Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables

Anonymous Author(s)
Affiliation
Address
email

Abstract

The problem of selecting optimal backdoor adjustment sets to estimate causal effects in graphical models with hidden and conditioned variables is addressed. Previous work has defined optimality as achieving the smallest asymptotic variance and identified a graphical criterion for an optimal set for the case without hidden variables. For the case with hidden variables currently a sufficient graphical criterion of limited applicability exists. Here optimality is characterized as maximizing a certain adjustment information which allows to derive a necessary and sufficient graphical criterion for the existence of an optimal adjustment set and a definition and algorithm to construct it. Further, the optimal set is valid if and only if a valid adjustment set exists and has higher (or equal) adjustment information than the Adjust-set proposed in Perković et al. [Journal of Machine Learning Research, 18: 1–62, 2018] for any graph. The results translate to minimal estimation error for a class of estimators whose asymptotic variance follows a certain information-theoretic relation. Numerical experiments indicate that the asymptotic results also hold for relatively small sample sizes. Surprisingly, among the randomly created setups more than 80% fulfill the optimality conditions indicating that also in many real-world scenarios graphical optimality may hold.

1 Introduction

A standard problem setting in causal inference is to estimate the causal effect between two variables given a causal graphical model that specifies the assumed qualitative causal relations among observed variables (Pearl, 2009), including a possible presence of hidden confounding variables. The graphical model then allows to employ graphical criteria to identify valid adjustment sets, the most well-known being the *backdoor criterion* (Pearl, 1993) and the *generalized adjustment criterion* (Shpitser et al., 2010; Perković et al., 2015, 2018) providing a complete identification of all valid adjustment sets. Estimators of causal effects based on such a valid adjustment set as a covariate are then consistent, but for different adjustment sets the estimation error may strongly vary. An *optimal adjustment set* may be characterized as one that has minimal estimation variance. Following work by Kuroki and Cai (2004) and Kuroki and Miyakawa (2003), Henckel et al. (2019) (abbreviated HPM19 in the following) showed graphical optimality for linear models in the causally sufficient case where all relevant variables are observed. In Witte et al. (2020) an alternative characterization of the optimal adjustment set is discussed and the approach was integrated into the IDA algorithm Maathuis et al. (2009, 2010) that does not require the causal graph to be known and is based on causal discovery (Spirtes et al., 2000). Rotnitzky and Smucler (2019) extended the results in HPM19 to asymptotically linear non-parametric graphical models. HPM19's optimal adjustment set holds for the causally sufficient case (no hidden variables) and the authors gave an example with hidden variables where

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

optimality does not hold in general, i.e., the optimal adjustment set depends on the coefficients and noise terms, rather than just the graph. Most recently, Smucler et al. (2020) (SSR20) partially extended these results to the non-parametric hidden variables case together with *dynamic treatment regimes*, i.e., conditional causal effects. SSR20 provide a sufficient criterion for an optimal set to exist and a definition based on a certain undirected graph-construction using a result by van der Zander et al. (2019). However, their sufficient criterion is very restrictive and a current major open problem is a *necessary* and sufficient condition for an optimal adjustment set to exist in the hidden variable case and a corresponding definition of an optimal set.

Here this problem is solved. Optimality for conditional causal effects in the hidden variables case is characterized by an information-theoretic approach based on relating the estimator's asymptotic variance to an expression involving a certain difference of conditional mutual informations (CMIs) among the observed variables termed the adjustment information. Maximizing the adjustment information formalizes the common intuition to choose adjustment sets that maximally constrain the effect variable and minimally constrain the cause variable. The derived optimal adjustment set also has the property of minimum cardinality, i.e., no node can be removed without sacrificing optimality. Further, the optimal set is valid if and only if a valid adjustment set exists and has higher (or equal) adjustment information than the Adjust-set proposed in Perković et al. (2018) for any graph, whether graphical optimality holds or not. The results translate to minimal estimation error for a class of estimators whose asymptotic variance follows a certain information-theoretic relation that, at present, we could only verify for the linear case. Numerical experiments corroborate the theoretical results. More detailed preliminaries, proofs, algorithms, and further numerical experiments are given in the Supplementary sections A–D. Code is available as part of the python package [anomymized].

1.1 Problem setting and preliminaries

We consider causal effects in causal graphical models over a set of variables $\mathbf{V}$ with a joint density $\mathcal{P} = \mathcal{P}(\mathbf{V})$ that is consistent with an acyclic directed mixed graph (ADMG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$. See Fig. 1A for an example graph. The results also hold for *Maximal Ancestral Graphs* (MAG) (Richardson and Spirtes, 2002) without selection variables. For detailed graphical models preliminaries see Supplementary Section A. Kinships are defined as usual: parents $pa(X)$ for "$\bullet\!\!\rightarrow X$", spouses $sp(X)$ for "$X\!\leftrightarrow\!\bullet$", children $ch(X)$ for "$X\!\rightarrow\!\bullet$", and correspondingly descendants des and ancestors an. The mediator nodes on causal paths from X to Y are denoted $\mathbf{M} = \mathbf{M}(X, Y)$ and exclude X and Y. In this work we only consider a univariate intervention variable X and effect variable Y. We simplify set notation and denote unions of variables as $\{W\} \cup \mathbf{M} \cup \mathbf{A} = W\mathbf{MA}$.

A (possibly empty) set of adjustment variables $\mathbf{Z}$ for the total causal effect of X on Y is called *valid* relative to (X, Y) if the interventional distribution for setting $do(X = \mathbf{x})$ (Pearl, 2009) factorizes as $p(Y|do(X = \mathbf{x})) = \int_{\mathbf{Z}} p(Y|\mathbf{x}, \mathbf{z})p(\mathbf{z})d\mathbf{z}$ for non-empty $\mathbf{Z}$ and as $p(Y|do(X = \mathbf{x})) = p(Y|\mathbf{x})$ for empty $\mathbf{Z}$. Valid adjustment sets, here denoted $\mathcal{Z}$, can be read off from a given causal graph using the generalized adjustment criterion (Perković et al., 2015, 2018) which generalizes Pearl's back-door criterion (Pearl, 2009). To this end define

$$\mathbf{forb}(X, Y) = X \cup des(Y\mathbf{M}) \tag{1}$$

(henceforth just denoted as $\mathbf{forb}$). A set $\mathbf{Z}$ is valid if all of the following conditions hold: (i) $\mathcal{G}$ is adjustment amenable relative to (X, Y) (defined in Sect. A), (ii) $\mathbf{Z} \cap \mathbf{forb} = \emptyset$, and (iii) all non-causal paths from X to Y are blocked by $\mathbf{Z}$. An adjustment set is called *minimal* if no strict subset of $\mathbf{Z}$ is still valid. Condition (i) is independent of any adjustment set $\mathbf{Z}$ and only relevant for MAGs since ADMGs and DAGs are always amenable. The validity conditions can in principle be manually checked directly from the graph, but, more conveniently, Perković et al. (2018) define an adjustment set called 'Adjust' that is valid if and only if a valid adjustment set exist. In our setting including conditioning variables $\mathbf{S}$ we call this set the *valid ancestors* defined as

$$\mathbf{vancs}(X, Y, \mathbf{S}) = an(XY\mathbf{S}) \setminus \mathbf{forb} \tag{2}$$

and refer to this set as $\mathbf{vancs}$ or Adjust-set.

