Graphical criteria for efficient total effect estimation via adjustment in causal linear models

Henckel, Leonard; Perkovi, Emilija; Maathuis, Marloes H.

Published in:
Journal of the Royal Statistical Society. Series B: Statistical Methodology

DOI: 10.1111/rssb.12451

Publication date: 2022

Document version
Publisher's PDF, also known as Version of record

Document license: CC BY

Citation for published version (APA):
Graphical criteria for efficient total effect estimation via adjustment in causal linear models

Leonard Henckel1 | Emilija Perković2 | Marloes H. Maathuis3

1University of Copenhagen, Copenhagen, Denmark | 2University of Washington, Seattle, Washington, USA | 3ETH Zurich, Zurich, Switzerland

Abstract
Covariate adjustment is a commonly used method for total causal effect estimation. In recent years, graphical criteria have been developed to identify all valid adjustment sets, that is, all covariate sets that can be used for this purpose. Different valid adjustment sets typically provide total causal effect estimates of varying accuracies. Restricting ourselves to causal linear models, we introduce a graphical criterion to compare the asymptotic variances provided by certain valid adjustment sets. We employ this result to develop two further graphical tools. First, we introduce a simple variance decreasing pruning procedure for any given valid adjustment set. Second, we give a graphical characterization of a valid adjustment set that provides the optimal asymptotic variance among all valid adjustment sets. Our results depend only on the graphical structure and not on the specific error variances or edge coefficients of the underlying causal linear model. They can be applied to directed acyclic graphs (DAGs), completed partially directed acyclic graphs (CPDAGs) and maximally oriented partially directed acyclic graphs (maximal PDAGs). We present simulations and a real data example to support our results and show their practical applicability.

Keywords
causal inference, covariate adjustment, efficiency, graphical models
1 | INTRODUCTION

Covariate adjustment is a popular method for estimating total causal effects from observational data. Given a causal graph, with nodes representing covariates and edges representing direct effects, graphical criteria have been developed to read off covariate sets that can be used for this purpose. We refer to such sets as valid adjustment sets. The best-known such criterion is probably the back-door criterion (Pearl, 1993), which is sufficient for adjustment. A necessary and sufficient criterion was developed by Shpitser et al. (2010) and Perković et al. (2018).

Given the complete identification of all valid adjustment sets, the following question naturally arises: If more than one valid adjustment set is available, which one should be used for estimation? In practice, this choice will often be affected by considerations such as the ease and cost of data collection. On the other hand, statistical aspects should also be taken into account as different valid adjustment sets provide estimates with varying accuracy. In this paper, we give an answer to this question from an asymptotic variance perspective under the assumptions that (i) the underlying causal model is linear and (ii) the total causal effects are estimated via adjusted ordinary least squares regression. Under these two assumptions, we graphically characterize a valid adjustment set and show that the corresponding linearly adjusted estimator is asymptotically optimal, meaning that it provides the smallest attainable asymptotic variance among all estimators adjusting for a valid adjustment set.

Our results are closely related to a growing literature on variable selection for efficient total causal effect estimation using either graphical criteria or conditional independence statements. For example, there are simulation studies (Brookhart et al., 2006; Lefebvre et al., 2008), results regarding minimum asymptotic variance bounds (Hahn, 2004; Rotnitzky & Robins, 1995; Rotnitzky et al., 2010) and theoretical results for certain estimators (Lunceford & Davidian, 2004; Robinson & Jewell, 1991; Schnitzer et al., 2016; Wooldridge, 2016). These results indicate that the following two notions hold: First, adding variables predictive of the treatment to a given valid adjustment set may harm the efficiency. Second, adding variables predictive of the outcome may improve the efficiency. Model selection procedures taking these notions into account have also been developed (Shortreed & Ertefaie, 2017; VanderWeele & Shpitser, 2011).

While the above notions provide useful heuristics, there are pitfalls to the approach of labeling individual covariates as either good or bad for efficiency. Whether adding a given covariate to an adjustment set is harmful or beneficial can vary depending on the starting adjustment set. Furthermore, adding or removing a covariate might render a valid adjustment set invalid. As a result, some care must be taken when sequentially applying these heuristics and doing so does not generally result in a valid adjustment set with an efficiency guarantee. In fact, none of the previous literature has even considered the possibility of a valid adjustment set with an efficiency guarantee, such as being asymptotically optimal.

Kuroki and Miyakawa (2003) and Kuroki and Cai (2004) adopt an approach that circumvents the difficulties of labeling individual covariates. They compare valid adjustment sets in terms of their respective estimator’s asymptotic variance. Both papers consider causal linear models with Gaussian errors and introduce graphical criteria that identify which of two valid adjustment sets provides the estimator with the smaller asymptotic variance, irrespective of the underlying causal linear model’s specific edge coefficients and error variances. However, the criterion from Kuroki and Miyakawa (2003) is restricted to adjustment sets of size two and the criterion from Kuroki and Cai (2004) is restricted to disjoint adjustment sets. This limits their practical applicability.

In this paper, we also adopt the approach of comparing valid adjustment sets. We first provide a new graphical criterion (Theorem 1) that can compare more pairs of valid adjustment sets than...
the criteria by Kuroki and Miyakawa (2003) and Kuroki and Cai (2004). In addition, our criterion can also be applied to cases where the treatment and outcome are multivariate. We note, however, that we still cannot compare all pairs of valid adjustment sets. This is in fact impossible with the graph alone (see Example 3).

We then use our novel graphical criterion to graphically characterize a valid adjustment set and show that it provides the smallest possible asymptotic variance among all valid adjustment sets (Theorem 3). We refer to this property as asymptotic optimality. Considering that not all pairs of valid adjustment sets can be compared with the graph alone, it is remarkable that such a graphically identifiable asymptotically optimal adjustment set exists.

In addition to our main results, we also provide a simple order-invariant pruning procedure that, given a valid adjustment set, returns a subset that is also valid and provides equal or smaller asymptotic variance (Algorithm 1 and Theorem 2). Such pruning has already been proposed by VanderWeele and Shpitser (2011), who conjectured a resulting efficiency gain. Our main contribution is that we formally establish this efficiency gain for causal linear models with Theorem 1 and show the order invariance of such a procedure.

We also provide two additional important extensions. First, we derive all results for graphs that are more general than DAGs and that can be learned from data under weaker assumptions than DAGs. We discuss this more thoroughly in the Preliminaries (Section 2). Second, we show that our results also hold for causal linear models with possibly non-Gaussian errors. This relies on an important and remarkable preparatory result (Proposition 1), where we show that for valid adjustment sets the misspecified ordinary least squares regressions that occur in causal linear models with non-Gaussian errors are nonetheless asymptotically well-behaved (see Example A.7 in the Supplement).

