since there can be many different estimands for a query, the question arises: which one should be selected?

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5643–5660.

Estimand expressions can vary significantly in their computational evaluation costs and statistical efficiency—factors that may hold varying degrees of importance depending on the setting. Indeed, recent work has studied methods for minimizing the computational cost [Dechter et al., 2025], as well as on minimizing the number of variables involved (reducing measurement work) [Tikka and Karvanen, 2018a,b, Guo et al., 2022]. Minimizing the asymptotic variance of the estimator has been studied in the special case of linear structural equation models [Taguchi and Kuroki, 2026], or restricted to backdoor conditions [Smucler et al., 2022, Rotnitzky and Smucler, 2020, Smucler and Rotnitzky, 2022]. **Our work here focuses on comparing the asymptotic variance of competing estimands.**

Results. In this paper we study variance estimation methods for estimand expressions, including front-door and other general-form expressions, focusing mainly on discrete variable models. We provide results of two types. First, we extend a dominance result obtained for back-door estimands [Smucler et al., 2022] to front-door estimands, allowing some estimands to be compared using the causal diagram alone. However, in many problems the variance comparison between two estimands can depend on the underlying distribution. We illustrate this point experimentally, and propose and study two simple variance estimators that make use of the observational data to determine which estimand to use.

2 BACKGROUND

Structural causal models. Our framework is based on *Structural Causal Models (SCM)* [Pearl, 2009]. An SCM is a quadruple $M = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$ where $\boldsymbol{V}$ and $\boldsymbol{U}$ are sets of endogenous and exogenous (latent) variables, respectively. $\boldsymbol{F}$ is a set of functions (called mechanisms): each $V \in \boldsymbol{V}$ is a function $f_V(PA_V, \boldsymbol{U}_V)$ of its endogenous and exogenous parents, $PA_V \subseteq \boldsymbol{V}$, and $\boldsymbol{U}_V \subseteq \boldsymbol{U}$, respectively. $P(u)$ is a joint distribution over $\boldsymbol{U}$. We assume that all the $\boldsymbol{U}$ variables are independent of each other.

The SCM M defines a probability distribution $P(\boldsymbol{V}, \boldsymbol{U})$,

$$P(\boldsymbol{V}, \boldsymbol{U}) = \prod_{V \in \boldsymbol{V}} \delta(V - f_V(PA_V, \boldsymbol{U}_V)) \cdot \prod_{U_V \in \boldsymbol{U}} P(\boldsymbol{U}_V).$$

The SCM induces an *observational distribution*, $P(\boldsymbol{V})$:

$$P(\boldsymbol{V}) = \sum_{\boldsymbol{U}} P(\boldsymbol{V}, \boldsymbol{U}).$$

Causal diagrams. The *causal diagram* of an SCM $M = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$ is a directed graph $\mathcal{G}$ whose nodes are the endogenous variables $\boldsymbol{V}$ and there is an edge $V_i \to V_j$ for every $V_i, V_j \in \boldsymbol{V}$ iff $V_i \in PA_{V_j}$. There is a bidirectional edge between V_i, V_j if $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$ share variables. The bidirectional edges indicate that there exists some latent variable pointing to both of the observed (endogenous) variables. See Figure 1 for an example.

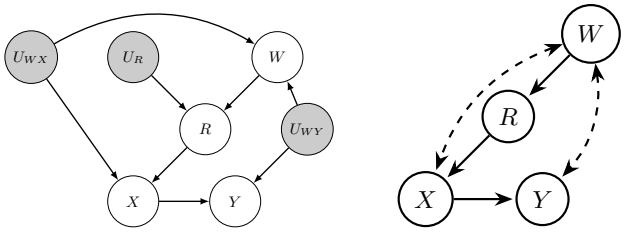


Figure 1: Causal diagrams for an SCM over both observed (endogenous) variables $\boldsymbol{V}$ and latent variables $\boldsymbol{U}$ (left), and over observed variables $\boldsymbol{V}$ only (right).

Causal effect and the truncation formula. An external intervention forcing variables $\boldsymbol{X}$ to take on value $\boldsymbol{x}$, called $do(\boldsymbol{X} = \boldsymbol{x})$, is modeled by replacing the mechanism for each $X \in \boldsymbol{X}$ with the function $X = x$. Formally,

$$P(\boldsymbol{V} \setminus \boldsymbol{X}, \boldsymbol{U} \mid do(\boldsymbol{X} = \boldsymbol{x})) = \prod_{V \notin \boldsymbol{X}} \delta(V - f_V(PA_V, \boldsymbol{U}_V)) \cdot P(\boldsymbol{U}) \Big|_{\boldsymbol{X} = \boldsymbol{x}}$$

The effect of $do(\boldsymbol{X})$ on variables $\boldsymbol{Y}$, denoted $P(\boldsymbol{Y}|do(\boldsymbol{X}))$, is defined by marginalizing over the truncated distribution:

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})) = \sum_{(\boldsymbol{V} \cup \boldsymbol{U}) \setminus \boldsymbol{Y}} P(\boldsymbol{V} \setminus \boldsymbol{X}, \boldsymbol{U} \mid do(\boldsymbol{X} = \boldsymbol{x})).$$

Identifiability and causal-effect query. The standard formulation of causal inference assumes that we only have access to the causal graph $\mathcal{G}$ and the observational distribution $P(\boldsymbol{V})$ (or a sample from it). The identifiability task is to determine if a given causal-effect query, $P(\boldsymbol{Y} \mid do(\boldsymbol{X}))$, can be *uniquely* answered from $\mathcal{G}$ and $P(\boldsymbol{V})$ [Pearl, 2009]. If it is, an *estimand* – an expression in terms of $P(\boldsymbol{V})$ that equals the causal effect query – is generated. Then the estimation task is to compute the value of $P(\boldsymbol{Y} \mid do(\boldsymbol{X}))$ given data samples drawn from $P(\boldsymbol{V})$.

A number of approaches have been applied to estimation. A simple approach for discrete-variable models, called the *plug-in estimator*, estimates each term in the estimand with its empirical probability value obtained from the data, though more sophisticated approaches have also been developed [Jung et al., 2020a,b, Raichev et al., 2024].

Estimand-based approaches. As noted above, once a causal-effect query is determined to be identifiable, a variety of estimands can be generated and used to estimate the answer. One such estimand expression can be derived via the do-calculus [Pearl, 2009]. A popular polynomial algorithm called *Identifiability* (ID) has been developed for this task, which generates a single estimand [Tian, 2002, Shpitser and

Pearl, 2006a]. Special expression types known as *back-door* and *front-door* estimands have been extensively explored, and they are also at the heart of our paper. Both are defined in terms of an *adjustment set* that satisfies certain graph conditions.

Back-door Adjustment. [Pearl, 2009, Spirtes et al., 2000] A set of variables $\boldsymbol{Z}$ satisfies the back-door criterion relative to (X, Y) if a) there is no node in $\boldsymbol{Z}$ which is a descendant of X; and b) $\boldsymbol{Z}$ blocks every path between X and Y that contains an arrow into X. If $\boldsymbol{Z}$ is a back-door adjustment set then the causal effect of X on Y is identifiable and the back-door estimand is:

$$P(y \mid \text{do}(x)) = \sum_{z} P(y \mid x, z)P(z) \tag{1}$$

Front-door Adjustment. A set $\boldsymbol{Z}$ satisfies the front-door criterion relative to (X, Y) iff a) $\boldsymbol{Z}$ intercepts all directed paths from X to Y, b) there is no unblocked back-door path from X to $\boldsymbol{Z}$; and c) all back-door paths from $\boldsymbol{Z}$ to Y are blocked by X. If also $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by:

$$P(y \mid do(x)) = \sum_{z} P(z \mid x) \sum_{x'} P(y \mid x', z)P(x') \tag{2}$$

Figure 2 shows a causal diagram of an SCM with a front door adjustment set.

As noted earlier, there may exist many possible adjustment sets for a single causal query. For example, for one of the diagrams used in the experiments (Fig. 3b) there exist several back-door adjustment sets, $\{Z_2, Z_4\}$ and $\{Z_4, Z_5\}$, as well as a front-door adjustment set, $\{Z_6, Z_7\}$.

Here, we focus on **choosing the most statistically efficient estimand** to minimize the asymptotic variance, which leads us to the problem of estimating the asymptotic variance of competing estimands.

3 DOMINATING ESTIMANDS

Recent work has shown that in some cases it is possible to compare the asymptotic variances of competing estimands based solely on the causal diagram and the query. In particular, Smucler et al. [2022] provide such conditions for back-door estimands. In this section we present those results, then extend them to the case of front-door estimands. Throughout this section, we restrict our attention to regular and asymptotically linear (RAL) estimators within the nonparametric model denoted $\mathcal{M}_{np}$, where no restrictions (such as functional form assumptions or linearity) are imposed on the observed data distribution beyond those implied by the causal graph itself. Informally, an estimator is regular if, asymptotically, its behavior does not change

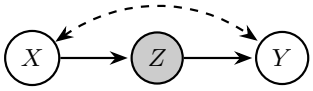


Figure 2: Causal Diagram for an SCM $\mathcal{M}$. Here Z is a valid front-door adjustment set for $\mathbb{E}[Y|do(X = x^*)]$, blocking the direct path from X to Y.

abruptly under small perturbations of the true distribution, and asymptotically linear if it behaves like an average of independent per-observation contributions; see the asymptotic expression (7). Within this class of "well-behaved" estimators we consider the minimum achievable asymptotic variance, called the *semiparametric efficiency bound* [Bickel et al., 1998], which provides a "Cramér-Rao"-like bound for nonparametric models [Newey, 1990] to characterize the lowest possible variance any RAL estimator can achieve. For any causal estimand $\psi(P)$, we refer to this efficiency bound as the (optimal asymptotic) *variance* of $\psi(P)$. First, we consider the following back-door estimand for $\mathbb{E}[Y|do(X = x^*)]$:

$$\Psi_{bd} = \sum_{z} \mathbb{E}[Y \mid x^*, z]\, P(z). \tag{3}$$

We denote the variance of Ψ_{bd} relative to the back-door adjustment set $\boldsymbol{Z}$ by $\sigma^2_{x^*,bd}(\boldsymbol{Z}; P)$, where the dependence on the distribution P is made explicit. Among multiple valid back-door adjustment sets $\boldsymbol{Z}$, the objective is to determine which choice of $\boldsymbol{Z}$ yields the lowest variance.