The quantity of interest is the average total causal effect of an intervention to set X to x vs. x' on the effect variable Y given a set of selected (conditioned) variables $\mathbf{S} = \mathbf{s}$ with $\mathbf{S} \cap des(X) = \emptyset$

$$\Delta_{yxx'|\mathbf{s}} = E(Y|do(x), \mathbf{s}) - E(Y|do(x'), \mathbf{s})\,. \tag{3}$$

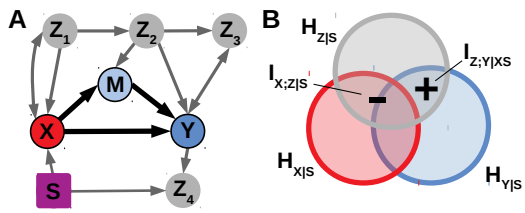

Figure 1: (**A**) Problem setting of optimal adjustment sets in causal graphs with hidden variables represented through bi-directed edges. Shown is an ADMG where there can be more than one edge between two nodes (here X and Z_1). The goal is to estimate the total causal effect of X on Y potentially through mediators $\mathbf{M}$, and given conditioned variables $\mathbf{S}$. The task is to select a valid adjustment set $\mathbf{Z}$ such that the estimator has minimal variance. (**B**) For a certain class of estimators a minimal estimation variance can be translated into an information-theoretical optimization problem, here visualized in a Venn diagram. An optimal adjustment set $\mathbf{Z}$ must maximize the adjustment information $I_{\mathbf{Z};Y|X\mathbf{S}} - I_{X;\mathbf{Z}|\mathbf{S}}$.

We denote an estimator given a valid adjustment set $\mathbf{Z}$ as $\widehat{\Delta}_{yxx'|\mathbf{S}.\mathbf{Z}}$. Current results on optimal adjustment sets consider two model classes for estimating causal effects, the causal linear model with possibly non-Gaussian error terms (linear regression estimator) in HPM19 and a class of regular asymptotically linear estimators (SSR20). For given $y, \mathbf{s}, x$ and x', the asymptotic distribution of estimators from these classes depends only on $\mathbf{Z}$ (see SSR20 for further details on such estimators). In the linear case $\Delta_{yxx'|\mathbf{s}}$ for $x = x' + 1$ corresponds to the regression coefficient $\beta_{YX\cdot\mathbf{ZS}}$ in the regression of Y on X, $\mathbf{Z}$, and $\mathbf{S}$. The ordinary least squares (OLS) estimator $\hat{\beta}_{YX\cdot\mathbf{ZS}}$ a consistent estimator of $\beta_{YX\cdot\mathbf{ZS}}$.

Our proposed approach to optimal adjustment sets is based on information theory (Cover and Thomas, 2006). The main quantity of interest there is the conditional mutual information (CMI) defined as a difference $I_{X;Y|Z} = H_{Y|Z} - H_{Y|ZX}$ of two (conditional) Shannon entropies $H_{Y|X} = -\int_{x,y} p(x,y) \ln p(y|x) dxdy$. Its main properties are non-negativity, $I_{X;Y|Z} = 0$ if and only if $X \perp\!\!\!\perp Y|Z$, and the chain rule $I_{XW;Y|Z} = I_{X;Y|Z} + I_{W;Y|ZX}$. All random variables in a CMI can be multivariate.

Figure 1A illustrates the problem setting: We are interested in the total causal effect of X on Y (conditioned on $\mathbf{S}$), which is here due to a direct link and an indirect causal path through a mediator M. There are six valid backdoor adjustment sets: $\mathbf{Z} = Z_1,\ Z_2,\ Z_1Z_2,\ Z_2Z_3,\ Z_1Z_3,\ Z_1Z_2Z_3$. Z_4 cannot be included in any set because it is a descendant of $Y\mathbf{M}$. Here $\mathbf{vancs}(X,Y) = Z_1Z_2S$. The question is which of these sets is statistically optimal in that it minimizes the estimation variance?

2 Optimal adjustment sets

2.1 Information-theoretic characterization

We information-theoretically formalize the intuition to choose an adjustment set $\mathbf{Z}$ that maximally constrains the effect variable Y and minimally constrains the cause variable X. In terms of CMIs and given selected fixed conditions $\mathbf{S}$ the quantity to maximize can be stated as follows.

Definition 1 (Adjustment information)**.** *Consider a causal effect of X on Y for an adjustment set $\mathbf{Z}$ given a condition set $\mathbf{S}$. The* (conditional) adjustment (set) information *is defined as*

$$J_{XY|\mathbf{S}.\mathbf{Z}} \equiv I_{\mathbf{Z};Y|X\mathbf{S}} - I_{X;\mathbf{Z}|\mathbf{S}}\,. \tag{4}$$

If the other variables are clear from the context, we abbreviate $J_{\mathbf{Z}} = J_{XY|\mathbf{S}.\mathbf{Z}}$.

$J_{\mathbf{Z}}$ is not necessarily positive if the dependence between X and $\mathbf{Z}$ (given $\mathbf{S}$) is larger than that between $\mathbf{Z}$ and Y given $X\mathbf{S}$. Using the CMI definition, $J_{\mathbf{Z}}$ can be expressed as

$$J_{\mathbf{Z}} = \underbrace{H_{Y|X\mathbf{S}} - H_{X|\mathbf{S}}}_{\text{not related to } \mathbf{Z}} - \underbrace{(H_{Y|X\mathbf{ZS}} - H_{X|\mathbf{ZS}})}_{\text{adjustment entropy}}\,. \tag{5}$$

Here $H_{Y|X\mathbf{S}} - H_{X|\mathbf{S}}$ is fixed by the problem setup. Now, given $(X, Y, \mathbf{S})$, choosing a valid optimal set $\mathbf{Z}_{\text{optimal}} \in \mathcal{Z}$ such that the adjustment information $J_{\mathbf{Z}}$ is maximal makes the adjustment entropy minimal:

$$\mathbf{Z}_{\text{optimal}} \in \text{argmax}_{\mathbf{Z}\in\mathcal{Z}} J_{\mathbf{Z}} \iff \mathbf{Z}_{\text{optimal}} \in \text{argmin}_{\mathbf{Z}\in\mathcal{Z}} (H_{Y|X\mathbf{ZS}} - H_{X|\mathbf{ZS}}) \,. \tag{6}$$

Fig. 1B illustrates the two CMIs in Eq. (4) in a Venn diagram from which one can read off the intuition to choose among the valid $\mathbf{Z}$ such that, given $\mathbf{S}$, the information between $\mathbf{Z}$ and Y given X is maximized while minimizing the information between $\mathbf{Z}$ and X.