We provide numerical experiments to quantify the efficiency that is gained by using our asymptotically optimal adjustment set in finite samples (Section 4) and apply our methods to single-cell data from Sachs et al. (2005) (Section 5). We outline how our results can be applied to cases with unmeasured covariates in the Discussion. All proofs can be found in the Supplement (Henckel et al., 2020) and we have made our code available at https://github.com/henckell/CodeEfficientVAS.

## 2 | PRELIMINARIES

In this paper, we use graphs where nodes represent random variables, and edges represent conditional dependencies and direct causal effects. We now give an overview of the main graphical objects used in this paper. We give the usual graphical definitions and define these objects more formally in Section A.1 of the Supplement.

We consider three classes of acyclic graphs: directed acyclic graphs (DAGs), completed partially directed acyclic graphs (CPDAGs) and maximally oriented partially directed acyclic graphs (maximal PDAGs) (see Example A.3 in the Supplement). DAGs are directed graphs, that is, graphs with all edges of the form \( \rightarrow \) and without directed cycles. They arise naturally to describe causal relationships under the assumption of no feedback loops (cf. Pearl, 2009). Generally, it is not possible to learn the causal DAG from observational data alone. Under the assumptions of causal sufficiency and faithfulness, one can, however, learn a Markov equivalence class of DAGs, which can be uniquely represented by a CPDAG (Andersson et al., 1997; Chickering, 2002; Colombo & Maathuis, 2014; Meek, 1995; Spirtes et al., 2000). Given explicit knowledge of some causal relationships between variables, access to interventional data, or some model restrictions, one can obtain a refinement of this class, uniquely represented by a maximal PDAG (Eigenmann et al., 2017; Hauser &
Bühlmann, 2012; Hoyer et al., 2008; Meek, 1995; Scheines et al., 1998; Wang et al., 2017). All three graph types encode conditional independence relationships that can be read off the graph by applying the well known d-separation criterion (see definition 1.2.3 in Pearl, 2009 for DAGs, definition 3.5 in Maathuis & Colombo, 2015 for CPDAGs and lemma C.1 of the Supplement for maximal PDAGs). If the d-separation criterion captures all conditional independences between the random variables, we say their joint distribution is faithful to the graph. We use the notation $X \perp_Z Y | \text{uni} \mathcal{G}$ to denote that $Z$ d-separates $X$ from $Y$ in $\mathcal{G}$, with $X$, $Y$ and $Z$ pairwise disjoint nodes sets in a graph $\mathcal{G}$.

Remark 1 DAGs and CPDAGs are special cases of maximal PDAGs. In the remainder of the paper, results are generally stated in terms of maximal PDAGs. Readers unfamiliar with CPDAGs and maximal PDAGs may also disregard this and simply think of all results as being with respect to DAGs.

We now introduce causal linear models, total effects and define some notation.

**Causal DAGs, CPDAGs, maximal PDAGs.** We consider interventions $\text{do}(x)$ (for $X \subseteq \mathcal{V}$), which represent outside interventions that set $X$ to $x$ uniformly for the entire population (Pearl, 1995). A density $f$ of $\mathcal{V} = \{V_1, \ldots, V_p\}$ is compatible with a causal DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if all post-intervention densities $f(v|\text{do}(x))$ factorize as:

$$f(v|\text{do}(x)) = \begin{cases} \prod_{v_i \in \mathcal{V} \setminus X} f(v_i|\text{pa}(V_i, \mathcal{G})), & \text{if } X = x, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Equation (1) is known as the truncated factorization formula (Pearl, 2009), manipulated density formula (Spirtes et al., 2000) or the g-formula (Robins, 1986). A density $f$ of $\mathcal{V} = \{V_1, \ldots, V_p\}$ is compatible with a causal maximal PDAG or a causal CPDAG $\mathcal{G}$ if it is compatible with a causal DAG $D \in [\mathcal{G}]$, where $[\mathcal{G}]$ is the class of DAGs represented by $\mathcal{G}$.

**Causal linear model.** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG. Then $\mathcal{V} = (V_1, \ldots, V_p)^T$, $p \geq 1$ follows a causal linear model compatible with $\mathcal{G}$ if the following two conditions hold:

1. The distribution $f$ of $\mathcal{V}$ is compatible with the causal DAG $\mathcal{G}$.
2. $V_1, \ldots, V_p$ follows a set of linear equations

$$V_i = \sum_{v_j \in \text{pa}(V_i, \mathcal{G})} \alpha_{ij} V_j + \epsilon_{v_i}, \tag{2}$$

where $i, j \in \{1, \ldots, p\}$, $i \neq j$, $\alpha_{ij} \in \mathbb{R}$ and $\epsilon_{v_1}, \ldots, \epsilon_{v_p}$ are jointly independent random variables with mean 0 and finite variance.

$\mathcal{V}$ follows a causal linear model compatible with a maximal PDAG or CPDAG $\mathcal{G}$ if it follows a causal linear model compatible with a DAG $D \in [\mathcal{G}]$.

We refer to $\epsilon_{v_1}, \ldots, \epsilon_{v_p}$ as errors and emphasize that we do not require them to be Gaussian. Furthermore, by construction $E[\mathcal{V}] = 0$. The coefficient $\alpha_{ij}$ corresponding to the edge $V_j \rightarrow V_i$ in the causal DAG $\mathcal{G}$ can be interpreted as the direct effect of $V_j$ on $V_i$ with respect to $\mathcal{V}$.

**Example 1** Consider the DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in Figure 1. Then the generating mechanism
with \( \epsilon_{v_1} \sim \text{Uniform}(-1, 1), \epsilon_{v_2} \sim \text{Uniform}(-2, 2), \epsilon_{v_3} \sim \mathcal{N}(0, 1), \epsilon_{v_4} \sim \mathcal{N}(0, 2), \epsilon_{v_5} \sim \text{Uniform}(-3, 3) \) and \( \epsilon_{v_6} \sim \mathcal{N}(0, 3) \) for \( \mathbf{V} \) is an example of a causal linear model compatible with \( \mathcal{G} \). We use the notation \( \leftarrow \) to emphasize that we are considering a generating mechanism and not just an equation.

A do-intervention, for example \( \text{do}(V_4 = 1) \), then corresponds to replacing the generating mechanism of the intervened on variables with the fixed intervention value, for example, \( V_4 \leftarrow 1 \).

**Causal and proper paths.** Let \( \mathcal{G} \) be a maximal PDAG. A path \((V_1, ..., V_m)\) in \( \mathcal{G} \) is called a causal path from \( V_1 \) to \( V_m \) if \( V_i \rightarrow V_{i+1} \) for every \( i \in \{1, ..., m-1\} \). Let \( \mathbf{X} \) and \( \mathbf{Y} \) be disjoint node sets in a causal maximal PDAG \( \mathcal{G} \). A path from \( \mathbf{X} \) to \( \mathbf{Y} \) is proper if only the first node on \( p \) is in \( \mathbf{X} \).