We say that $\boldsymbol{Z}$ *dominates* $\boldsymbol{Z}'$, and write $\boldsymbol{Z} \preccurlyeq \boldsymbol{Z}'$, iff:

$$\sigma^2_{x^*,bd}(\boldsymbol{Z}; P) \leq \sigma^2_{x^*,bd}(\boldsymbol{Z}'; P) \qquad \forall x^* \in \mathcal{X},\ \forall P \in \mathcal{M}(\mathcal{G})$$

where $\mathcal{M}(\mathcal{G})$ are all probability distributions consistent with the causal diagram $\mathcal{G}$.

We next present results from Smucler et al. [2022], restricted to single variable deterministic intervention.

Lemma 1. *[Smucler et al., 2022] Assume a given causal diagram $\mathcal{G}$ and a causal-effect identifiable query on X and Y, and let $\boldsymbol{B}$ be a* **back-door** *adjustment set and let $\boldsymbol{A} \subset \boldsymbol{V}$ satisfy $X \perp\!\!\!\perp \boldsymbol{A} \mid \boldsymbol{B}$. Then $\boldsymbol{A} \cup \boldsymbol{B}$ is also a back-door adjustment set, and for any $x^* \in \mathcal{X}$ and all $P \in \mathcal{M}(\mathcal{G})$:*

$$\sigma^2_{x^*,bd}(\boldsymbol{B}; P) - \sigma^2_{x^*,bd}(\boldsymbol{A} \cup \boldsymbol{B}; P) \geq 0 \tag{4}$$

That is: $\boldsymbol{A} \cup \boldsymbol{B} \preccurlyeq \boldsymbol{B}$.

Lemma 2. *[Smucler et al., 2022] Let $\boldsymbol{A} \cup \boldsymbol{B}$ be a* **back-door** *adjustment set with $\boldsymbol{A}$ and $\boldsymbol{B}$ disjoint, and suppose $Y \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{A}, X$. Then $\boldsymbol{A}$ is also a back-door adjustment set, and for any $x^* \in \mathcal{X}$ and all $P \in \mathcal{M}(\mathcal{G})$:*

$$\sigma^2_{x^*,bd}(\boldsymbol{A} \cup \boldsymbol{B}; P) - \sigma^2_{x^*,bd}(\boldsymbol{A}; P) \geq 0. \tag{5}$$

That is: $\boldsymbol{A} \preccurlyeq \boldsymbol{A} \cup \boldsymbol{B}$.

Lemmas 1-2 summarize two structural conditions based only on the causal diagram for when adding or removing variables to a given back-door adjustment set decreases the variance. Intuitively, they tell us that, given a valid back-door set, adding variables "closer" to the outcome Y will decrease our variance, while adding variables "closer" to the intervention X can only increase it.

We now show that Lemma 2 can be extended to provide a front-door analogue, giving a condition for dominance between two front-door estimands. On the other hand, Lemma 1 cannot be directly extended to the front-door case, as we will illustrate via counterexample in the experiments.

Specifically, we consider the following front-door estimands for $\mathbb{E}[Y|do(X = x^*)]$:

$$\Psi_{fd} = \sum_z P(z \mid x^*) \sum_x \mathbb{E}[Y \mid x, z]\, P(x). \tag{6}$$

We denote the variance of Ψ_{fd} relative to the front-door adjustment set $\boldsymbol{Z}$ by $\sigma^2_{x^*,fd}(\boldsymbol{Z}; P)$. We derive $\sigma^2_{x^*,fd}(\boldsymbol{Z}; P)$ using the influence function approach [Hampel, 1974]. In statistics the influence function characterizes the first order contribution of a single data point to the asymptotic distribution of an estimator. Among all influence functions, we focus on the efficient influence function (EIF), ϕ, which achieves the minimum variance and thus determines the semiparametric efficiency bound.

Specifically, any RAL estimator $\hat{\psi}$ of a causal parameter ψ based on independent identically distributed observations $V^{(1)} \dots V^{(m)}$ of $\boldsymbol{V}$, admits the asymptotic expansion

$$\sqrt{m}(\hat{\psi} - \psi) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \phi(V^{(i)}) + o_p(1). \tag{7}$$

By the central limit theorem, $\sqrt{m}(\hat{\psi} - \psi) \xrightarrow{d} \mathcal{N}(0,\ \mathrm{Var}_P\{\phi(\boldsymbol{V})\})$. Therefore, all efficient RAL estimators of ψ are asymptotically equivalent and attain the semiparametric efficiency bound given by $\mathrm{Var}_P\{\phi(\boldsymbol{V})\}$ [Vaart, 2000, Bickel et al., 1998].

The EIF for Ψ_{fd} has been derived in Fulcher et al. [2020] for a slightly more general case, incorporating observed pre-exposure covariates C. Setting $C = \varnothing$ gives,

Proposition 1. *[Fulcher et al., 2020] Let $f(\cdot)$[1] denote the probability density function with respect to a dominating measure. The EIF of Ψ_{fd} in the nonparametric model $\mathcal{M}_{np}$, denoted $\varphi_{P,x^*}(X, Y, \boldsymbol{Z})$, is given by*

$$\varphi_{P,x^*}(X, Y, \boldsymbol{Z}) = \big(Y - \mathbb{E}[Y|X, \boldsymbol{Z}]\big)\tfrac{f(\boldsymbol{Z}|x^*)}{f(\boldsymbol{Z}|X)} + \frac{\mathbb{I}(X = x^*)}{f(X)}\Big(\sum_x \mathbb{E}[Y|x, \boldsymbol{Z}]f(x) - \sum_{x\overline{z}} \mathbb{E}[Y|x, \overline{z}]f(\overline{z}|X)f(x)\Big) + \sum_z \mathbb{E}[Y|X, z]f(z|x^*) - \Psi_{fd}$$

and the semiparametric efficiency bound of Ψ_{fd} in $\mathcal{M}_{np}$ is given by $\mathrm{Var}\{\varphi_{P,x^*}(X, Y, \boldsymbol{Z})\}$.

By Proposition 1, we obtain

$$\sigma^2_{x^*,fd}(\boldsymbol{Z}; P) = \mathrm{Var}_P\{\varphi_{P,x^*}(X, Y, \boldsymbol{Z})\}. \tag{8}$$

As with the back-door case, among multiple valid front-door adjustment sets $\boldsymbol{Z}, \boldsymbol{Z}'$, we say that $\boldsymbol{Z}$ *dominates* $\boldsymbol{Z}'$, and write $\boldsymbol{Z} \preccurlyeq \boldsymbol{Z}'$, iff:

$$\sigma^2_{x^*,fd}(\boldsymbol{Z}; P) \leq \sigma^2_{x^*,fd}(\boldsymbol{Z}'; P) \qquad \forall x^* \in \mathcal{X},\ \forall P \in \mathcal{M}(\mathcal{G})$$

Based on Proposition 1, we can prove an analogue to Lemma 2 for front-door adjustment, which tells us that removing unnecessary (conditionally independent) variables $\boldsymbol{B}$ from an adjustment set $\boldsymbol{A} \cup \boldsymbol{B}$ cannot increase the variance.

Our main result is stated in the following theorem.

Theorem 1. *Given causal diagram $\mathcal{G}$ and identifiable query on X and Y, let $\boldsymbol{A} \cup \boldsymbol{B}$ be a* **front-door** *adjustment set with $\boldsymbol{A}$ and $\boldsymbol{B}$ disjoint, and $Y \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{A}, X$. Then $\boldsymbol{A}$ is also a front-door adjustment set and for any intervention $X = x^*$ and for all $P \in \mathcal{M}(\mathcal{G})$,*

$$\sigma^2_{x^*,fd}(\boldsymbol{A} \cup \boldsymbol{B}; P) - \sigma^2_{x^*,fd}(\boldsymbol{A}; P) = \mathbb{E}\Big[\mathrm{Var}(Y|\, X, \boldsymbol{A})\, \mathrm{Var}\big[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \mid X, \boldsymbol{A}\big]\Big] \geq 0.$$

Consequently, $\boldsymbol{A} \preccurlyeq \boldsymbol{A} \cup \boldsymbol{B}$.

Proof. See Appendix A. □

For example, in causal diagram $\mathcal{G}_3$ (Fig. 3c) we can let $\boldsymbol{A} = \{V_5\}$ and $\boldsymbol{B} = \{V_3\}$. Since $Y \perp\!\!\!\perp \{V_3\} \mid \{V_5\}, X$, adding variable V_3 to the front-door adjustment set $\boldsymbol{A}$ is not beneficial for lowering the asymptotic variance. Thus, adjustment set $\boldsymbol{A}$ dominates $\boldsymbol{A} \cup \boldsymbol{B}$. Similarly since V_5 is the closest variable to our treatment Y, it is not beneficial to add any of the preceding variables $\{V_1, \dots, V_4\}$ to adjustment set $\boldsymbol{A}$, as the same conditional independence holds. In the empirical section we will illustrate that extending Lemma 1 to the front-door case is not valid.