Now the question is for which causal effect estimators $\widehat{\Delta}_{yxx'|\mathbf{s}.\mathbf{z}}$ the intuition of maximizing the adjustment information $J_{\mathbf{Z}}$ leads to a minimal asymptotic estimation variance where we assume that $\widehat{\Delta}_{yxx'|\mathbf{s}.\mathbf{z}}$ is consistent due to a valid adjustment set and correct functional model specification. Consider the class of estimators whose (square-root of the) asymptotic variance can be expressed as

$$\sqrt{E[(\Delta_{yxx'|\mathbf{s}} - \widehat{\Delta}_{yxx'|\mathbf{s}.\mathbf{z}})^2]} = f(H_{Y|X\mathbf{ZS}} - H_{X|\mathbf{ZS}}) \,, \tag{7}$$

for a real-valued, strictly monotonously increasing function f of the adjustment entropy. One candidate is $f(\cdot) = \frac{1}{\sqrt{n}} e^{H_{Y|X\mathbf{ZS}} - H_{X|\mathbf{ZS}}}$ which is related to the lower bound of the estimation counterpart to Fano's inequality (Theorem 8.6.6 in Cover and Thomas (2006)). At present, we were only able to derive the corresponding relation to (7) for a linear Gaussian causal model. The lower bound of Fano's inequality is obtained for Gaussian distributed estimators. Then, for the coefficient $\beta_{YX\cdot\mathbf{ZS}}$ with the entropies given by $H(Y|X\mathbf{ZS}) = \frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2_{Y|X\mathbf{ZS}})$ and $H(X|\mathbf{ZS}) = \frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2_{X|\mathbf{ZS}})$ relation (7) becomes $f(\cdot) = \frac{1}{\sqrt{n}} \frac{\sigma_{Y|X\mathbf{ZS}}}{\sigma_{X|\mathbf{ZS}}}$ where $\sigma(\cdot|\cdot)$ denotes the square-root of the conditional variance. This relation is also the basis of the results in Henckel et al. (2019) and holds more generally for causal linear models that do not require the noise terms to be Gaussian. Further research will show whether relation (7) also holds for the class of regular asymptotically linear estimators. The intuition for the latter comes from the fact (see SSR20) that their asymptotic distribution depends only on $\mathbf{Z}$ (and in the present context also on $\mathbf{S} = \mathbf{s}$). Throughout the present paper we will assume the following.

Assumptions 1 (General setting and assumptions)**.** *We assume a causal graphical model over a set of variables $\mathbf{V}$ with a joint distribution $\mathcal{P} = \mathcal{P}(\mathbf{V})$ that is consistent with an ADMG $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ or a MAG without selection variables. We assume a non-zero causal effect from X on Y, potentially through a set of mediators $\mathbf{M}$, and given selected conditioned variables $\mathbf{S}$, where $\mathbf{S} \cap des(XY\mathbf{M}) = \emptyset$. If $\mathcal{G}$ is a MAG, then we assume that $\mathcal{G}$ is adjustment amenable relative to (X, Y). We denote the set of valid adjustment sets with $\mathcal{Z}$ and assume that at least one valid adjustment set (given $\mathbf{S}$) exists and, hence, the causal effect is identifiable (except when stated otherwise). Finally, we assume the usual Causal Markov Condition (implicit in semi-Markovian models) and Faithfulness.*

Assumptions 2 (Estimator class assumption)**.** *The model class of the causal effect estimator* (3) *is correctly specified and its asymptotic variance can be expressed as in relation* (7)*.*

Our goal is to provide graphical criteria for optimal adjustment sets, i.e., criteria that depend only on the structure of the graph $\mathcal{G}$ and not on the distribution.

Definition 2 (Graphical optimality)**.** *Given Assumptions 1 we say that* graphical optimality holds *if there is a $\mathbf{Z} \in \mathcal{Z}$ such that either there is no other $\mathbf{Z}' \neq \mathbf{Z} \in \mathcal{Z}$ or for all other $\mathbf{Z}' \neq \mathbf{Z} \in \mathcal{Z}$ and* all *distributions $\mathcal{P}$ consistent with $\mathcal{G}$ we have $J_{\mathbf{Z}} \geq J_{\mathbf{Z}'}$.*

This general definition only relates optimality of adjustment sets to the adjustment information J. defined in Eq. (4), but not to any particular estimator. It holds in principle also for multivariate $\mathbf{X}$ and $\mathbf{Y}$. The following lemma then relates $J_{\mathbf{Z}} \geq J_{\mathbf{Z}'}$ to the corresponding asymptotic variances of a given estimator.

Lemma 1 (Asymptotic variance and adjustment information)**.** *Given Assumptions 1 and an estimator fulfilling Assumptions 2, if and only if for two different adjustment sets $\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}$ we have $J_{\mathbf{Z}} \geq J_{\mathbf{Z}'}$, then the adjustment set $\mathbf{Z}$ has a smaller or equal asymptotic variance compared to $\mathbf{Z}'$.*

Proof. By Equations (5) and (7), $J_{\mathbf{Z}} \geq J_{\mathbf{Z}'}$ (for fixed $X, Y, \mathbf{S}$) is directly related to a smaller or equal asymptotic variance for $\mathbf{Z}$ compared to $\mathbf{Z}'$, and vice versa. □

Our main result builds on the following lemma which relates graphical optimality to information-theoretic inequalities.

Lemma 2 (Necessary and sufficient comparison criterion for existence of an optimal set)**.** *Given Assumptions 1, if and only if there is a $\mathbf{Z} \in \mathcal{Z}$ such that either there is no other $\mathbf{Z}' \neq \mathbf{Z} \in \mathcal{Z}$ or for all other $\mathbf{Z}' \neq \mathbf{Z} \in \mathcal{Z}$ and all probability densities $\mathcal{P}$ it holds that*

$$\underbrace{I_{\mathbf{Z}\setminus\mathbf{Z}';Y|\mathbf{Z}'X\mathbf{S}}}_{(i)} \geq \underbrace{I_{\mathbf{Z}'\setminus\mathbf{Z};Y|\mathbf{Z}X\mathbf{S}}}_{(iii)} \quad \textit{and} \quad \underbrace{I_{X;\mathbf{Z}'\setminus\mathbf{Z}|\mathbf{Z}\mathbf{S}}}_{(ii)} \geq \underbrace{I_{X;\mathbf{Z}\setminus\mathbf{Z}'|\mathbf{Z}'\mathbf{S}}}_{(iv)}, \tag{8}$$

then graphical optimality holds implying $J_{\mathbf{Z}} \geq J_{\mathbf{Z}'}$.