**Total effects.** (Pearl, 2009) Let \( \mathbf{V} = (\mathbf{X}^T, \mathbf{Y}^T, \mathbf{Z}^T)^T \) be a random vector, where \( \mathbf{X} = (X_1, ..., X_{k_x})^T \) and \( \mathbf{Y} = (Y_1, ..., Y_{k_y})^T \). Then the total effect of \( \mathbf{X} \) on \( \mathbf{Y} \) is defined as the matrix \( \tau_{\mathbf{Y} \mathbf{X}} \), where

\[
(\tau_{\mathbf{Y} \mathbf{X}})_{j,i} = \frac{\partial}{\partial x_i} E[Y_j | \text{do}(x_1, ..., x_{k_x})], \quad i \in \{1, ..., k_x\}, \quad j \in \{1, ..., k_y\}
\]

represents the effect of \( X_i \) on \( Y_j \) in the joint intervention of \( \mathbf{X} \) on \( \mathbf{Y} \). In general, \( \tau_{\mathbf{Y} \mathbf{X}} \) is a matrix of functions but in causal linear models, the partial derivatives do not depend on \( x_i \). Hence, \( \tau_{\mathbf{Y} \mathbf{X}} \) reduces to a matrix of numbers, whose values are determined by the coefficients in Equation (2) (Nandy et al., 2017; Wright, 1934). We can thus give an equivalent definition of the total effect specific to this setting. Consider disjoint node sets \( \mathbf{X} = \{X_1, ..., X_{k_x}\} \) and \( \mathbf{Y} = \{Y_1, ..., Y_{k_y}\} \) in a causal DAG \( \mathcal{G} = (\mathbf{V}, \mathbf{E}) \), such that \( \mathbf{V} \) follows a causal linear model compatible with \( \mathcal{G} \). The total effect along a causal path \( p \) from \( \mathbf{X} \) to \( \mathbf{Y} \) in \( \mathcal{G} \) is the product of the edge coefficients along \( p \). The total effect of \( \mathbf{X} \) on \( \mathbf{Y} \) is then the matrix \( \tau_{\mathbf{Y} \mathbf{X}} \in \mathbb{R}^{k_y \times k_x} \) whose \((j, i)\)th value \( (\tau_{\mathbf{Y} \mathbf{X}})_{j,i} \) is equal to the sum of the total effects along all proper causal paths from \( \mathbf{X} \) to \( Y_j \) starting with \( X_i \) in \( \mathcal{G} \).

If \( \mathbf{V} \) follows a causal linear model compatible with a causal CPDAG or maximal PDAG \( \mathcal{G} \), the total effect of \( \mathbf{X} \) on \( \mathbf{Y} \) is identifiable if it is the same for every DAG in \( \mathcal{G} \).

**FIGURE 1** Directed acyclic graph from Examples 1, 2 and 6
Remark 2 Consider the total effect \( \tau_{yx} \) of \( X \) on \( Y \). If for some \( Y_j \in Y \) and \( X_i \in X \), \( Y_j \) is a non-descendant of \( X_i \) then \( (\tau_{yx})_{i,j} = 0 \). Further, the total effect \( \tau_{yx} \) of \( X_i \) on \( Y_j \) will generally differ from the partial total effect \( (\tau_{yx})_{i,j} \) in the joint intervention on \( X \). This is due to the latter effect not considering causal paths from \( X_i \) to \( Y_j \) that contain other nodes in \( X \setminus \{X_i\} \). The total effect of \( X \) on any \( Y_j \), however, does not depend on the remaining \( Y \setminus \{Y_j\} \).

Notation for covariance matrices and regression coefficients. Consider random vectors \( S = (S_1, \ldots, S_k)^T \), \( T = (T_1, \ldots, T_k)^T \) and \( W = (W_1^T, \ldots, W_m^T)^T \), with \( W_1, W_2, \ldots, W_m \) themselves random vectors. We denote the covariance matrix of \( S \) with \( \Sigma_{ss} \in \mathbb{R}^{k \times k} \) and the covariance matrix between \( S \) and \( T \) with \( \Sigma_{st} \in \mathbb{R}^{k_1 \times k_2} \), where its \( (i,j) \)th element equals \( \text{Cov}(S_i, T_j) \). We further define \( \Sigma_{ss, t} = \Sigma_{ss} - \Sigma_{st} \Sigma_{tt}^{-1} \Sigma_{ts} \). If \( k_s = 1 \), we write \( \sigma_{ss, t} \) instead. Let \( \beta_{st, w} \in \mathbb{R}^{k_s \times k_t} \) represent the ordinary least squares regression coefficient matrix whose \((i,j)\)th element is the regression coefficient of \( T_j \) in the regression of \( S_i \) on \( T \) and \( W \), with \( \hat{\beta}_{st, w} \) denoting the corresponding estimator. We also use the notation that \( \beta_{st, w1^t w2^t \ldots w_m^t} = \beta_{st, w} \) and \( \Sigma_{st, w1^t w2^t \ldots w_m^t} = \Sigma_{st, w} \). Given a set \( X = \{X_1, \ldots, X_k \} \) we use the notation \( X_{-i} \) to denote \( X \setminus \{X_i\} \).

3 \quad MAIN RESULTS

3.1 \quad Total effect estimation via covariate adjustment

In causal linear models, total effects can be estimated via ordinary least squares regression given an appropriate adjustment set. This result is well known in the Gaussian case with Shpitser et al. (2010) and Perković et al. (2018) having fully characterized the class of valid adjustment sets (see Definition A.4).

The fact that total effects can be estimated via ordinary least squares regression has been shown to generalize to causal linear models with arbitrary error distributions for a singleton \( X \) with the adjustment set \( \text{pa}(X, G) \) (Proposition 3.1 from the supplement of Nandy et al., 2017). We now extend this property to arbitrary valid adjustment sets and derive the estimator’s asymptotic distribution.

Proposition 1 Let \( X = \{X_1, \ldots, X_k \} \) and \( Y = \{Y_1, \ldots, Y_k \} \) be disjoint node sets in a causal DAG \( G = (V, E) \) and let \( V \) follow a causal linear model compatible with \( G \). Let \( Z \) be a valid adjustment set relative to \( (X, Y) \) in \( G \). Then

\[
\sqrt{n}((\hat{\beta}_{yxz})_{j,i} - (\tau_{yx})_{j,i}) \xrightarrow{d} \mathcal{N} \left( 0, \begin{pmatrix} \sigma_{jyj, xz} \\ \sigma_{xix, xz} \end{pmatrix} \right),
\]

for all \( i = 1, \ldots, k_x \) and \( j = 1, \ldots, k_y \) with \( \xrightarrow{d} \) denoting convergence in distribution.