[1] We make the same positivity assumptions as in Fulcher et al. (2020) for probability density function $f(.)$:

1. There exists $m_1 > 0$ such that $f(\boldsymbol{Z}|X) > m_1$ almost surely
2. There exists $m_2 > 0$ such that $f(X) > m_2$ almost surely

4 ESTIMATING THE VARIANCE OF AN ESTIMAND

When the variance relationship of two estimands can be determined solely from the causal diagram $\mathcal{G}$, it holds for any associated distribution $P \in \mathcal{M}(\mathcal{G})$. However, there are SCMs for which the estimates' relative variance may depend on P, with neither estimand always better than the other. In such settings, how can we select the estimand to use? A natural answer is to use the available data from $P(\boldsymbol{V})$ to estimate and compare the variances of our estimands. We consider two simple methods: a bootstrap estimator and a plug-in estimator of the variances.

Bootstrap Estimator. The bootstrap [Efron and Tibshirani, 1993] is a classic frequentist statistical technique for estimating the distribution of a statistic by repeatedly resampling with replacement from the observed dataset $\mathcal{D}$. In particular, to estimate the variance of a statistic $\hat{\psi}(\mathcal{D})$ – in our setting, an estimand comprised of probabilities estimated from $\mathcal{D}$ – we compute the empirical variance of B data sets drawn with replacement from the empirical distribution, $\hat{P}_D$, of our observed data $\mathcal{D}$:

$$\mathcal{D}^{(b)} \overset{\text{i.i.d.}}{\sim} \hat{P}_D(\boldsymbol{V}) = \frac{1}{|\mathcal{D}|} \sum_{v \in \mathcal{D}} \delta(\boldsymbol{V} - v), \quad b = 1 \dots B$$

This has the advantage of making no additional distributional assumptions about the data, and can be applied even to complex statistics (such as general-form estimands). However, its many resampled computations of the estimand expression can be computationally intensive. When comparing two estimand expressions, we use a paired bootstrap scheme in which the same resampled datasets $D^{(b)}$ are used to assess both candidate estimands; this preserves the sample-level covariance between the competing estimators. Given bootstrap estimates of the variance of two estimand expressions, we can simply select the estimand with smaller variance for use. This method is summarized in Algorithm 1 (see also a more detailed version in Algorithm 3 in the Appendix).

Algorithm 1 Bootstrap Variance Estimator

Require: Observed dataset $\mathcal{D}$, number of bootstrap resamples B, estimands $\hat{\psi}_A, \hat{\psi}_B$

Ensure: Bootstrap variance estimate $\widehat{\mathrm{Var}}_{\mathrm{boot}}(\hat{\psi}_A)$ and $\widehat{\mathrm{Var}}_{\mathrm{boot}}(\hat{\psi}_B)$

1: **Resample:** Draw B bootstrap datasets $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(B)}$, each of size $|\mathcal{D}|$, i.i.d. with replacement from the empirical distribution $\hat{P}_{\mathcal{D}}$.
2: **Evaluate:** For each $k \in \{A, B\}$, evaluate $\hat{\psi}_k^{(b)} = \hat{\psi}_k(\mathcal{D}^{(b)})$ for $b = 1, \dots, B$.
3: **Compute variance:** For each $k \in \{A, B\}$, estimate $\widehat{\mathrm{Var}}_{\mathrm{boot}}(\hat{\psi}_k)$ as the sample variance of $\{\hat{\psi}_k^{(b)}\}_{b=1}^{B}$ across resamples.
4: **return** $\widehat{\mathrm{Var}}_{\mathrm{boot}}(\hat{\psi}_A)$ and $\widehat{\mathrm{Var}}_{\mathrm{boot}}(\hat{\psi}_B)$

Algorithm 2 Plug-in Variance Estimator

Require: Data $\mathcal{D} = \{x^{(1)}, \dots, x^{(m)}\}$, ordering $V_1, \dots, V_n$, estimand $\hat{\psi}$

Ensure: estimate of variance $\widehat{\mathrm{Var}}[\hat{\psi}(\hat{\rho})]$

1: **Empirical distribution:** Estimate each conditional probability table $\hat{\rho}_{\cdot,i,v} = \hat{P}(V_i \mid \boldsymbol{V}_{<i} = v)$ from $\mathcal{D}$ by relative frequency.
2: **Fisher information:** Form the block-diagonal Fisher information matrix $I(\hat{\rho})$
3: **Delta method:** Evaluate $\hat{\psi}(\hat{\rho})$ and its gradient $\nabla_\rho \hat{\psi}(\hat{\rho})$ via automatic differentiation, and compute $\widehat{\mathrm{Var}}[\hat{\psi}(\hat{\rho})] = \nabla\hat{\psi}(\hat{\rho})^\top I(\hat{\rho})^{-1} \nabla\hat{\psi}(\hat{\rho})$.
4: **return** $\widehat{\mathrm{Var}}[\hat{\psi}(\hat{\rho})]$

Plug-In Estimator. An alternative variance estimate is provided by the fact that, were the true $P(\boldsymbol{V})$ known, it would be straightforward to determine the asymptotic variance of its plug-in estimate. For a discrete distribution $P(X = x) = \rho_x$, with $\rho_x > 0$ and $\sum_{x=1}^{d} \rho_x = 1$, the maximum likelihood estimate given m independent samples $\mathcal{D} = \{x^{(1)}, \dots, x^{(m)}\}$ is simply the empirical frequency, $\hat{\rho}_x = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{x^{(j)} = x\}$. This estimator is consistent, regular, and asymptotically linear [Vaart, 2000], and its asymptotic (co)variance is characterized by its Fisher information matrix:

$$I(\rho) = m\Big(\mathrm{diag}(\underline{\rho}^{-1}) + \frac{1}{1 - \mathbf{1}^\top \underline{\rho}} \mathbf{1}\mathbf{1}^\top\Big)$$

where $\underline{\rho} = [\rho_1, \dots, \rho_{d-1}]$ are the $d-1$ linearly independent parameters and $\mathbf{1}$ is a vector of ones of the same size.

Any estimand is made up of an expression of conditional distributions, $P(V_i|\boldsymbol{V}_{<i} = v) = \rho_{\cdot,i,v}$ along some ordering of the observed variables $\boldsymbol{V}$. Since the maximum likelihood estimator decomposes across each of these conditional probability vectors, the Fisher information matrix of the full parameter set ρ is block diagonal, with blocks given by each $\rho_{\cdot,i,v}$, where $m_{i,v} = m \cdot P(\boldsymbol{V}_{<i} = v)$ is the expected number of samples with $\boldsymbol{V}_{<i} = v$. Then, the asymptotic variance of the estimand's value $\hat{\psi}(\hat{\rho})$ is given by

$$\mathrm{Var}[\hat{\psi}(\hat{\rho})] = \nabla_\rho \hat{\psi}(\rho)^\top I(\rho)^{-1} \nabla_\rho \hat{\psi}(\rho),$$

where $\nabla_\rho \hat{\psi}(\rho)$ is the derivative of estimand $\hat{\psi}$ with respect to ρ. In practice, given an estimand expression composed of probability tables $P(V_i|\boldsymbol{V}_{<i})$, we can evaluate this derivative using now-standard autograd computations [Abadi et al., 2015] to evaluate the asymptotic variance with little more computation than required for the estimand itself.

The plug-in estimator of variance, then, is straightforward: we assume that the empirical estimates $\hat{\rho}$ of $P(V_i|\boldsymbol{V}_{<i})$

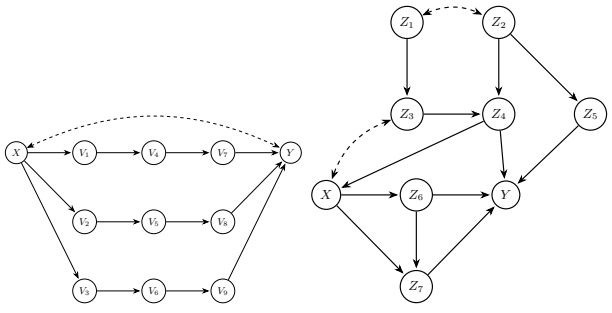
are correct, and evaluate the asymptotic variance of this distribution. This method is summarized in Algorithm 2 (see also a more detailed version in Algorithm 4 in the Appendix). Compared to the bootstrap estimator, this is significantly easier computationally, but estimates only the asymptotic variance, and is also not itself unbiased.

5 EMPIRICAL EVALUATION

We evaluate both the plug-in and bootstrap estimators for asymptotic variance empirically. First, we verify using the exact asymptotic variance, that structural graph-based comparisons between adjustment sets are often insufficient on their own, motivating the need for distribution-dependent estimates. We then assess how well our plug-in and bootstrap estimators determine the more efficient estimand between two candidate estimands given finite samples. All experimental code is available on GitHub [Raichev et al., 2026].

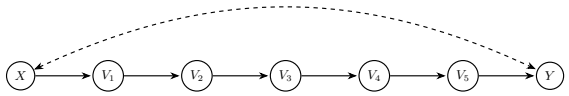
Benchmarks Our experiments focus on structural causal models with binary variables that correspond to the three causal diagrams in Figure 3. To create a full model including the distribution, P, over both observed and latent variables, we generate the conditional probability tables (CPTs), one per variable and its parents. The parameters, ρ, of each CPT were generated by sampling from a Dirichlet distribution with concentration parameters $\alpha \equiv 1$. Across the three graphs, we then selected a number of both front-door and back-door estimands with different adjustment sets for the query $P(Y \mid do(X = x^*))$.

To analyze our estimators given a graph $\mathcal{G}$ and distribution P, we generate data sets by forward sampling [Darwiche, 2009] and then discard the values of the latent variables, yielding samples from the observed distribution only. We calculate the true asymptotic variance for each model and estimand using Fisher information matrices as described in Section 4.