In SSR20 and HPM19 the corresponding conditional independence statements to the terms (iii) and (iv) in the inequalities (8) are used to prove their sufficient optimality criterion. Lemma 2 provides a way to check optimality, but not a very efficient one since (in principle) all valid subsets have to be compared with each other. SSR20 provide a sufficient (but not necessary) criterion purely based on ancestral relationships (see below). The inequalities (8) guide a construction of an optimal adjustment set $\mathbf{O}$ in comparison with any other valid adjustment set $\mathbf{Z}$. In essence, $\mathbf{O}$ should satisfy that all other sets have a weaker dependence with Y and larger dependence with X. The idea will be to construct $\mathbf{O}$ based on the parents and valid collider path nodes of $Y\mathbf{M}$ and to also add parents of collider paths. This approach maximizes the CMI $I_{\mathbf{O};Y|X\mathbf{S}}$ while minimizing $I_{X;\mathbf{O}|\mathbf{S}}$ (see Fig. 1).

The optimal adjustment set for the causally sufficient case is simply $\mathbf{P} = pa(Y\mathbf{M}) \setminus \mathbf{forb}$ and was derived in HPM19 and Rotnitzky and Smucler (2019). In Section B.2 the derivation is discussed from an information-theoretic perspective. In the case with hidden variables we need to account for bi-directed edges "$\leftrightarrow$" which considerably complicate the situation. Then the parents of $Y\mathbf{M}$ are not sufficient to block all non-causal paths. Further, just like conditioning on parents of $Y\mathbf{M}$ leads to optimality in the sufficient case since parents constrain information in $Y\mathbf{M}$, in the hidden variables case also conditioning on spouses of $Y\mathbf{M}$ constrains information about $Y\mathbf{M}$.

Example A. A simple graph (ADMG) to illustrate this is $X \rightarrow Y \leftrightarrow Z_1$ (shown with an additional $\mathbf{S}$ in Fig. 2A below, or Fig. 4 in SSR20). Here $\mathbf{Z} = \emptyset = \mathbf{vancs}$ is a valid set, but it is not optimal. Consider $\mathbf{O} = Z_1$, then term (iii) $= 0$ since $\mathbf{Z} \setminus \mathbf{O} = \emptyset$. Even though not needed to block non-causal paths (there is none), Z_1 still constrains information in Y while being independent of X (hence, term (iv) $= 0$) which leads to $J_{\mathbf{O}} > J_{\emptyset}$ according to the inequalities (8).

2.2 Definition of O-set

Not only direct spouses can constrain information in Y as Fig. 2B below illustrates. Since for $W \in Y\mathbf{M}$ the motif "$W \leftrightarrow \boxed{C} \leftarrow *$" is open, we can further constrain information by conditioning also on subsequent spouses and this chain of colliders only ends if we reach a tail again or there is no further adjacency. This leads to the notion of a *collider path* (related to the notion of a *district* in Evans and Richardson (2014)).

Definition 3 ((Valid) collider paths)**.** *Given a graph $\mathcal{G}$, a* collider path *of W for $k \geq 1$ is defined by a sequence of edges $W \leftrightarrow C_1 \leftrightarrow \cdots \leftrightarrow C_k$. We denote the set of path nodes (excluding W) along a path indexed by i as π^i_W. Using the set of valid ancestors $\mathbf{vancs} = an(XY\mathbf{S}) \setminus \mathbf{forb}$ for the causal effect of X on Y given $\mathbf{S}$ we call a collider path node set π^i_W for $W \in Y\mathbf{M}$* valid *wrt. to $(X, Y, \mathbf{S})$ if for each path node $C \in \pi^i_W$ both of the following conditions are fulfilled:*

$$(1)\ C \notin \mathbf{forb}, \quad \textit{and} \quad (2a)\ C \in \mathbf{vancs}\ \textit{or}\ (2b)\ C \perp\!\!\!\perp X \mid \mathbf{vancs}\,. \tag{9}$$

Condition (1) is required for any valid adjustment set. If jointly (2a) and (2b) are not fulfilled, i.e. $C \notin \mathbf{vancs}$ and $C \not\perp\!\!\!\perp X \mid \mathbf{vancs}$, then the collider path stops before C. In the results developed here (in particular Lemmas B.6,B.7,B.8) it will become clear that this choice leads to an optimal set if graphical optimality holds. Our candidate optimal adjustment set is now constructed based on the parents of $Y\mathbf{M}$, valid collider path nodes of $Y\mathbf{M}$, and their parents to 'close' these collider paths.

Definition 4 (O-set)**.** *Given Assumptions 1, define the set $\mathbf{O}(X, Y, \mathbf{S}) = \mathbf{P} \cup \mathbf{C} \cup \mathbf{P_C}$ where*

$$\mathbf{P} = pa(Y\mathbf{M}) \setminus \mathbf{forb}, \;\; \mathbf{C} = \cup\!\!\!\!\cup_{W \in Y\mathbf{M}} \cup\!\!\!\!\cup_i \{\pi^i_W :\ \pi^i_W \textit{ is valid wrt. to } (X, Y, \mathbf{S})\}, \;\; \mathbf{P_C} = pa(\mathbf{C})$$

In the following we will abbreviate $\mathbf{O} = \mathbf{O}(X, Y, \mathbf{S})$. Algorithm C.1 states efficient pseudo-code to construct the $\mathbf{O}$-set and detect whether a valid adjustment set exists. Since none of the conditions of Def. 3 for adding parents or collider nodes depends on previously added nodes, the algorithm is