The key aspect to Proposition 1, is that it does not require the considered regression of \( Y \) on \( X \) and \( Z \) to be well-specified, in the sense of being linear and having homoskedastic residuals. One may think this generality is not needed given that we consider causal linear models. However, in a causal linear model with non-Gaussian errors, adjusted regressions other than that of a node on its parents, are not generally well-specified (see Example A.7). We note that for Proposition 1 to hold for misspecified regressions, it is essential that \( Z \) is a valid adjustment set. Of course, Proposition 1 corresponds to what we know for well-specified regressions, in which case the
restriction to valid adjustment sets is not needed. In causal linear models with Gaussian errors, all regressions are well-specified.

Due to the result in Proposition 1, we use the notation $\hat{\tau}_{yx}^Z$ to denote the least squares estimate $\hat{\beta}_{yxz}$ of $\tau_{yx}$, for any valid adjustment set $Z$ relative to $(X, Y)$. We also write

$$a. \text{var}((\hat{\tau}_{yx}^Z)_{j,i}) = a. \text{var}((\hat{\beta}_{yxz})_{j,i}) = \frac{\sigma_{yjyjxz}}{\sigma_{xixi,xixi,z}}$$

(3)

and use $a. \text{var}(\hat{\tau}_{yx}^Z)$ to denote the matrix with entries

$$a. \text{var}(\hat{\tau}_{yx}^Z)_{j,i} = a. \text{var}((\hat{\tau}_{yx}^Z)_{j,i}), \quad i = 1, \ldots, k_x \text{ and } j = 1, \ldots, k_y.$$

Remark 3 The terms in Equation (3) depend on the distribution of $V = \{V_1, \ldots, V_p\}$ only through the covariance matrix $\Sigma_{vv}$, which in turn only depends on the underlying causal linear model through the edge coefficients $a_{ij}$ and error variances $\text{var}(\epsilon_i), \quad i, j \in \{1, \ldots, p\}, \quad i \neq j$ (cf. Nandy et al., 2017). In particular, this implies that the asymptotic variance $a. \text{var}(\hat{\tau}_{yx}^Z)$ does not depend on the error distribution families.

Example 2 Consider the causal DAG $G = (V, E)$ in Figure 1 and assume that $V$ follows the causal linear model from Example 1. The total effect of $V_4$ on $V_6$ in $G$ is $\tau_{64} = \alpha_{64} + \alpha_{65}\alpha_{54}$. By Proposition 1, $\tau_{64}$ also equals the population level regression coefficient of $V_4$ in the regression of $V_6$ on $V_4$ and any adjustment set of the form $A \cup B$, with $A \subseteq \{V_1, V_2\}$ non-empty and $B \subseteq \{V_3\}$ possibly empty (see Definition A.4).

3.2 Comparing valid adjustment sets

We now introduce a new graphical criterion for qualitative comparisons between the asymptotic variances resulting from certain pairs of valid adjustment sets, which is more general than the criteria of Kuroki and Miyakawa (2003) and Kuroki and Cai (2004).

Theorem 1 Let $X$ and $Y$ be disjoint node sets in a maximal PDAG $G = (V, E)$, such that $V$ follows a causal linear model that is compatible with $G$. Let $Z_1$ and $Z_2$ be two valid adjustment sets relative to $(X, Y)$ in $G$ and let $T = Z_1 \backslash Z_2$ and $S = Z_2 \backslash Z_1$. If $Y \perp_G T|X \cup Z_2$ and $X \perp_G S|Z_1$, then

$$a. \text{var}(\hat{\tau}_{yx}^Z) \leq a. \text{var}(\hat{\tau}_{yx}^Z),$$

with the matrix inequality denoting entry-wise inequality.

The proof of Theorem 1 relies on Equation (3). The intuition behind Theorem 1 is that a conditioning set $B$ with more information on a target variable $A$ leads to a smaller conditional variance $\sigma_{aa,b}$. Thus, the conditional independence statements in Theorem 1 imply that

$$\sigma_{xixi,xixi} \leq \sigma_{xixi,xixi} \quad \text{and} \quad \sigma_{yjyj,yjz2} \leq \sigma_{yjyj,yjz1},$$

for all $X_i \in X$ and $Y_j \in Y$. 

---

HENCKEL ET AL. 585

---
We stress that when a causal linear model with non-Gaussian errors is considered, Theorem 1 holds only for pairs of valid adjustment sets. This is due to Proposition 1, which only holds for misspecified regressions when a valid adjustment set is considered.

Example 3  Consider the DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure 2a and assume that $\mathbf{V}$ follows a causal linear model compatible with $\mathcal{G}$. Any valid adjustment set $\mathbf{Z}$ relative to $(X, Y)$ in $\mathcal{G}$ is of the form $\{B\} \cup \mathbf{S}$, where $\mathbf{S} \subseteq \{A, C, D\}$ (see Definition A.4). Fix any such set $\mathbf{Z}$. One can check that (i) $Y \perp_{\mathcal{G}} A/X \cup \mathbf{Z} - \{A\}$, (ii) $Y \perp_{\mathcal{G}} D/X \cup \mathbf{Z} - \{D\}$ and (iii) $X \perp_{\mathcal{G}} C/\mathbf{Z} - \{C\}$. We can thus apply Theorem 1 to the following pairs of valid adjustment sets:

1. $\mathbf{Z}_1 = \mathbf{Z}$ and $\mathbf{Z}_2 = \mathbf{Z} - \{A\}$,
2. $\mathbf{Z}_1 = \mathbf{Z}$ and $\mathbf{Z}_2 = \mathbf{Z} - \{D\}$ and
3. $\mathbf{Z}_1 = \mathbf{Z} - \{C\}$ and $\mathbf{Z}_2 = \mathbf{Z}$.

We can conclude that adding $A$ or $D$ to any conditioning set worsens the asymptotic variance, while the converse holds for $C$. Consequently, $\{B, C\}$ provides the best asymptotic variance, while the set $\text{pa}(X, \mathcal{G}) = \{A, B\}$ does not fare well.