(a) Causal Diagram $\mathcal{G}_1$

(b) Causal Diagram $\mathcal{G}_2$



(c) Causal Diagram $\mathcal{G}_3$

Figure 3: Causal diagrams used in the experiments. Each diagram has multiple adjustment sets for $P(Y \mid do(X = x))$, from which we select two to compare (see Table 2).

Exact asymptotic variance. As shown in Section 3, in some circumstances, structural (graph-based) rules can be used to compare two adjustment sets $\boldsymbol{A}, \boldsymbol{B}$, regardless of the model probabilities. However, such dominance results are often applicable only to limited settings. For example, Lemmas 1-2 and Theorem 1 apply to compare subsumed sets ($\boldsymbol{A} \subseteq \boldsymbol{B}$ or vice versa) under specific conditional independence relations. Comparing arbitrary adjustment sets, including disjoint sets or when the estimands belong to different families (e.g., front-door vs. back-door), is more difficult and in some cases impossible based solely on graph structure. Indeed, as shown by Gorbach et al. [2023], neither the front-door nor the back-door strategy uniformly dominates the other in terms of asymptotic variance. Therefore in many cases, the variance comparison and optimal choice of estimand depends on the underlying distribution P.

Invalidity of Lemma 1 for front-door adjustment sets. To illustrate the dependence on the distribution, we demonstrate via counterexample that, although Lemma 2 could be adapted to the front-door case, Lemma 1's criteria are not sufficient for a similar front-door rule. In Table 1 we present results comparing three pairs of adjustment sets for front-door estimands over the diagram $\mathcal{G}_1$ (Fig. 3a). We sampled $\omega = 100$ different probability distributions $P_i \in \mathcal{M}(\mathcal{G}_1)$, and show the percentage of runs for which the asymptotic variance using front-door adjustment set $\boldsymbol{A}$ is less than or equal to the asymptotic variance when using adjustment set $\boldsymbol{B}$. To have some idea about the quantities involved we also compute the average absolute log ratio of the variances, $\frac{1}{\omega}\sum_{i=1}^{\omega} |\log \frac{\sigma^2_{x^*,fd}(\boldsymbol{B};P_i)}{\sigma^2_{x^*,fd}(\boldsymbol{A};P_i)}|$, where $\sigma^2_{x^*,fd}(\boldsymbol{A};P_i)$ corresponds to the asymptotic variance of front-door estimand A under distribution P_i.

We see, for example, that while adjustment sets $\boldsymbol{A} = \{V_1, V_2, V_3\}$ and $\boldsymbol{B} = \{V_1, V_2, V_3, V_9\}$, satisfy the conditions of Lemma 1 – since $\{V_1, V_2, V_3\}$ is a valid adjustment set and we add $\{V_9\}$ with $X \perp\!\!\!\perp \{V_9\} \mid \{V_1, V_2, V_3\}$ – neither front-door adjustment set dominates the other. For some distributions P_i, $\boldsymbol{A}$ has lower asymptotic variance, while for others $\boldsymbol{B}$ does. We also see the effect of Theorem 1 in the table: adjustment sets $\boldsymbol{A} = \{V_7, V_8, V_9\}$ vs $\boldsymbol{B} = \{V_1, V_7, V_8, V_9\}$ satisfy its conditions, and we see empirically that $\boldsymbol{A}$ always has asymptotic variance at least as small as $\boldsymbol{B}$.

Experimental Setup for Estimators. When one estimand is not known to dominate another, we can use our empirical variance estimates to decide which one we should use.

Comparison: $\boldsymbol{A} = \{V_1, V_2, V_3\}$ **vs.** $\boldsymbol{B} = \{V_1, V_2, V_3, V_9\}$		
$X = x^*$	Freq: $\sigma^2(\boldsymbol{A}) \leq \sigma^2(\boldsymbol{B})$	Mean of $\lvert \log(\frac{\sigma^2(\boldsymbol{B})}{\sigma^2(\boldsymbol{A})}) \rvert$
$x^* = 0$	30	0.0214
$x^* = 1$	31	0.0257

Comparison: $\boldsymbol{A} = \{V_7, V_8, V_9\}$ **vs.** $\boldsymbol{B} = \{V_1, V_7, V_8, V_9\}$		
$X = x^*$	Freq: $\sigma^2(\boldsymbol{A}) \leq \sigma^2(\boldsymbol{B})$	Mean of $\lvert \log(\frac{\sigma^2(\boldsymbol{B})}{\sigma^2(\boldsymbol{A})}) \rvert$
$x^* = 0$	100	0.3880
$x^* = 1$	100	0.4216

Comparison: $\boldsymbol{A} = \{V_4, V_5, V_6\}$ **vs.** $\boldsymbol{B} = \{V_7, V_8, V_9\}$		
$X = x^*$	Freq: $\sigma^2(\boldsymbol{A}) \leq \sigma^2(\boldsymbol{B})$	Mean of $\lvert \log(\frac{\sigma^2(\boldsymbol{B})}{\sigma^2(\boldsymbol{A})}) \rvert$
$x^* = 0$	59	0.2045
$x^* = 1$	64	0.1948

Table 1: Here we look at $\omega = 100$ different models matching causal diagram $\mathcal{G}_1$ (Figure 3). We calculate the frequency of the true asymptotic variance of adjustment set $\boldsymbol{A}$ in a front-door estimand being better than (less than or equal to) the variance of set $\boldsymbol{B}$ per intervention. We also include the average (over distributions P_i) of the absolute log ratio of the asymptotic variances for the two adjustment sets. Here we suppress the subscripts of $\sigma^2(\cdot)$ and the second argument P_i for compactness. The intervention is explicit in the table and they are all front-door adjustment sets.

Causal Diagram $\mathcal{G}_1$	
A:	$\sum_{v_4, v_5, v_6} P(v_6 \mid v_4, v_5, X) P(v_5 \mid v_4, X) P(v_4 \mid X) \times \sum_{x'} P(Y \mid V_4, V_5, V_6, x') P(x')$
B:	$\sum_{v_7, v_8, v_9} P(v_9 \mid v_7, v_8, X) P(v_8 \mid v_7, X) P(v_7 \mid X) \times \sum_{x'} P(Y \mid V_7, V_8, V_9, x') P(x')$

Causal Diagram $\mathcal{G}_2$	
A:	$\sum_{z_4, z_5} P(Y \mid X, z_4, z_5) P(z_5 \mid z_4) P(z_4)$
B:	$\sum_{z_6, z_7} P(z_7 \mid z_6, X) P(z_6 \mid X) \sum_{x'} P(Y \mid x', Z_6, Z_7) P(x')$

Causal Diagram $\mathcal{G}_3$	
A:	$\sum_{v_5} P(v_5 \mid X) \left(\sum_{x'} P(Y \mid V_5, x') P(x') \right)$
B:	$\sum_{v_1, v_5} P(v_5 \mid v_1, X) P(v_1 \mid X) \left(\sum_{x'} P(Y \mid V_1, V_5, x') P(x') \right)$

Table 2: Estimands for $P(Y \mid do(X = x))$ in the experiments. Each graph $\mathcal{G}$ from Figure 3 admits at least two estimands A, B whose quality may vary with the distribution P.

To evaluate the two methods, we expand our experimental framework to the three diagrams in Figure 3, defining a pair of estimands (of various types) for each; see Table 2. For diagram $\mathcal{G}_1$, we compare two front-door adjustment sets, with neither dominating the other. For diagram $\mathcal{G}_2$, we compare a front-door estimand to a back-door estimand. Finally, for diagram $\mathcal{G}_3$ we compare two front-door sets, in which $\boldsymbol{A}$ provably dominates $\boldsymbol{B}$.

We then generate a number of different ("true") distributions P_i for each diagram. Given a particular P_i, we sample a collection of data sets $\{D_{ij}\}$ that could be observed, each of size m. Our task is to decide, given D_{ij}, whether we should prefer estimand A or B to answer $P(Y|do(X = 0))$ for its underlying distribution P_i.

In this task, we expect our results to depend on the relative quality of A vs. B under P_i, as well as on the sample size m. When one estimand has much lower variance than the other, it is easier to detect which is better; and as m increases, our estimates of P_i improve, helping discern even smaller differences in variance.

An idealized version of our problem provides some guidance on what we might expect. For two sets of independent samples from populations A, B, it is a classic statistical problem to test or compare the populations' variance. In such a setting, the standard estimates $\hat{\sigma}^2_A, \hat{\sigma}^2_B$ admit a scaled F-distribution, whose degrees of freedom are determined by the sample sizes used for A and B. Then, the probability of deciding $\sigma^2_A < \sigma^2_B$ would be characterized as a function of the F-distribution cdf, with threshold determined by the relative strength of A over B as quantified by the log-ratio, $s = \log(\sigma^2_B / \sigma^2_A)$.

To evaluate the performance of a selection procedure, we can plot the probability (across data sets j) that our procedure selects estimand A over B, as a function of their relative strength $s_i = \log(\sigma^2_{x^*}(B; P_i) / \sigma^2_{x^*}(A; P_i))$. [2] When $s_i \approx 0$ we are unlikely to distinguish superiority, but with little significance since both estimators are equally good. For $s_i \gg 0$, estimand A is much better than B, and we hope to choose A with high probability; for $s_i \ll 0$, it is much worse and we hope to choose A rarely.

For each probability distribution P_i, we evaluate the relative estimand strength s_i, and plot this against the fraction of data sets D_{ij} of size m that preferred A over B (square markers in Figure 4 and Figure 5). For reference, we also show the F-distribution curves corresponding to the theoretically predicted probability were we to compare m independent samples with the same relative variance (dotted lines).