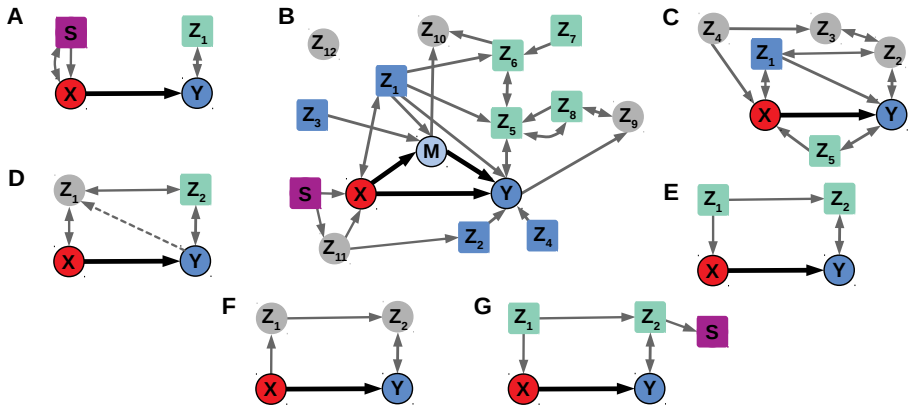

Figure 2: Examples illustrating (optimal) adjustment sets. In all examples the causal effect along causal paths (thick black edges) between X (red circle) and Y (blue circle) potentially through mediators $\mathbf{M}$ (light blue circle), and conditioned on some variables $\mathbf{S}$ (purple box), is considered. The adjustment set $\mathbf{O}$ consists of $\mathbf{P}$ (blue boxes) and $\mathbf{C}$ (green boxes). See main text for details.

order-independent. The statements occurring in lines 11 and 21 that no valid adjustment set exists are proven in Thm. 1. If the graph is a DAG, then lines 5-23 can be omitted. The most time-consuming part is checking for a path in line 13, Def. 3(2b) $C \perp\!\!\!\perp X \mid \mathbf{vancs}$, which can be implemented with (bi-directional) breadth-first search as proposed in van der Zander et al. (2019).

Our numerical experiments in Section 3 will show that further interesting adjustment sets are the *minimized* $\mathbf{O}$-set $\mathbf{O}_{\min}$, where $\mathbf{O}$ is minimized such that no subset can be removed without making $\mathbf{O}_{\min}$ invalid, and the *collider-minimized* $\mathbf{O}$-set $\mathbf{O}_{\mathrm{Cmin}}$ where only $\mathbf{CP_C} \setminus \mathbf{P} \subseteq \mathbf{O}$ is minimized such that no collider-subset can be removed without making $\mathbf{O}_{\mathrm{Cmin}}$ invalid. Both adjustment sets can be constructed with Alg. C.2 similar to the efficient algorithms in van der Zander et al. (2019). Also the minimized sets are order-independent since the nodes are removed only after the for-loops. Based on the idea in $\mathbf{O}_{\mathrm{Cmin}}$, in the numerical experiments we also consider $\mathrm{Adjust}_{\mathrm{Xmin}}$, where only $\mathrm{Adjust} \setminus pa(Y\mathbf{M})$ is minimized and $pa(Y\mathbf{M})$ is always included. Finally, we also evaluate $\mathrm{Adjust}_{\min}$ where Adjust is fully minimized.

Before discussing the optimality of the $\mathbf{O}$-set, we need to assure that it is a valid adjustment set. Similar to the proof given in Perković et al. (2018) for the validity of the $\mathbf{vancs}$-set (for the case without $\mathbf{S}$), we can state that the $\mathbf{O}$-set is valid if and only if a valid adjustment set exists.

Theorem 1 (Validity of O-set)**.** *Given Assumptions 1 but* without *a priori assuming that a valid adjustment set exists (apart from the requirements of amenability and* $\mathbf{S} \cap \mathbf{forb} = \emptyset$*). If and only if a valid adjustment set exists, then* $\mathbf{O}$ *is a valid adjustment set.*

2.3 Graphical optimality

We now move to the question of optimality. It is known that there are graphs where no graphical criterion exists to determine optimality. Examples, discussed later, are the graphs in Figs. 2E,F. Before stating necessary and sufficient conditions for graphical optimality, we mention that next to the $\mathbf{O}$-set defined above and the Adjust set $\mathbf{vancs}$ (Perković et al., 2018), we are not aware of any other systematically constructed set that will yield a valid adjustment set for the case with hidden variables (MAGs or ADMGs). van der Zander et al. (2019) provide algorithms to list all valid adjustment sets, but the question is which of these a user should choose. In principle, Lemma 2 can be used to cross-compare all pairs of sets, but this is not really feasible. Hence, for automated causal effect estimation, rather than the question of whether graphical optimality holds, it is crucial to have a set with better properties than other systematically constructable sets. In the following we state that the adjustment informations follow $J_{\mathbf{O}} \geq J_{\mathbf{vancs}}$ for *any* graph (whether graphical optimality holds or not).

Theorem 2 (O-set vs Adjust-set)**.** *Given Assumptions 1 with* $\mathbf{O}$ *defined in Def. 4 and the Adjust-set defined in Eq.* (2)*, it holds that* $J_{\mathbf{O}} \geq J_{\mathbf{vancs}}$ *for any graph* $\mathcal{G}$ *with* $J_{\mathbf{O}} = J_{\mathbf{vancs}}$ *only if (1)* $\mathbf{O} = \mathbf{vancs}$*, or (2)* $\mathbf{O} \subseteq \mathbf{vancs}$ *and* $X \perp\!\!\!\perp \mathbf{vancs} \setminus \mathbf{O} \mid \mathbf{OS}$.

In the following examples we illustrate the $\mathbf{O}$-set construction and explore conditions for graphical optimality. SSR20 provide a sufficient condition for optimality, which states that either all nodes are observed (no bi-directed edges exist) or for all observed nodes $\mathbf{V} \subset \mathbf{vancs}$. This is a very strict assumption and not fulfilled for any of the example graphs (except for Example G) discussed in the following in Fig. 2.

Example B. Figure 2B depicts a larger example to illustrate the $\mathbf{O}$-set $\mathbf{O} = \mathbf{PCP_C}$ with $\mathbf{P} = Z_1Z_2Z_3Z_4$ (blue boxes) and $\mathbf{CP_C} = Z_5Z_6Z_7Z_8$ (green boxes). We also have a conditioned variable $\mathbf{S}$. Among $\mathbf{P}$, only Z_1Z_2 are needed to block non-causal paths to X, Z_3Z_4 are only there to constrain information in Y. Here the same holds for the whole set $\mathbf{CP_C}$ which was constructed from the paths $Y \leftrightarrow Z_5 \leftrightarrow Z_6 \leftarrow Z_7$ and $Y \leftrightarrow Z_5 \leftrightarrow Z_8$ which does not include Z_9 since it is a descendant of $Y\mathbf{M}$. Here the collider-minimized set is $\mathbf{O}_{\mathrm{Cmin}} = Z_1Z_2Z_3Z_4$. Including an independent variable like Z_{12} in $\mathbf{O}$ would not decrease the adjustment information $J_{\mathbf{O}}$, but then $\mathbf{O}$ would not be of minimum cardinality anymore (proven in Cor. B.1). Here, again, the condition of SSR20 does not hold (e.g., Z_5 is not an ancestor of $XY\mathbf{S}$). $\mathbf{O}$ is optimal here which can be seen as follows: For term (iii) in the inequalities (8) to even be non-zero, we would need a valid $\mathbf{Z}$ such that $\mathbf{Z} \setminus \mathbf{O}$ has a path to Y given $\mathbf{OS}X$. But these are all blocked. Note that while Z_{10} or $Z_9 \in \mathbf{Z}$ would open a path to Y, both of these are descendants of $\mathbf{M}$ or Y and, hence, cannot be in a valid $\mathbf{Z}$. For term (iv) to even be non-zero $\mathbf{O} \setminus \mathbf{Z}$ would need to have a path to X given $\mathbf{ZS}$. But since a valid $\mathbf{Z}$ has to contain Z_1 and Z_2 (or Z_{11}), which is part of $\mathbf{O}_{\mathrm{Cmin}}$, all those paths are blocked. Hence, $\mathbf{O}$ is optimal here.