In order to empirically verify these results, we randomly drew six causal linear models compatible with $\mathcal{G}$ and computed the asymptotic variances $a\var(\hat{\theta}_{xy})$ for each valid adjustment set $\mathbf{Z}$ relative to $(X, Y)$ in $\mathcal{G}$, across these six models. Specifically, we did the following for each model. We drew error variances $\sigma_{vv}$ for each node $V \in \mathbf{V}$ independently from a standard uniform distribution and edge coefficients $\alpha_{vw}$ for each edge $(W, V) \in \mathbf{E}$ independently from a standard normal distribution. From these parameters we computed the causal linear model’s covariance matrix and then, in accordance with Proposition 1, the asymptotic variances corresponding to each valid adjustment set. We did not consider error properties other than the variance (and mean 0) as they are irrelevant for the asymptotic variances (see Remark 3).

The asymptotic variances are given in Table 1. They show that the three proven trends do in fact hold and that $\{B, C\}$ provides the best asymptotic variance in the considered models. Interestingly, the order of the asymptotic variances corresponding to any two sets that cannot be compared using Theorem 1, such as $\{A, B, C\}$ and $\{B\}$, or $\{A, B\}$ and $\{B, D\}$, are in fact inconsistent throughout the considered models.

We now give a simple corollary of Theorem 1. It shows that superfluous parents of $X$ are harmful to the asymptotic variance, while parents of $Y$ are beneficial.

\begin{figure}[h]
\centering
\begin{subfigure}{0.4	extwidth}
\centering
\begin{tikzpicture}
\node (A) at (0,0) {$A$};
\node (B) at (1,0) {$B$};
\node (C) at (2,0) {$C$};
\node (D) at (0,-1) {$D$};
\node (X) at (1,-1) {$X$};
\node (Y) at (2,-1) {$Y$};
\draw (A) -- (B);
\draw (B) -- (C);
\draw (D) -- (X);
\draw (X) -- (Y);
\end{tikzpicture}
\caption{(a) Causal directed acyclic graph (DAG) from Examples 3, 4 and 6 and (b) causal DAG from Examples 5 and 6}
\end{subfigure}
\end{figure}
Corollary 1  Let $X$ and $Y$ be disjoint node sets in a maximal PDAG $\mathcal{G} = (V, E)$ and let $V$ follow a causal linear model compatible with $\mathcal{G}$. Let $Z$ be a valid adjustment set relative to $(X, Y)$ in $\mathcal{G}$. Then the following two statements hold:

1. Let $P \in \text{pa}(X, \mathcal{G})$. If $Z' = Z \setminus \{P\}$ is a valid adjustment set relative to $(X, Y)$ in $\mathcal{G}$, then
   \[ a.\text{var}(\hat{\tau}_{yx}^{Z'}) \leq a.\text{var}(\hat{\tau}_{yx}^Z). \]

2. Let $R \in \text{pa}(Y, \mathcal{G})$. If $Z' = Z \cup \{R\}$ is a valid adjustment set relative to $(X, Y)$ in $\mathcal{G}$, then
   \[ a.\text{var}(\hat{\tau}_{yx}^{Z'}) \leq a.\text{var}(\hat{\tau}_{yx}^Z). \]

We now give a second corollary of Theorem 1 that is especially relevant for randomized trials. It shows that if $\text{pa}(X, \mathcal{G}) = \emptyset$, adding covariates not in $\text{de}(X, \mathcal{G})$ to an adjustment set can only be beneficial for the asymptotic variance. In particular, this result applies to pre-treatment covariates in a randomized trial.

Corollary 2  Let $X$ and $Y$ be disjoint node sets in a DAG $\mathcal{G} = (V, E)$, such that $\text{pa}(X, \mathcal{G}) = \emptyset$ and let $V$ follow a causal linear model compatible with $\mathcal{G}$. Let $Z$ and $Z'$ be two node sets in $\mathcal{G}$, such that $Z \cap (\text{de}(X, \mathcal{G}) \cup Y) = \emptyset$ and $Z' \cap (\text{de}(X, \mathcal{G}) \cup Y) = \emptyset$. Then $Z$ and $Z'$ are valid adjustment sets relative to $(X, Y)$ in $\mathcal{G}$ and if $Z \subseteq Z'$,

\[ a.\text{var}(\hat{\tau}_{yx}^{Z'}) \leq a.\text{var}(\hat{\tau}_{yx}^Z). \]

### 3.3 Pruning procedure

The result from Theorem 1 can be used to prune a valid adjustment set to obtain a subset that is still valid and yields a smaller asymptotic variance. Generally, which of the subsets $\tilde{Z} \subseteq Z$ provides the optimal asymptotic variance depends on the edge coefficients in the underlying acyclic graph in Figure 2a.

#### TABLE 1  Asymptotic variances for six randomly drawn causal linear models compatible with the directed acyclic graph in Figure 2a

<table>
<thead>
<tr>
<th>Adjustment set</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
<th>Case 5</th>
<th>Case 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[A, B]$</td>
<td>5.38</td>
<td>5.47</td>
<td>0.85</td>
<td>0.57</td>
<td>5.42</td>
<td>0.64</td>
</tr>
<tr>
<td>$[A, B, C]$</td>
<td>1.44</td>
<td>4.44</td>
<td>0.51</td>
<td>0.39</td>
<td>2.02</td>
<td>0.61</td>
</tr>
<tr>
<td>$[B]$</td>
<td>3.49</td>
<td>4.40</td>
<td>0.54</td>
<td>0.26</td>
<td>2.76</td>
<td>0.39</td>
</tr>
<tr>
<td>$[B, C]$</td>
<td>0.94</td>
<td>3.58</td>
<td>0.32</td>
<td>0.18</td>
<td>1.03</td>
<td>0.37</td>
</tr>
<tr>
<td>$[A, B, D]$</td>
<td>7.20</td>
<td>7.39</td>
<td>12.65</td>
<td>0.65</td>
<td>5.72</td>
<td>0.69</td>
</tr>
<tr>
<td>$[A, B, C, D]$</td>
<td>1.93</td>
<td>6.01</td>
<td>7.59</td>
<td>0.45</td>
<td>2.13</td>
<td>0.65</td>
</tr>
<tr>
<td>$[B, D]$</td>
<td>5.31</td>
<td>6.33</td>
<td>12.34</td>
<td>0.35</td>
<td>3.05</td>
<td>0.44</td>
</tr>
<tr>
<td>$[B, C, D]$</td>
<td>1.42</td>
<td>5.15</td>
<td>7.41</td>
<td>0.24</td>
<td>1.14</td>
<td>0.41</td>
</tr>
</tbody>
</table>
causal linear model (see Example 5). However, we can use Theorem 1 to identify a subset such that there is no other subset for which Theorem 1 guarantees a better asymptotic variance. This is formalized in Algorithm 1 and Theorem 2. In practice, such pruning is advisable as it reduces the number of variables that need to be measured while also improving precision.