[2] To simplify notation, we omit the fd and bd, which is implicitly determined by the graph $\mathcal{G}$ and the estimands in Table 2. When clear from context, we use the estimand (e.g., "A") as the argument to the variance, as each adjustment set corresponds to a unique front-door or back-door estimand.

Results. In Figure 4, we present the results for the plug-in estimator on each diagram and its pair of estimands. Causal Diagram $\mathcal{G}_1$, uses non-dominating front-door estimands. We see that, for the distribution P_i with $s_i \approx 0$, we select A vs. B with approximately equal probability, and tend to select A more often as it becomes stronger than B ($s_i > 0$) and less often when it is weaker ($s_i < 0$). Our predictions grow more accurate with data size, with almost perfect predictions when $m \geq 1000$.

In causal diagram $\mathcal{G}_2$, estimand A is a back-door expression and estimand B is a front-door expression. Here we see a similar trend, where the selection performance becomes more accurate when $s_i \ll 0$ or $s_i \gg 0$.

Causal diagram $\mathcal{G}_3$ also illustrates this point – here, estimand A and B are both front-doors, but A dominates B (hence all P_i have $s_i \geq 0$). We again see that the plug-in approach is highly accurate. In fact, for data sizes $m \geq 200$ most probabilities of selecting estimand A over B exceed 90% except when the estimands are of nearly equal strength ($s_i \approx 0$).

In Figure 5, we display the results of the same comparisons using the bootstrap estimator with 100 bootstrap samples. In diagram $\mathcal{G}_1$, we see similar results to our plug-in approach, with prediction quality increasing with more data. However, in diagrams $\mathcal{G}_2$ and $\mathcal{G}_3$, prediction quality is noticeably worse than the plug-in. A potential benefit of the bootstrap is that it estimates the finite-sample variance rather than an asymptotic approximation. However, because we evaluate correctness against the asymptotic variance, this mismatch in targets can lead to the appearance of more error.

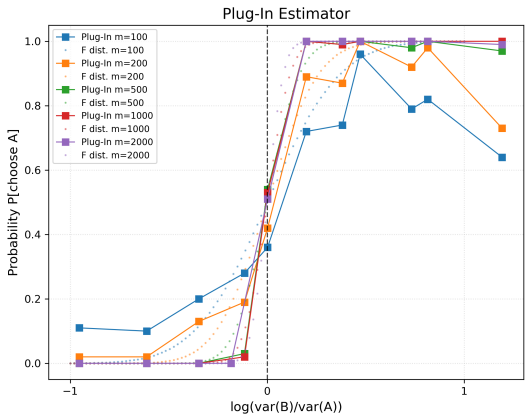
Given the significant computational savings and equal or superior performance, these experiments suggest that the plug-in estimator is a reliable and practical technique for selecting among potential estimands.

6 CONCLUSION

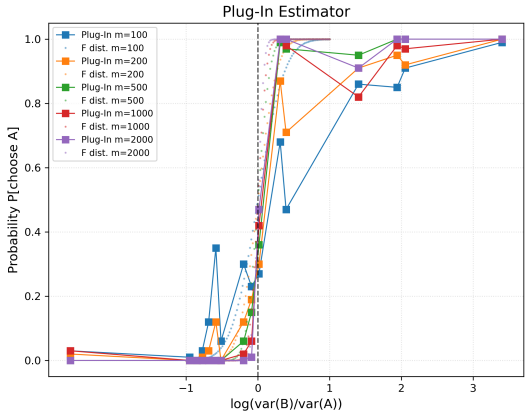
Our paper examines the statistical efficiency of causal effect estimation by analyzing and comparing the asymptotic variance for front-door and back-door estimands. We establish a dominance result for front-door estimands and their adjustment sets, which is inspired by previous work on back-door estimands. Specifically, given a particular conditional independence relation, we prove that removing extraneous variables from a front-door adjustment set cannot increase the asymptotic variance of the front-door estimand. We also show via counterexample that another known back-door dominance result does *not* hold for front-door adjustment.

In many settings however, the graph alone is insufficient to establish which estimand should be preferred – the asymptotic relationship between their variances depends on the underlying probabilities P. To address this case, we introduce two data driven methods for estimating and comparing the asymptotic variances of estimands, based on both the causal diagram and observed data. Our experiments show that it is feasible to approximate the asymptotic variance of competing estimands directly from observational samples, and suggest that the plug-in estimator may be both more accurate and computationally simpler. We find that, as sample size increases and variance gaps widen, we can reliably distinguish superior estimands.

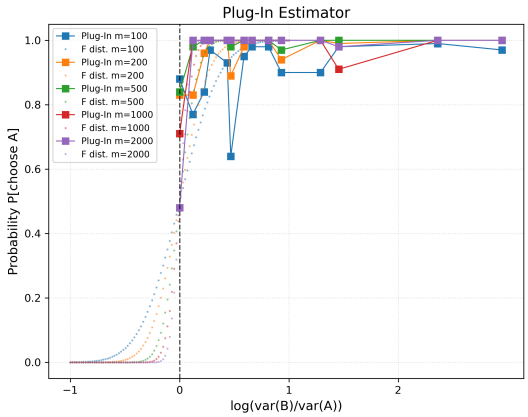
An interesting open direction is to characterize the efficiency of more complex and general estimands derived via the ID algorithm or do-calculus. While our empirical techniques should remain applicable, the full class of identifiable queries often yield nested functionals which are more difficult to analyze. In particular, comparing the variance of estimand expressions based off graph structure, and even enumerating the collection of valid estimand expressions for a given query in the first place, become substantially harder problems.



(a) Causal Diagram 1 ($\mathcal{G}_1$)

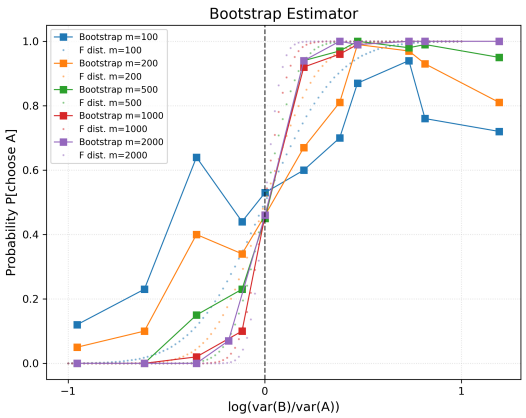


(b) Causal Diagram 2 ($\mathcal{G}_2$)

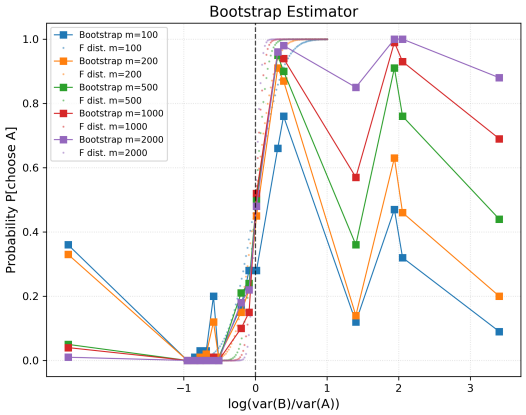


(c) Causal Diagram 3 ($\mathcal{G}_3$)

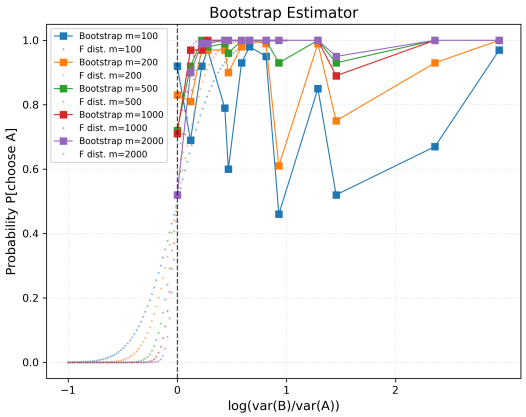
Figure 4: Results for the plug-in estimator for various data set sizes m. The F-distribution is plotted as dotted lines. Each square point indicates a particular true probability distribution $\{P_i\}$ corresponding to the causal diagram, and the probability is calculated from counting the frequency of choosing $\sigma^2_{x^*}(A; P_i) < \sigma^2_{x^*}(B; P_i)$ out of 100 data draws of size m.



(a) Causal Diagram 1 ($\mathcal{G}_1$)



(b) Causal Diagram 2 ($\mathcal{G}_2$)



(c) Causal Diagram 3 ($\mathcal{G}_3$)

Figure 5: Results for the bootstrap estimator for various data set sizes m. The F-distribution is plotted as dotted lines. Each square point indicates a particular true probability distribution $\{P_i\}$ corresponding to the causal diagram, and the probability is calculated from counting the frequency of choosing $\sigma^2_{x^*}(A; P_i) < \sigma^2_{x^*}(B; P_i)$ out of 100 data draws of size m.

Acknowledgements

This work was supported in part by NSF grant CNS-2321786.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research*, 23:1–76, 2022.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1998.

Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

Rina Dechter. *Reasoning with Probabilistic and Deterministic Graphical Models: Exact Algorithms*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2013.

Rina Dechter, Anna Raichev, Jin Tian, and Alexander Ihler. Graph-based complexity for causal effect by empirical plug-in. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan, editors, *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, pages 4798–4806. PMLR, 2025.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, 1993.

Isabel R. Fulcher, Ilya Shpitser, Stella Marealle, and Eric 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)*, 82(1):199–214, Feb 2020.

Tetiana Gorbach, Xavier de Luna, Juha Karvanen, and Ingeborg Waernbaum. Contrasting identifying assumptions of average causal effects: Robustness and semiparametric efficiency. *Journal of Machine Learning Research*, 24(197): 1–65, 2023. URL `https://jmlr.org/papers/v24/21-1392.html`.