Example C. In Fig. 2C a case is shown where $\mathbf{O} = \mathbf{O}_{\mathrm{Cmin}} = Z_1Z_5$. Z_2 is not part of $\mathbf{O}$ because none of the conditions in Def. 3(2) is fulfilled: $Z_2 \notin \mathbf{vancs} = Z_1Z_4Z_5$ and $Z_2 \not\perp\!\!\!\perp X \mid \mathbf{vancs}$. We call Z_2 an N-node. But Z_2 cannot be part of any valid $\mathbf{Z}$ because it has a collider path to X through Z_1 which is always open because it is part of $\mathbf{vancs}$. Hence, term (iii) is always zero. Term (iv) is zero because $\mathbf{O} \setminus \mathbf{Z}$ is empty for any valid $\mathbf{Z}$ here. Here even $J_{\mathbf{O}} > J_{\mathbf{Z}}$ since $\mathbf{O}$ is minimal and term (ii) $I_{X;\mathbf{Z}\setminus\mathbf{O}|\mathbf{O}} > 0$ for any $\mathbf{Z} \neq \mathbf{O}$ (generally proven in Corollary B.1).

Example D. The example in Fig. 2D depicts a case with $\mathbf{O} = Z_2$ where Z_1 is an N-node (without the dashed link). Next to $\mathbf{Z} = \emptyset$ another valid set is $\mathbf{Z} = Z_1$. Then term (iii) is non-zero, but always $\leq$ than term (i) because the dependence between Z_1 and Y given X is always smaller than the dependence between Z_2 and Y given X. In the same way term (iv) is non-zero but always smaller than term (ii). Hence, $\mathbf{O}$ is optimal here. If the dashed link exists, then the only other valid set is $\mathbf{Z} = \emptyset$ and both terms are strictly zero. Note, however, that then, if the graph is interpreted as a MAG, it is not amenable.

Example E. The example in Fig. 2E (Fig. 3 in SSR20 and also discussed in HPM19) is not graphically optimal. Here $\mathbf{O} = Z_1Z_2$. Other valid adjustment sets are Z_1 or the empty set. From using $Z_1 \perp\!\!\!\perp Y|X$ and $X \perp\!\!\!\perp Z_2|Z_1$ in the inequalities (8) one can derive in information-theoretic terms that both Z_1Z_2 and $\emptyset$ are better than $\mathbf{vancs} = Z_1$, but since $J_{Z_1Z_2} = J_{\emptyset} + I_{Z_2;Y|XZ_1} - I_{X;Z_1}$, a superior adjustment set depends on how strong the link $Z_1 \rightarrow X$ vs. $Z_2 \leftrightarrow Y$ is. The graph stays non-optimal also with a link $Z_1 \leftrightarrow Z_2$.

Example F. The example in Fig. 2F is also not graphically optimal. Here $\mathbf{O} = \emptyset$ and Z_2 is an N-node with a non-collider path to X. Other valid adjustment sets are Z_1 and Z_1Z_2. Higher adjustment information here depends on the distribution. Also the same graph with the link $Z_1 \leftrightarrow X$ is non-optimal. If, however, there is another link $Z_1 \rightarrow Y$, then $\mathbf{O} = \emptyset$ is optimal (then Z_1 is a mediator).

Example G. The example in Fig. 2G is only a slight modification of Example E with an added selected condition $\mathbf{S}$. Then $Z_1, Z_2 \in \mathbf{vancs}$. We still get $\mathbf{O} = Z_1Z_2$ and this is now optimal since any valid set has to contain Z_1 and $X \perp\!\!\!\perp Z_2 \mid Z_1$.

The main result of this work is a set of necessary and sufficient conditions for the existence of graphical optimality and the proof of optimality of the $\mathbf{O}$-set which is based on the intuition gained in the preceding examples. A relevant Lemma for this theorem states that $\mathbf{O}_{\mathrm{Cmin}}$ is a subset of $\mathbf{vancs}$ (Lemma B.1), similar to corresponding Lemmas in van der Zander et al. (2019).

Theorem 3 (Necessary and sufficient graphical conditions for optimality and optimality of O-set)**.** *Given Assumptions 1 and with* $\mathbf{O} = \mathbf{PCP_C}$ *defined in Def. 4 and* $\mathbf{O}_{\mathrm{Cmin}}$ *constructed by Alg. C.2. Denote* $\mathbf{C}_u = \mathbf{O} \setminus \mathbf{O}_{\mathrm{Cmin}}$. *If and only if exactly one valid adjustment set exist, or both of the following conditions are fulfilled, then graphical optimality holds and* $\mathbf{O}$ *is optimal:*

(I.1) There are no spouses $N \in sp(Y\mathbf{MC}) \setminus (\mathbf{forbOS})$ *or for* all N *it holds that (I.2)* $X \ast\!\!\rightarrow N$ *or* X *has a collider path (except for the first link)* $X \ast\!\!\rightarrow K \leftrightarrow \cdots \leftrightarrow N$ *where all colliders* $K \in \mathbf{vancs}$*, and*

(II) Either (II.1) there are no valid collider path nodes ($\mathbf{CP_C} \setminus \mathbf{P} = \emptyset$), or (II.2) $\mathbf{O} = \mathbf{O}_{\mathrm{Cmin}}$ and, hence, $\mathbf{C}_u = \emptyset$, or (II.3) $X \perp\!\!\!\perp \mathbf{C}_u \mid \mathbf{O}_{\mathrm{Cmin}}\mathbf{S}$.