Algorithm 1: Pruning procedure

\[\text{input} : \text{Causal maximal PDAG } G \text{ and disjoint node sets } X, Y \text{ and } Z \text{ in } G, \text{ such that } Z \text{ is a valid adjustment set relative to } (X, Y) \text{ in } G\]

\[\text{output}: \text{Valid adjustment set } Z' \subseteq Z \text{ relative to } (X, Y) \text{ in } G, \text{ such that } a.var(\hat{r}_{yx}^Z) \leq a.var(\hat{r}_{yx}^{Z'}) \]

1 begin
2 \[Z' = Z;\]
3 foreach \(Z \in Z\) do
4 \[\text{if } Y \perp_G Z | X \cup (Z' \setminus \{Z\}) \text{ then}\]
5 \[Z' = Z' \setminus \{Z\};\]
6 return \(Z'\);

Theorem 2  Let \(X\) and \(Y\) be disjoint node sets in a maximal PDAG \(G = (V, E)\) and let \(V\) follow a causal linear model compatible with \(G\). Let \(Z\) be a valid adjustment set relative to \((X, Y)\) in \(G\). Applying Algorithm 1 then yields a valid adjustment set \(Z' \subseteq Z\), such that \(a.var(\hat{r}_{yx}^Z) \leq a.var(\hat{r}_{yx}^{Z'})\) and there is no other subset of \(Z\) for which Theorem 1 guarantees a better asymptotic variance than \(Z'\). Further, Algorithm 1 outputs the same set \(Z'\), regardless of the order in which the nodes in \(Z\) are considered.

The idea of iteratively discarding covariates not associated with the outcome has previously been considered by VanderWeele and Shpitser (2011). They showed that such pruning preserves the adjustment sets validity and conjectured that it is good for efficiency. We expand on this result in two ways. First, we establish that such pruning is order invariant, which allows us to formalize it into the order-agnostic Algorithm 1. Second, we formally establish the efficiency gain for causal linear models. For completeness, we also provide a graphical proof for the fact that our pruning procedure preserves validity (Lemma D.1 in the Supplement).

Example 4  We now return to Example 3 and the DAG \(G = (V, E)\) in Figure 2a to illustrate Algorithm 1. Fix some valid adjustment \(Z\) relative to \((X, Y)\) in \(G\). As (i) \(Y \perp_G A | X \cup Z \setminus \{A\}\), (ii) \(Y \perp_G D | X \cup Z \setminus \{D\}\) and (iii) \(X \perp_G C | Z \setminus \{C\}\), Algorithm 1 will discard the nodes \(A\) and \(D\), while keeping the nodes \(B\) and \(C\) whenever these nodes are in \(Z\). This is done independently of the order in which the nodes are considered. Hence, \(Z\) will either be pruned to \([B]\) or \([B, C]\). Both these sets are valid adjustment sets relative to \((X, Y)\) in \(G\) and \([B, C]\) yields the optimal asymptotic variance among all valid adjustment sets, while \([B]\) yields the optimal asymptotic variance among all valid adjustment sets that do not contain \(C\).

Example 5  We now give an example in which one cannot use Theorem 1 to decide which subset \(Z' \subseteq Z\) of a valid adjustment set \(Z\) provides the optimal asymptotic variance. Instead,
the optimal subset depends on the edge coefficients and error variances of the underlying causal linear model.

Consider the DAG $\mathcal{G}$ in Figure 2b and two sets of possible edge coefficients for $\mathcal{G}$. Let all edge coefficients that are not explicitly mentioned be 1 and let $\alpha_{ba} = 0.5$, $\alpha_{xa} = 0.25$ and $\alpha_{yx} = 2$ in Case (i), while $\alpha_{xa} = 0.7$ and $\alpha_{yc} = 0.5$ in Case (ii). With all error variances equal to 1 in both cases, one obtains the asymptotic variances shown in Table 2, where we ignore error properties other than variance (and mean 0) in accordance with Remark 3.

The set $\{C\}$ provides the smallest asymptotic variance in both cases and will also be the output of Algorithm 1 applied to any valid adjustment set containing $C$. If we instead consider valid adjustment sets that do not contain $C$ the situation is more complex. If, for example, we apply Algorithm 1 to $\{A, B, D\}$, the output is $\{A, B\}$, which is the subset that yields the optimal asymptotic variance in Case (i), but is outperformed by the empty set in Case (ii). These two sets cannot be compared with Theorem 1. However, Theorem 1 still implies that the valid adjustment sets $\{A\}$, $\{D\}$, $\{A, D\}$ and $\{A, B, D\}$ provide worse asymptotic variances than both the empty set and $\{A, B\}$. Algorithm 1 will prune these sets to either $\{A, B\}$ or the empty set, depending on whether they originally included $\{A, B\}$.

### 3.4 The optimal adjustment set

We will now define a set that provides the optimal asymptotic variance among all valid adjustment sets. This is remarkable since Theorem 1 can only compare the asymptotic variance provided by certain valid adjustment sets (see Examples 3 and 5). Nevertheless, it allows us to define this optimal set, whose optimality only depends on the underlying causal graph. We first give some preparatory definitions which for simplicity we restrict to the DAG setting. The general definitions for maximal PDAGs are given in Section A.1. This includes the definition of the set of possible descendants $\text{possde}(X, \mathcal{G})$, which in the case that $\mathcal{G}$ is a DAG, reduces to the set of descendants $\text{de}(X, \mathcal{G})$.

**Causal and forbidden nodes.** Consider a DAG $\mathcal{G}$ and two disjoint node sets $X, Y$ in $\mathcal{G}$. We define causal nodes relative to $(X, Y)$ in $\mathcal{G}$, denoted $\text{cn}(X, Y, \mathcal{G})$, as all nodes on proper causal paths from $X$ to $Y$, excluding nodes in $X$. For singleton $X$, causal nodes are also called mediating nodes. We then define the forbidden nodes relative to $(X, Y)$ in $\mathcal{G}$ as

$$\text{forb}(X, Y, \mathcal{G}) = \text{de}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) \cup X.$$ 

Note that, differently from Perković et al. (2018), we also include $X$ in $\text{forb}(X, Y, \mathcal{G})$ to simplify notation. The forbidden set characterizes those covariates that may never be included in a valid adjustment set (see Definition A.4).