F Richard Guo, Emilija Perković, and Andrea Rotnitzky. Variable elimination, graph reduction and the efficient g-formula. *Biometrika*, 110:739–761, November 2022.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Yimin Huang and Marco Valtorta. On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4): 363–408, 2008.

Hyunchai Jeong, Jin Tian, and Elias Bareinboim. Finding and listing front-door adjustment sets. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pages 33173–33185, 2022.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating causal effects using weighting-based estimators. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 10186–10193, 2020a.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. In *Advances in Neural Information Processing Systems 33*, 2020b.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 12113–12122, 2021a.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects on markov equivalence class through double machine learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 2021b.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

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

Anna Raichev, Alexander Ihler, Jin Tian, and Rina Dechter. Estimating causal effects from learned causal networks. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024)*, 2024.

Anna Raichev, Alexander Ihler, Jin Tian, and Rina Dechter. VarianceEstimation: Data-driven framework for selecting efficient causal estimands. `https://github.com/anniemeow/VarianceEstimation`, 2026. Accessed: July 2026.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, page 1219, 2006a.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Uncertainty in Artificial Intelligence*, pages 442–449, 2006b.

Ezequiel Smucler and Andrea Rotnitzky. A note on efficient minimum cost adjustment sets in causal graphical models. *Journal of Causal Inference*, 10(1):174–189, 2022.

Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, 109(1):49–65, 2022.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000.

Chie Taguchi and Manabu Kuroki. Variance-based difference between graphical identification conditions of causal effects in linear structural equation models. *Statistics & Probability Letters*, 228:110561, 2026.

Jin Tian. *Studies in Causal reasoning and Learning*. PhD thesis, University of California, Los Angeles, 2002.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI)*, pages 567–573, 2002.

Santtu Tikka and Juha Karvanen. Enhancing identification of causal effects by pruning. *Journal of Machine Learning Research*, 18:1–23, 06 2018a.

Santtu Tikka and Juha Karvanen. Simplifying probabilistic expressions in causal inference. *Journal of Machine Learning Research*, 18(36):1–30, 2018b.

A. W. van der Vaart. *Asymptotic Statistics*. Number 9780521784504 in Cambridge Books. Cambridge University Press, none edition, January 2000.

Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the UAI 2014 Workshop Causal Inference: Learning and Prediction co-located with 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014), Quebec City, Canada, July 27, 2014*, volume 1274, pages 11–24. CEUR-WS.org, 2014.

Marcel Wienöbst, Benito van der Zander, and Maciej Liskiewicz. Linear-time algorithms for front-door adjustment in causal graphs. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *AAAI*, pages 20577–20584. AAAI Press, 2024.

Appendix: Variance Estimation and Selecting Estimands for Causal Effect Queries

Anna K Raichev[1] **Rina Dechter**[1] **Jin Tian**[2] **Alexander Ihler**[1]

[1]Computer Science Dept., University of California, Irvine, Irvine, California, USA
[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates,

A PROOF OF THEOREM 1

Here we prove:

Theorem 1. *Given causal diagram $\mathcal{G}$ and identifiable query on X and Y, let $\boldsymbol{A} \cup \boldsymbol{B}$ be a front-door adjustment set with $\boldsymbol{A}$ and $\boldsymbol{B}$ disjoint, and $Y \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{A}, X$. Then $\boldsymbol{A}$ is also a front-door adjustment set and for any intervention $X = x^*$ and for all $P \in \mathcal{M}(\mathcal{G})$,*

$$\sigma^2_{x^*,fd}(\boldsymbol{A} \cup \boldsymbol{B}; P) - \sigma^2_{x^*,fd}(\boldsymbol{A}; P) \;=\; \mathbb{E}\Big[\operatorname{Var}(Y|\, X, \boldsymbol{A}) \operatorname{Var}\big[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \mid X, \boldsymbol{A}\big]\Big] \;\geq\; 0.$$

Consequently, $\boldsymbol{A} \preccurlyeq \boldsymbol{A} \cup \boldsymbol{B}$.

Proof. We start with the definition of the efficient influence function for adjustment set $\boldsymbol{Z}$

$$\varphi_{P,x^*}(X, Y, \boldsymbol{Z}) = \big(Y - \mathbb{E}[Y|X, \boldsymbol{Z}]\big)\frac{f(\boldsymbol{Z}|x^*)}{f(\boldsymbol{Z}|X)} + \frac{\mathbb{I}(X = x^*)}{f(X)}\Big(\sum_x \mathbb{E}[Y|x, \boldsymbol{Z}]f(x) - \sum_{x\overline{z}} \mathbb{E}[Y|x, \overline{z}]f(\overline{z}|X)f(x)\Big) + \sum_z \mathbb{E}[Y|X, z]f(z|x^*) - \psi \quad (9)$$

Since $Y \perp\!\!\!\perp \boldsymbol{B} \mid X, \boldsymbol{A}$, we have that $\mathbb{E}[Y|X, \boldsymbol{A}, \boldsymbol{B}] = \mathbb{E}[Y|X, \boldsymbol{A}]$.

By the law of total variance we can write,

$$\operatorname{Var}_P\{\varphi_{P,x^*}(X, Y, \boldsymbol{A}, \boldsymbol{B})\} = \operatorname{Var}_P\{\mathbb{E}_P(\varphi_{P,x^*}(X, Y, \boldsymbol{A}, \boldsymbol{B}) \mid X, Y, \boldsymbol{A})\} + \mathbb{E}_P\{\operatorname{Var}_P(\varphi_{P,x^*}(X, Y, \boldsymbol{A}, \boldsymbol{B}) \mid X, Y, \boldsymbol{A})\} \quad (10)$$

We will focus on simplifying both terms separately; we first show that,

$$\operatorname{Var}_P\{\mathbb{E}_P(\varphi_{P,x^*}(X, Y, \boldsymbol{A}, \boldsymbol{B}) \mid X, Y, \boldsymbol{A})\} = \operatorname{Var}_P\{\varphi_{P,x^*}(X, Y, \boldsymbol{A})\} = \sigma^2_{x^*,fd}(\boldsymbol{A}; P).$$

From the definition of $\varphi_{P,x}$ in Eq. (9) we have,

$$\begin{aligned}&\mathbb{E}_P[\varphi_{P,x}(X, Y, \boldsymbol{A}, \boldsymbol{B}) \mid X, Y, \boldsymbol{A}] \\ &= \mathbb{E}_p\Big[\big(Y - \mathbb{E}[Y|X, \boldsymbol{A}, \boldsymbol{B}]\big)\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} + \frac{\mathbb{I}(X = x^*)}{f(X)}\Big(\sum_x \mathbb{E}[Y|x, \boldsymbol{A}, \boldsymbol{B}]f(x) - \sum_{x,\overline{b},\overline{g}} \mathbb{E}[Y|x, \overline{b}, \overline{g}]f(\overline{b}, \overline{g}|X)f(x)\Big) \\ &\qquad + \sum_{b,g} \mathbb{E}[Y|X, b, g]f(b, g|x^*) - \psi \;\Big|\; X, Y, \boldsymbol{A}\Big].\end{aligned}$$

Then, since $Y \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{A}, X$, we have $\mathbb{E}_P(Y \mid X, \boldsymbol{A}, \boldsymbol{B}) = \mathbb{E}_P(Y \mid X, \boldsymbol{A})$:

$$= \mathbb{E}_p\Big[\big(Y - E[Y|X,\boldsymbol{A}]\big)\frac{f(\boldsymbol{A},\boldsymbol{B}|x^*)}{f(\boldsymbol{A},\boldsymbol{B}|X)} + \frac{\mathbb{I}(X=x^*)}{f(X)}\Big(\sum_x \mathbb{E}[Y|x,\boldsymbol{A}]f(x) - \sum_{x,\bar{b},\overline{g}} \mathbb{E}[Y|x,\overline{g}]f(\bar{b},\overline{g}|X)f(x)\Big) + \sum_{b,g}\mathbb{E}[Y|X,g]f(b,g|x^*) - \psi \;\Big|\; X,Y,\boldsymbol{A}\Big]$$

The term $(Y - \mathbb{E}_p(Y|X,\boldsymbol{A})$ is a constant when conditioning on X, Y, G, yielding

$$= \big(Y - \mathbb{E}[Y|X,\boldsymbol{A}]\big)\mathbb{E}_P\Big[\frac{f(\boldsymbol{A},\boldsymbol{B}|x^*)}{f(\boldsymbol{A},\boldsymbol{B}|X)} + \frac{\mathbb{I}(X=x^*)}{f(X)}\Big(\sum_x \mathbb{E}[Y|x,\boldsymbol{A}]f(x) - \sum_{x,\bar{b},\overline{g}} \mathbb{E}[Y|x,\overline{g}]f(\bar{b},\overline{g}|X)f(x)\Big) + \sum_{b,g}\mathbb{E}_p[Y|X,g]f(b,g|x^*) - \psi \;\Big|\; X,Y,\boldsymbol{A}\Big]$$

Applying linearity of expectation and $\frac{f(\boldsymbol{A}|x^*)}{f(\boldsymbol{A}|X)} = \mathbb{E}_p[\frac{f(\boldsymbol{A},\boldsymbol{B}|x^*)}{f(\boldsymbol{A},\boldsymbol{B}|X)}\Big|X,Y,\boldsymbol{A}]$,

$$\begin{aligned}= \big(Y - \mathbb{E}[Y|X,\boldsymbol{A}]\big)\,\frac{f(\boldsymbol{A}|x^*)}{f(\boldsymbol{A}|X)} \\ + \frac{\mathbb{I}(X=x^*)}{f(X)}\left(\mathbb{E}_p\Big[\sum_x \mathbb{E}[Y|x,\boldsymbol{A}]f(x) \;\Big|\; X,Y,\boldsymbol{A}\Big] - \mathbb{E}_p\Big[\sum_{x,\bar{b},\overline{g}} \mathbb{E}[Y|x,\overline{g}]f(\bar{b},\overline{g}|X)f(x) \;\Big|\; X,Y,\boldsymbol{A}\Big]\right) \\ + \mathbb{E}_p\Big[\sum_{b,g}\mathbb{E}[Y|X,g]f(b,g|x^*) - \psi \;\Big|\; X,Y,\boldsymbol{A}\Big]\end{aligned}$$