Applied to the examples, we obtain that in Example A Cond. (I) holds since no N-node exists and Cond. (II.3) holds since $X \perp\!\!\!\perp Z_1 \mid S$. In Example B also no N-node exists and Cond. (II.3) holds with $\mathbf{O}_{\mathrm{Cmin}} = Z_1 Z_2 Z_3 Z_4$. In example C Z_2 is an N-node, but there is a collider path to X through Z_1 which is in $\mathbf{vancs}$. In example D (without the dashed link) Z_1 is an N-node, but it has a bidirected link with X and Cond. (II.3) holds with $\mathbf{O}_{\mathrm{Cmin}} = \emptyset$. In Example E optimality does not hold, but Cond. (I) actually holds since there is no N-node. Cond. (II) is not fulfilled since $\mathbf{O} \neq \mathbf{O}_{\mathrm{Cmin}}$ and there is a link to X. Example F has an N-node Z_2 with no collider path to X implying that Cond. (I) does not hold, while Cond. (II) is actually fulfilled with $\mathbf{O} = \emptyset = \mathbf{O}_{\mathrm{Cmin}}$. Example G is optimal since there are no N-nodes and $X \perp\!\!\!\perp Z_2 \mid \mathbf{O}_{\mathrm{Cmin}}\mathbf{S} = Z_1 S$.

Similar to SSR20, HPM19, and Witte et al. (2020), we also provide results regarding minimality and minimum cardinality for the hidden variables case in Corollary B.1.

3 Numerical experiments

We now investigate graphical optimality empirically to answer three questions: Firstly, whether for a linear estimator under Assumptions 2 the asymptotically optimal variance also translates into better *finite-sample* variance. Secondly, how the $\mathbf{O}$-set performs in non-optimal settings (according to Thm. 3). Thirdly, whether the results also hold for nonlinear dependencies not captured by the class of estimators for which our theoretical results were derived (under Assumptions 2 leading to Lemma 1). To this end, we compare the performance of $\mathbf{O}$, Adjust, $\mathbf{O}_{\mathrm{Cmin}}$, $\mathbf{O}_{\mathrm{min}}$, $\mathrm{Adjust}_{\mathrm{Xmin}}$, and $\mathrm{Adjust}_{\mathrm{min}}$ (see definitions in Section 2.2) on two experimental setups: (1) linear models using linear least squares estimation (LinReg) and in the Supplement (2) nonlinear models using nearest neighbor (kNN) and gaussian process (GP) estimation. The setup is based on a generalized additive model and described in detail in Section D. The configurations cover $N \in \{10, 20, 30\}$ nodes and $L \in \{2N, 3N, 5N\}$ links (in total 9,000 models). For each model we randomly pick an observed pair (X, Y) connected by a causal path, set $\mathbf{S} = \emptyset$, and consider the *soft* intervention $do(X = x + 1)$ relative to the unperturbed data (x) as ground truth, which corresponds to the linear regression coefficient in the linear case. We further assert that the following criteria hold: (1) the effect is identifiable, (2) the minimal adjustment cardinality is $|\mathbf{vancs}_{\mathrm{min}}(X, Y)| > 0$, and (3) the (absolute) causal effect is $\geq 10^{-3}$ to make sure that Faithfulness holds (if these criteria cannot be fulfilled, another model is generated). Surprisingly, among these 9,000 randomly created configurations more than 80% fulfill the optimality conditions in Thm. 3. This may indicate that also in many real-world scenarios graphical optimality actually holds.

The results in Fig. S9 confirm our first hypothesis that for linear experiments with an estimator fulfilling Assumptions 2 and in settings where graphical optimality is fulfilled (Thm. 3) the $\mathbf{O}$-set either has similar RMSE or significantly outperforms all other tested variants. In particular, $\mathbf{O}_{\mathrm{min}}$ and $\mathrm{Adjust}_{\mathrm{min}}$ are bad choices for this setting. Adjust is intermediate and $\mathbf{O}_{\mathrm{Cmin}}$ and $\mathrm{Adjust}_{\mathrm{Xmin}}$ come closest to $\mathbf{O}$, but may still yield significantly higher variance. $\mathbf{O}_{\mathrm{Cmin}}$ seems to always outperform $\mathrm{Adjust}_{\mathrm{Xmin}}$, albeit in the experiments they yield similar adjustment sets in most cases. Interestingly, $\mathrm{Adjust}_{\mathrm{Xmin}}$ is almost always better than Adjust and $\mathrm{Adjust}_{\mathrm{min}}$ (here not further analyzed theoretically).

Secondly, in non-optimal settings the $\mathbf{O}$-set still outperforms Adjust (as expected by Thm. 2), but compared to all other variants there is no clear winner although $\mathbf{O}_{\mathrm{min}}$ and $\mathrm{Adjust}_{\mathrm{min}}$ are still bad choices (where still $\mathbf{O}_{\mathrm{min}}$ always outperforms $\mathrm{Adjust}_{\mathrm{min}}$). $\mathbf{O}$ vs $\mathbf{O}_{\mathrm{Cmin}}$ vs $\mathrm{Adjust}_{\mathrm{Xmin}}$ is rather undecided, but as in the optimal case $\mathbf{O}_{\mathrm{Cmin}}$ seems to always outperform $\mathrm{Adjust}_{\mathrm{Xmin}}$ in the few cases where their adjustment sets are not identical. Cardinality is slightly higher for $\mathbf{O}$ (average cardinality ≈ 7) compared to Adjust (average cardinality ≈ 6). As expected $\mathbf{O}_{\mathrm{Cmin}}$ (average cardinality ≈ 4, almost same as $\mathrm{Adjust}_{\mathrm{Xmin}}$) has slightly lower cardinality and the fully minimal sets are almost all at 1 (the minimum criterion in our configurations). As expected, for very small sample sizes $n = 30$ (see Appendix D.3) that become comparable to the adjustment set cardinality, there tends to be a trade-off and smaller cardinality helps. Then $\mathbf{O}_{\mathrm{Cmin}}$ and $\mathrm{Adjust}_{\mathrm{Xmin}}$ (with $\mathbf{O}_{\mathrm{Cmin}}$ still outperforming $\mathrm{Adjust}_{\mathrm{Xmin}}$) tend to be better than $\mathbf{O}$, but here this effect is only present for $n = 30$ and for $n = 50$ already negligible compared to the gain in $J_{\mathbf{O}}$. In general, results are very similar for other sample sizes (see Appendix D.3).