<table>
<thead>
<tr>
<th>Adjustment set</th>
<th>${C}$</th>
<th>${B, C}$</th>
<th>${A, C}$</th>
<th>$\emptyset$</th>
<th>${A, B}$</th>
<th>${A}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case (i)</td>
<td>0.48</td>
<td>0.49</td>
<td>0.5</td>
<td>0.97</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>Case (ii)</td>
<td>0.4</td>
<td>0.45</td>
<td>0.5</td>
<td>0.5</td>
<td>0.56</td>
<td>0.62</td>
</tr>
</tbody>
</table>

**TABLE 2** Non-exhaustive table of the asymptotic variances corresponding to valid adjustment sets in Example 5
Definition 1 Let $X$ and $Y$ be disjoint node sets in a maximal PDAG $G$. We define $O(X, Y, G)$ as:

$$O(X, Y, G) = \text{pa}(\text{cn}(X, Y, G), G) \setminus \text{forb}(X, Y, G).$$

Theorem 3 Let $X$ and $Y$ be disjoint node sets in a causal maximal PDAG $G = (V, E)$, such that $Y \subseteq \text{possde}(X, G)$. Let the density $f$ of $V$ be compatible with $G$ and let $O = O(X, Y, G)$. Then the following three statements hold:

(i) The set $O$ is a valid adjustment set relative to $(X, Y)$ in $G$ if and only if there exists a valid adjustment set relative to $(X, Y)$ in $G$.

(ii) Let $Z$ be a valid adjustment set relative to $(X, Y)$ in $G$. If $V$ follows a causal linear model compatible with $G$ then

$$a. \text{var}(\hat{\tau}_{yx}^O) \leq a. \text{var}(\hat{\tau}_{yx}^Z).$$

(iii) Let $Z$ be a valid adjustment set relative to $(X, Y)$ in $G$, such that

$$a. \text{var}(\hat{\tau}_{yx}^O) = a. \text{var}(\hat{\tau}_{yx}^Z).$$

If $V$ follows a causal linear model compatible with $G$ and $f$ is faithful to $G$ then $O \subseteq Z$.

Remark 4 In Theorem 3 we assume that $Y \subseteq \text{possde}(X, G)$. If $Y \not\subseteq \text{possde}(X, G)$ we can instead consider the total effect of $X$ on $Y = Y \cap \text{possde}(X, G)$ since the total effect of $X$ on $Y \setminus Y$ is 0 (see Remark 2). Hence, this restriction simply avoids superfluous estimation of some zero values.

Statement (i) implies that our optimal set, similarly to the $\text{Adjust}(X, Y, G)$ set from Definition 12 in Perković et al. (2018), can be used to check if there exists a valid adjustment set, albeit with the added qualifier that $Y$ has to be appropriately pruned in advance (see Remark 4). Due to statement (ii) in Theorem 3, we call $O(X, Y, G)$ asymptotically optimal. Statement (iii) implies that in faithful models no other asymptotically optimal set is of smaller or equal size than $O(X, Y, G)$.

As a corollary to Theorem 3 jointly with Theorem 2, the output of Algorithm 1 is $O(X, Y, G)$, whenever the starting valid adjustment set $Z$ is a superset of $O(X, Y, G)$. It is of course simpler to compute $O$ directly rather than via pruning.

Remarkably, given a maximal PDAG $G$ amenable relative to some tuple of node sets $(X, Y)$ such that $Y \subseteq \text{possde}(X, G)$, $O(X, Y, G)$ is not only the optimal set amongst all valid adjustment sets in $G$ but also among all valid adjustment sets in any DAG $D \in [G]$.

In fact, $\text{forb}(X, Y, D') = \text{forb}(X, Y, G)$ and $O(X, Y, D') = O(X, Y, G)$ for all DAGs $D' \in [G]$ (see Lemmas E.7 and E.8).

Intuitively, $O(X, Y, G)$ is constructed to maximize information on $Y$, while minimizing information on $X$ and preserving validity. Although one may think that a simpler set, such as $\text{pa}(Y, G) \setminus (X \cup Y)$ would suffice for this purpose, this is not generally the case. We illustrate this in Example 6. Interestingly, Witte et al. (2020) have shown that $O(X, Y, G)$ can indeed be characterized as $\text{pa}(Y, G) \setminus (X \cup Y)$ in a specific latent projection graph $G$ of $G$.

Example 6 Consider the DAG in Figure 1 and the two DAGs from Figure 2, denoted, respectively, as $G_1$, $G_{2,a}$ and $G_{2,b}$. We now illustrate how to construct $O(X, Y, G)$ and the results from Theorem 3.
Single interventions

FIGURE 3  Violin plots of the ratios of the mean squared errors provided by $O(X, Y, G)$ and the three alternative adjustment sets (i)–(iii) from Section 4, respectively. The true causal directed acyclic graph cases are on the left and the graph estimate cases on the right. The single intervention cases ($|X| = 1$) are at the top and the joint intervention cases ($|X| > 1$) are at the bottom. The red squares show the geometric average of the ratios, the black squares the median. While the summary measures are computed with all ratios, ratios larger than 1.5 are not shown in the plots. From left to right the percentage of ratios larger than 1.5 in the single intervention setting is 0.2, 0.0, 0.0, 7.1, 1.1 and 1.5 percent, respectively. For the joint interventions setting the percentages are 0.2, 0.0, 0.0, 8.8, 2.2 and 1.8 percent, respectively [Colour figure can be viewed at wileyonlinelibrary.com]
Consider first $G_1$ and suppose we are interested in the total effect of $V_4$ on $V_6$ as in Example 2. Here, $\text{pa}(\text{cn}(V_4, V_6, G_1), G_1) = \{V_2, V_3, V_4, V_5\}$ and $\text{forb}(V_4, V_6, G_1) = \{V_4, V_5, V_6\}$. Therefore,

$$O(V_4, V_6, G) = \{V_2, V_3\}.$$ 

As shown in Example 2, $\{V_2, V_3\}$ is a valid adjustment set relative to $(V_4, V_6)$ in $G_1$. It can also easily be verified with Theorem 1 that $\{V_2, V_3\}$ provides a smaller asymptotic variance than any of the alternative valid adjustment sets, as $V_1 \perp_{G_1} V_6 | \{V_2, V_3, V_4\}$ and $V_3 \perp_{G_1} V_4 | \{V_2\}$. Hence, it is asymptotically optimal as claimed in Theorem 3.

Consider now $G_{2,a}$. Here, $\text{pa}(\text{cn}(X, Y, G_{2,a}), G_{2,a}) = \{X, B, C\}$ and $\text{forb}(X, Y, G_{2,a}) = \{X, Y\}$. Therefore,

$$O(X, Y, G_{2,a}) = \{B, C\}.$$ 

Consider now $G_{2,b}$. Here, $\text{pa}(\text{cn}(X, Y, G_{2,b}), G_{2,b}) = \{X, C\}$ and $\text{forb}(X, Y, G_{2,b}) = \{X, Y\}$. Therefore,

$$O(X, Y, G_{2,b}) = \{C\}.$$ 

In these two cases, the results from Theorem 3 are corroborated by Examples 3 and 4, respectively.