Finally, taking expectations and using the property that $\mathbb{E}[f(x)|x] = f(x)$,

$$\begin{aligned}&= \big(Y - \mathbb{E}[Y|X,\boldsymbol{A}]\big)\,\frac{f(\boldsymbol{A}|x^*)}{f(\boldsymbol{A}|X)} + \frac{\mathbb{I}(X=x^*)}{f(X)}\Big(\sum_x \mathbb{E}_p[Y|x,\boldsymbol{A}]f(x) - \sum_{x,\overline{g}} \mathbb{E}_p[Y|x,\overline{g}]f(\overline{g}|X)f(x)\Big) \\ &\qquad + \sum_g \mathbb{E}[Y|X,g]f(g|x^*) - \psi \\ &= \varphi_{P,x^*}(X,Y,\boldsymbol{A})\end{aligned}$$

by definition (9). Thus for the first term,

$$\mathrm{Var}_P\{\mathbb{E}_P(\varphi_{P,x^*}(X,Y,\boldsymbol{A},\boldsymbol{B}) \mid X,Y,\boldsymbol{A})\} = \mathrm{Var}_P\{\varphi_{P,x^*}(X,Y,\boldsymbol{A})\} = \sigma^2_{x^*,fd}(\boldsymbol{A};P).$$

Next we consider the second term in the law of total variance:

$$\begin{aligned}&\mathrm{Var}(\varphi_{P,x^*}(X,Y,\boldsymbol{A},\boldsymbol{B})|X,Y,\boldsymbol{A})) \\ &= \mathrm{Var}\left[(Y - \mathbb{E}[Y|X,\boldsymbol{A},B])\frac{f(\boldsymbol{A},\boldsymbol{B}|x^*)}{f(\boldsymbol{A},\boldsymbol{B}|X)} + \frac{\mathbb{I}(X=x^*)}{f(X)}\Big(\sum_x \mathbb{E}[Y|x,\boldsymbol{A},\boldsymbol{B}]f(x) - \sum_{x,\bar{b},\overline{g}} \mathbb{E}[Y|x,\bar{b},\overline{g}]f(\bar{b},\overline{g}|X)f(x)\Big)\right. \\ &\qquad \left. + \sum_{b,g}\mathbb{E}[Y|X,b,g]f(b,g|x^*) - \psi \;\Big|\; X,Y,\boldsymbol{A}\right].\end{aligned}$$

Applying $\mathbb{E}_P[Y \mid X,\boldsymbol{A},\boldsymbol{B}] = \mathbb{E}_P[Y \mid X,\boldsymbol{A}]$ and simplifying sums,

$$= \mathrm{Var}\left[\big(Y - \mathbb{E}[Y|X,\boldsymbol{A}]\big)\frac{f(\boldsymbol{A},\boldsymbol{B}|x^*)}{f(\boldsymbol{A},\boldsymbol{B}|X)} + \frac{\mathbb{I}(X=x^*)}{f(X)}\Big(\sum_x \mathbb{E}[Y|x,\boldsymbol{A}]f(x) - \sum_{x,\overline{g}} \mathbb{E}[Y|x,\overline{g}]f(\overline{g}|X)f(x)\Big) + \sum_g \mathbb{E}[Y|X,g]f(g|x^*) - \psi \;\Big|\; X,Y,\boldsymbol{A}\right]$$

Since, when conditioning on $X, Y, \boldsymbol{A}$, the only non-constant terms are those containing $\boldsymbol{B}$, and $\mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X)$ for α, β constant:

$$= \mathrm{Var}\left[\big(Y - \mathbb{E}[Y|X, \boldsymbol{A}]\big) \frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \;\middle|\; X, Y, \boldsymbol{A} \right]$$
$$= \big(Y - \mathbb{E}[Y|X, \boldsymbol{A}]\big)^2 \; \mathrm{Var}\left[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \;\middle|\; X, \boldsymbol{A} \right]$$

Now, taking the expected value gives,

$$\mathbb{E}_p[\mathrm{Var}(\varphi_{P,x^*}(X, Y, \boldsymbol{A}, \boldsymbol{B})|X, Y, \boldsymbol{A}))] = \mathbb{E}_p\Big[\big(Y - \mathbb{E}[Y|X, \boldsymbol{A}]\big)^2 \mathrm{Var}[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \,|X, \boldsymbol{A}] \Big]$$
$$= \mathbb{E}_{X,\boldsymbol{A}}\Big[\mathbb{E}_p\Big[\big(Y - \mathbb{E}[Y|X, \boldsymbol{A}]\big)^2 \mathrm{Var}[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \,|X, \boldsymbol{A}] \mid X, \boldsymbol{A} \Big] \Big].$$

Since $Y \perp\!\!\!\perp \boldsymbol{B} \mid X, \boldsymbol{A}$, this is,

$$= \mathbb{E}\left[\mathbb{E}_p\Big[\big(Y - \mathbb{E}[Y|X, \boldsymbol{A}]\big)^2 | X, \boldsymbol{A} \Big] \mathbb{E}_p\Big[\mathrm{Var}[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \,|X, \boldsymbol{A}] \,|X, \boldsymbol{A} \Big] \right]$$

and applying the identities $\mathrm{Var} = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$ and $\mathbb{E}[f(X, A)|X, A] = f(X, A)$,

$$= E\Big[\mathrm{Var}(Y|\, X, \boldsymbol{A}) \mathrm{Var}[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \mid X, \boldsymbol{A}] \Big]$$

Variance is always nonnegative, and so this expectation is also nonnegative.

Finally, combining these two results gives,

$$\sigma^2_{x^*,fd}(\boldsymbol{A} \cup \boldsymbol{B}; P) = \mathrm{Var}_P\{\varphi_{P,x^*}(X, Y, \boldsymbol{A}, \boldsymbol{B})\} = \sigma^2_{x^*,fd}(\boldsymbol{A}; P) + E\Big[\mathrm{Var}(Y|\, X, \boldsymbol{A}) \mathrm{Var}[\frac{f(\boldsymbol{A}, \boldsymbol{B}|x^*)}{f(\boldsymbol{A}, \boldsymbol{B}|X)} \mid X, \boldsymbol{A}] \Big]$$

showing

$$\sigma^2_{x^*,fd}(\boldsymbol{A} \cup \boldsymbol{B}; P) - \sigma^2_{x^*,fd}(\boldsymbol{A}; P) \geq 0.$$

□

B BOOTSTRAP VARIANCE ESTIMATOR ALGORITHM

Our bootstrap variance estimation is a straightforward application of the bootstrap technique to the statistic defined by evaluating the estimand. We provide it in algorithmic form here for clarity.

Algorithm 3 Bootstrap Variance Estimator [Efron and Tibshirani, 1993]

Require: Observed dataset $\mathcal{D}$, number of bootstrap resamples B, estimand function $\psi(\cdot)$
Ensure: Bootstrap variance estimate $\widehat{\text{Var}}_{\text{boot}}(\hat{\psi})$, optionally selected estimand $\hat{\psi}^*$

1: **Construct** the empirical distribution of the observed data:

$$\hat{P}_{\mathcal{D}}(\boldsymbol{V}) \;=\; \frac{1}{|\mathcal{D}|}\sum_{v\in\mathcal{D}}\delta(\boldsymbol{V}-v)$$

2: **for** $b = 1, 2, \ldots, B$ **do**
3: **Resample** a bootstrap dataset $\mathcal{D}^{(b)}$ of size $|\mathcal{D}|$ by drawing i.i.d. samples **with replacement** from $\hat{P}_{\mathcal{D}}$:

$$\mathcal{D}^{(b)} \overset{\text{i.i.d.}}{\sim} \hat{P}_{\mathcal{D}} \quad (|\mathcal{D}| \text{ draws})$$

4: **Compute** both estimands on the bootstrap sample:

$$\hat{\psi}_A^{(b)} = \hat{\psi}_A\left(\mathcal{D}^{(b)}\right), \quad \hat{\psi}_B^{(b)} = \hat{\psi}_B\left(\mathcal{D}^{(b)}\right)$$

5: **end for**
6: **Compute** the bootstrap sample means

$$\bar{\psi}_{A,\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\psi}_A^{(b)}, \qquad \bar{\psi}_{B,\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\psi}_B^{(b)}$$

7: **Compute** the bootstrap variances:

$$\widehat{\text{Var}}_{\text{boot}}\left(\hat{\psi}_A\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\psi}_A^{(b)} - \bar{\psi}_{A,\text{boot}}\right)^2$$

$$\widehat{\text{Var}}_{\text{boot}}\left(\hat{\psi}_B\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\psi}_B^{(b)} - \bar{\psi}_{B,\text{boot}}\right)^2$$

(Optional – Estimand Selection)
8: **Select** the estimand minimizing the bootstrap variance:

$$\hat{\psi}^* = \begin{cases}\psi_A(\cdot) & \text{if} \quad \widehat{\text{Var}}_{\text{boot}}(\hat{\psi}_A) \le \widehat{\text{Var}}_{\text{boot}}(\hat{\psi}_B)\\ \psi_B(\cdot) & \text{otherwise}\end{cases}$$

9: **Return** variance estimates $\widehat{\text{Var}}_{\text{boot}}(\hat{\psi}_A)$, $\widehat{\text{Var}}_{\text{boot}}(\hat{\psi}_B)$, and the selected estimand $\hat{\psi}^*$.