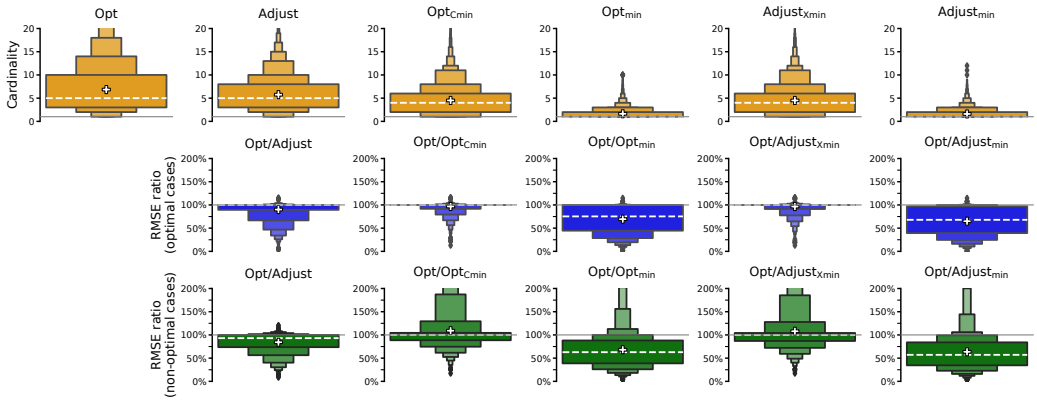

Figure 3: Results of linear experiments with LinReg and sample size $n = 100$. Shown are letter-value plots (Hofmann et al., 2017) of adjustment set cardinalities (top row), as well as RMSE ratios for the $\mathbf{O}$-set vs. other approaches for optimal configurations (middle row) and non-optimal configurations (bottom row). RMSE was estimated from 100 realizations. The dashed horizontal line denotes the median of the RMSE ratios, and the white 'plus' their average. The letter-value plots are interpreted as follows: The largest box shows the 25%–75% range. The next smaller box above (below) shows the 75%–87.5% (12.5%–25%) range and so forth.

Thirdly, for the nonlinear setup with the kNN and GP-estimators (all results shown in Section D) we clearly find that the RMSE-entropy-relation (7) for $f(\cdot) = \frac{1}{\sqrt{n}} e^{H_{Y|X\mathbf{Z}S} - H_{X|\mathbf{Z}S}}$ (bottom row) does not hold. Consequently, we find no clear results and observe only tendencies of better performance. For example, $\text{Adjust}_{\text{Xmin}}$ seems to often work better than Adjust, $\mathbf{O}_{\text{min}}$ better than $\text{Adjust}_{\text{min}}$, and $\text{Adjust}_{\text{Xmin}}$ better than $\mathbf{O}_{\text{min}}$. Overall there is a slight tendency that $\mathbf{O}_{\text{Cmin}}$ works better in more configurations compared to all others. Results are similar for 'optimal' and 'non-optimal' configurations.

4 Discussion and Conclusions

The proposed adjustment information formalizes the common intuition to choose adjustment sets that maximally constrain the effect variable and minimally constrain the cause variable. For a class of causal effect estimators the asymptotic variance is directly related to the adjustment information. Basic properties of information theory yield inequalities that were used to investigate graphical optimality. The main **theoretical contributions** are a necessary and sufficient graphical criterion for the existence of an optimal adjustment set and a definition and algorithm to construct it. Further, the optimal set is valid if and only if a valid adjustment set exists and has smaller (or equal) asymptotic variance compared to the Adjust-set proposed in Perković et al. (2018) for any graph, whether graphical optimality holds or not. Our **practical contributions** comprise Python code to construct adjustment sets and check optimality, as well as extensive numerical experiments that demonstrate that the theoretical results also hold for relatively small sample sizes.

The **results are currently limited** to estimators for which the asymptotic variance can be expressed as in relation (7) (Assumptions 2). This result holds for least-squares estimators, but it is unclear whether this also holds for more general classes. We conjecture this is the case for asymptotically linear estimators considered in SSR20 and Rotnitzky and Smucler (2019) since there it has been shown that the asymptotic distribution depends only on $\mathbf{Z}$, and potentially it holds even more generally.

The proposed information-theoretic approach can guide **further research**, for example to address other types of graphs as emerge from the output of causal discovery algorithms and the setting where the graph is unknown (Witte et al., 2020; Maathuis et al., 2009, 2010). At present, the approach only applies to causal graphical models described by ADMGs and MAGs without selection variables. We note that selection-variables linked to mediators or the effect immediately render the causal effect non-identifiable. Last, it remains an open problem to identify optimal adjustment estimands for other criteria such as the front-door formula and Pearl's general do-calculus (Pearl, 2009).

References

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, 2006.

Robin J Evans and Thomas S Richardson. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, pages 1452–1482, 2014.

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

Heike Hofmann, Hadley Wickham, and Karen Kafadar. Letter-value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, 26(3):469–477, 2017.

L. F. Kozachenko and Nikolai N. Leonenko. Sample estimate of the entropy of a random vector. *Probl. Peredachi Informatsii*, 23(2):9–16, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):16, 2004. ISSN 1063651X.

Manabu Kuroki and Zhihong Cai. Selection of identifiability criteria for total effects by using path diagrams. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 333–340, 2004.

Manabu Kuroki and Masami Miyakawa. Covariate selection for estimating the causal effect of control plans by using causal diagrams. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):209–222, 2003.

Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–8, 2010.

KV Mardia, JT Kent, and JM Bibby. Multivariate analysis, 1979. *Probability and mathematical statistics. Academic Press Inc*, 1979.

Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statist. Sci.*, 8(3):266–269, 08 1993. doi: 10.1214/ss/1177010894.

Judea Pearl. *Causality: Models, reasoning, and inference.* Cambridge University Press, 2009.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. A complete generalized adjustment criterion. In *Uncertainty in Artificial Intelligence-Proceedings of the Thirty-First Conference (2015)*, pages 682–691. AUAI Press, 2015.

Emilija Perković, Johannes Textor, and Markus Kalisch. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18:1–62, 2018.

CE Rasmussen and CKI Williams. *Gaussian processes for machine learning*. MIT Press, Cambridge, MA, USA, 2006.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *Ann. Statist.*, 30(4):962–1030, 08 2002.

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

Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 527–536, 2010.

Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521*, 2020.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Boston, 2000.

Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270:1–40, 2019.

Janine Witte, Leonard Henckel, Marloes H Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.

Jiji Zhang. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, Citeseer, 2006.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] All theoretical claims are substantiated by corresponding theorems.
   (b) Did you describe the limitations of your work? [Yes] The Discussion section discusses the limitations.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a rather theoretical work with no direct societal implications.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Yes, see Assumptions 1,2.
   (b) Did you include complete proofs of all theoretical results? [Yes] All full proofs are given in the Appendix. There are no 'proof sketches'.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The Supplement contains code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The Supplement details all experimental parameters.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] There are only letter-valued plots and scatter plots without error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] For the considered graphs the computation time to construct adjustment sets is very short and we deem it neglegible to the actual cost of fitting methods that use these adjustment sets.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Scikit-learn was used to fit models.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Own code to construct adjustment sets is provided.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] No data from other sources was used.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No data from other sources was used.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No human experiments.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No human experiments.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No human experiments.