C PLUG-IN VARIANCE ESTIMATOR ALGORITHM

In this section we expand on the plug-in procedure for estimating the asymptotic variance using the empirical distribution. We are given i.i.d. samples $\mathcal{D} = \{v^{(1)}, \ldots, v^{(m)}\}$ from a discrete distribution over variables $V = (V_1, \ldots, V_n)$, and an estimand $\hat{\psi}(\rho)$ that is a differentiable function of the conditional probability tables $P(V_i \mid \boldsymbol{V}_{<i})$, and wish to compute a plug-in estimate of the asymptotic variance of $\hat{\psi}$.

Input.

- An ordering $V_1, \ldots, V_n$.
- A differentiable functional $\hat{\psi}(\rho)$ expressed in terms of conditional probabilities

  $$\rho_{\cdot,i,v} \equiv P(V_i \mid \boldsymbol{V}_{<i} = v),$$

  for all i and all configurations v of $\boldsymbol{V}_{<i}$. (Note that these conditional distributions may depend only on some subset of $\boldsymbol{V}_{<i}$.)
- Data $\mathcal{D} = \{v^{(1)}, \ldots, v^{(m)}\}$ with $v^{(j)} = (v_1^{(j)}, \ldots, v_n^{(j)})$.

Output. The plug-in estimate of the asymptotic variance, $\widehat{\mathrm{Var}}[\hat{\psi}(\rho)]$.

Step 1: Estimate conditional probability tables (plug-in $\hat{\rho}$).

For each variable V_i and each configuration v of $\boldsymbol{V}_{<i}$, compute the number of observations

$$m_{i,v} = \sum_{j=1}^{m} \mathbb{1}\{\boldsymbol{V}_{<i}^{(j)} = v\}.$$

For each value x in the support of V_i, define the empirical estimator

$$\hat{\rho}_{x,i,v} = \frac{\sum_{j=1}^{m} \mathbb{1}\{\boldsymbol{V}_{<i}^{(j)} = v,\, V_i^{(j)} = x\}}{m_{i,v}},$$

whenever $m_{i,v} > 0$. Collect all such conditional probability vectors into the parameter vector $\hat{\rho}$.

Step 2: Construct the block-diagonal Fisher information matrix.

For each conditional probability vector $\rho_{\cdot,i,v}$ of dimension $d_{i,v}$, write the free parameters as

$$\underline{\rho}_{i,v} = [\rho_{1,i,v}, \ldots, \rho_{d_{i,v}-1,i,v}],$$

with the final component determined by the normalization constraint. The Fisher information for this block is

$$I_{i,v}(\underline{\rho}) = m_{i,v} \left(\mathrm{diag}(\underline{\rho}_{i,v}^{-1}) + \frac{1}{1 - \mathbf{1}^\top \underline{\rho}_{i,v}} \mathbf{1}\mathbf{1}^\top \right),$$

where $\mathbf{1}$ is a vector of ones of length $d_{i,v} - 1$.

Because the likelihood factorizes across (i, v), the full Fisher information matrix is block diagonal:

$$I(\rho) = \mathrm{blockdiag}\big(I_{i,v}(\rho) : \text{all } (i,v)\big).$$

Evaluate this matrix at $\rho = \hat{\rho}$.

Step 3: Evaluate the estimand and its gradient.

Compute the plug-in estimate of the estimand,

$$\hat{\psi} = \psi(\hat{\rho}).$$

along with the gradient with respect to the free parameters,

$$g(\hat{\underline{\rho}}) = \nabla_{\underline{\rho}} \psi(\rho)\big|_{\underline{\rho}=\hat{\underline{\rho}}}.$$

In practice, this gradient can be evaluated using automatic differentiation applied to the computation of $\psi(\rho)$. The free parameters $\underline{\rho}$ are set to their empirical values $\hat{\underline{\rho}}_{i,v}$, which form the leaves of the computation. We next construct the full probabilities $\hat{\rho}_{i,v}$ by $\hat{\rho}_{d_{i,v},i,v} = 1 - \sum_{j=1}^{d_{i,v}-1} \hat{\underline{\rho}}_{j,i,v}$, then apply the estimand expression to evaluate $\psi(\hat{\rho})$.

Step 4: Compute the plug-in asymptotic variance.

By the multivariate delta method, the asymptotic variance of $\hat{\psi}(\rho)$ under $\hat{\rho}$ is

$$\widehat{\mathrm{Var}}[\psi(\rho)] = g(\hat{\underline{\rho}})^\top I(\hat{\underline{\rho}})^{-1} g(\hat{\underline{\rho}}).$$

The block diagonal form of I ensures that we do not need to store or multiply any matrices of dimensions larger than the individual $d_{i,v}$:

$$\widehat{\mathrm{Var}}[\psi(\rho)] = g(\hat{\underline{\rho}})^\top I(\hat{\underline{\rho}})^{-1} g(\hat{\underline{\rho}}) = \sum_{i,v} g(\hat{\underline{\rho}}_{i,v})^\top I_{i,v}(\hat{\underline{\rho}})^{-1} g(\hat{\underline{\rho}}_{i,v})$$

Return $\widehat{\mathrm{Var}}[\psi(\rho)]$.

Condensing this procedure in a more concrete, algorithmic form gives Algorithm 4 below.

Algorithm 4 Plug-in Variance Estimator for $\psi(\rho)$

Require: Data $\mathcal{D} = \{v^{(1)}, \ldots, v^{(m)}\}$, ordering $V_1, \ldots, V_p$, estimands $\psi_A(\rho), \psi_B(\rho)$
Ensure: Plug-in variance estimates $\widehat{\mathrm{Var}}[\psi_A(\rho)]$, $\widehat{\mathrm{Var}}[\psi_B(\rho)]$, and optionally selected estimand $\hat{\psi}^*$

Estimate conditional probability tables:
for $i = 1$ to n **do**
  for each configuration v of $\boldsymbol{V}_{<i}$ **do**
    $m_{i,v} \leftarrow \sum_{j=1}^{m} \mathbb{1}\{\boldsymbol{V}_{<i}^{(j)} = v\}$
    for each value x in support of V_i **do**
      $\hat{\rho}_{x,i,v} \leftarrow \dfrac{1}{m_{i,v}} \displaystyle\sum_{j=1}^{m} \mathbb{1}\{\boldsymbol{V}_{<i}^{(j)} = v,\, V_i^{(j)} = x\}$
    end for
  end for
end for

Construct block-diagonal Fisher information
Initialize $I(\hat{\rho}) \leftarrow 0$
for each (i, v) **do**
  Let $\underline{\rho}_{i,v} = [\hat{\rho}_{1,i,v}, \ldots, \hat{\rho}_{d_{i,v}-1,i,v}]$
  Define block $\quad I_{i,v}(\underline{\hat{\rho}}) = m_{i,v} \left(\mathrm{diag}(\underline{\hat{\rho}}_{i,v}^{-1}) + \dfrac{1}{1 - \mathbf{1}^\top \underline{\hat{\rho}}_{i,v}} \mathbf{1}\mathbf{1}^\top \right)$
  Insert $I_{i,v}(\underline{\hat{\rho}})$ as a diagonal block of $I(\underline{\hat{\rho}})$
end for

Compute gradients of estimands
Extend $\underline{\hat{\rho}}_{i,v}$ to $\hat{\rho}_{i,v}$ using the normalization constraint
Evaluate $\hat{\psi}_A \leftarrow \psi_A(\hat{\rho})$
Compute gradients $g_A(\underline{\hat{\rho}}_{i,v}) \leftarrow \nabla_{\underline{\rho}_{i,v}} \psi_A(\underline{\rho})\big|_{\underline{\rho}=\underline{\hat{\rho}}}$ (via autodiff)
Evaluate $\hat{\psi}_B \leftarrow \psi_B(\hat{\rho})$
Compute gradients $g_B(\underline{\hat{\rho}}_{i,v}) \leftarrow \nabla_{\underline{\rho}_{i,v}} \psi_B(\underline{\rho})\big|_{\underline{\rho}=\underline{\hat{\rho}}}$ (via autodiff)

Compute plug-in variances
$\widehat{\mathrm{Var}}[\psi_A(\rho)] = g_A(\underline{\hat{\rho}})^\top I(\underline{\hat{\rho}})^{-1} g_A(\underline{\hat{\rho}}) = \displaystyle\sum_{i,v} g_A(\underline{\hat{\rho}}_{i,v})^\top I_{i,v}(\underline{\hat{\rho}})^{-1} g_A(\underline{\hat{\rho}}_{i,v})$
$\widehat{\mathrm{Var}}[\psi_B(\rho)] = g_B(\underline{\hat{\rho}})^\top I(\underline{\hat{\rho}})^{-1} g_B(\underline{\hat{\rho}}) = \displaystyle\sum_{i,v} g_B(\underline{\hat{\rho}}_{i,v})^\top I_{i,v}(\underline{\hat{\rho}})^{-1} g_B(\underline{\hat{\rho}}_{i,v})$

(Optional — Estimand Selection)
Select the estimand minimizing the plug-in variance:

$$\hat{\psi}^* = \begin{cases} \psi_A(\cdot) & \text{if} \quad \widehat{\mathrm{Var}}[\psi_A(\rho)] \le \widehat{\mathrm{Var}}[\psi_B(\rho)] \\ \psi_B(\cdot) & \text{otherwise} \end{cases}$$

return $\widehat{\mathrm{Var}}[\psi_A(\rho)]$, $\widehat{\mathrm{Var}}[\psi_B(\rho)]$, and the selected estimand $\hat{\psi}^*$