We now discuss why $O(X, Y, G)$ takes its distinctive form by considering these three examples. By the result from Theorem 1, an asymptotically optimal valid adjustment set with respect to $(X, Y)$ in $G$ must contain less or equal information on $X$ and more or equal information on $Y$ than any other valid adjustment set.

One might intuitively expect that $\text{pa}(Y, G) \setminus (X \cup Y)$ satisfies these properties. This is indeed the case for two of the three examples considered here, with

$$O(X, Y, G_{2,a}) = \{B, C\} = \text{pa}(Y, G_{2,a}) \setminus \{X, Y\}$$

and

$$O(X, Y, G_{2,b}) = \{C\} = \text{pa}(Y, G_{2,b}) \setminus \{X, Y\}.$$ 

This pattern, however, fails to hold for $G_1$. Here, $\text{pa}(V_6, G_1) \setminus \{V_4, V_6\} = \{V_5\}$ is not a valid adjustment set relative to $(V_4, V_6)$ in $G_1$, as the mediator $V_5 \notin \text{forb}(V_4, V_6, G_1)$.

The construction of $O(V_4, V_6, G_1)$ solves this problem by using the next-closest non-forbidden nodes instead, that is, the non-forbidden parents $\{V_2, V_3\}$ of the causal node $V_5$. This ensures validity while maximizing information on $V_6$ and not providing unnecessary information on $V_4$. Specifically, $V_2$, as the non-forbidden node closest to $V_5$ and furthest from $V_4$ on the non-causal path $(V_4, V_1, V_2, V_5, V_6)$, is the most efficient choice to block this path. Moreover, $V_3$, although superfluous for validity, contains only information on $V_6$ and therefore improves precision. Interestingly, it does so even though $V_3 \notin \text{pa}(V_6, G_1)$.

4 | SIMULATION STUDY

We investigate the finite sample performance of adjusting for $O(X, Y, G)$ by sampling data from randomly generated causal linear models and comparing the empirical mean squared error
provided by $O(X, Y, G)$ to three alternative adjustment sets. We do so both under the assumption that (i) the true causal graph is known and (ii) the true causal graph has to be estimated. A detailed explanation of our simulation setup is given in Section F.1 of the Supplement (Henckel et al., 2020).

We randomly generate a total of 10,000 DAGs, with the number of nodes chosen from \{10, 20, 50, 100\} and the expected neighborhood size from \{2, 3, 4, 5\}. Each graph is associated with a causal linear model. The edge coefficients of the model are drawn independently from a uniform distribution on \([-2, -0.1] \cup [0.1, 2]\), and the errors are either drawn from a Gaussian distribution, a $t$-distribution with 5 degrees of freedom, a logistic distribution or a uniform distribution, with variances in the range of \([0.5, 1.5]\).

For each DAG $D$, we randomly draw $(X, Y)$ such that $|X| \in \{1, 2, 3\}$ and $Y \in \cap_{X_i \in X} \text{de}(X_i, D)$, where we use $|X|$ to denote the size of $X$. We do this for two reasons. First, for a multivariate $Y$ we could simply consider each $Y \in Y$ separately (see Remark 2). Second, if $Y \notin \text{de}(X_i, D)$ for some $X_i \in X$ then the corresponding entry of the total effect $(r_{XY})_i = 0$ (see Remark 2). We then verify whether there exists a valid adjustment set with respect to $(X, Y)$ in both the DAG $D$ and its CPDAG $C$. If not, we resample $X$ and $Y$.

For each causal linear model, we generate 100 data sets with sample sizes $n \in \{125, 500, 2000, 10,000\}$. We simultaneously consider both the case that (i) we know the true causal DAG $D$ and (ii) we have to estimate a causal graph $\hat{G}$ from the data. If the errors are drawn from a Gaussian distribution $\hat{G}$ is estimated with the Greedy Equivalence Search (GES) algorithm (Chickering, 2002), otherwise with the Linear Non-Gaussian Acyclic Models (LiNGAM) algorithm (Shimizu et al., 2006). In both cases, we use the algorithms as implemented in the `pcalg` R-package (Kalisch et al., 2012).

We then compute total effect estimates by adjusting for $O(X, Y, G)$ and three alternative adjustment sets. This is done with respect to both the true causal DAG $D$ and the estimated causal graph $\hat{G}$, with two special cases for the estimates with respect to $\hat{G}$. First, no estimate is returned if there was no valid adjustment set relative to $(X, Y)$ in $\hat{G}$, that is, these cases are discarded for the mean squared error computation. In such cases, we recommend the use of alternative total effect estimators such as the IDA algorithm by Maathuis et al. (2009) and the jointIDA algorithm by Nandy et al. (2017). Second, $0$ is returned as the estimate whenever $Y \notin \text{possde}(X, \hat{G})$, since the total effect on a non-descendant is $0$. The pair $(X, Y)$ is sampled, ensuring that these two special cases do not occur in either the true DAG or its corresponding CPDAG.

The three alternative adjustments sets are:

(i) The empty set, representing a non-causal baseline. It is generally not a valid adjustment set and is denoted by “em”.

(ii) The set $\text{pa}(X, G) \setminus \text{forb}(X, Y, G)$, which in the setting $|X| = 1$ is the valid adjustment set $\text{pa}(X, G)$ popular with practitioners (e.g., Gascon et al., 2015; Sunyer et al., 2015; Williamson et al., 2014). If $|X| > 1$, it is not generally a valid adjustment set. It is denoted by “pa”.

(iii) The valid adjustment set $\text{Adjust}(X, Y, G)$ from Perković et al., (2018). It is denoted by “adj”.

For each causal linear model, we thus have four adjustment sets in two graphical settings. In each of these cases, we compute the empirical mean squared error of our respective estimates with respect to the true total effect. We do so to circumvent any possible issues with post-selection inference and emphasize that we do not consider the estimated standard errors or residuals from the regression analyses. To quantify the advantage of $O(X, Y, G)$, we compute the ratio of the
mean squared error corresponding to $O(X, Y, G)$ and each of the three alternative adjustment sets. This is done separately for the two graphical settings.

Figure 3 is a violin plot of these ratios. We plot the single intervention ($|X| = 1$) and joint intervention ($|X| > 1$) cases separately, as $pa(X, G) \setminus forb(X, Y, G)$ is only guaranteed to be a valid adjustment set if $|X| = 1$. We see that $O(X, Y, G)$ provides consistently smaller mean squared errors than any of the considered alternatives. Except for $\text{Adjust}(X, Y, G)$, all alternative sets are clearly outperformed by $O(X, Y, G)$, with geometric averages below 0.5. As might be expected, the gain becomes smaller when the underlying causal DAG has to be estimated, but it remains respectable. Notably, the proportion of ratios larger than 1.5 is small, even negligible when the true DAG is known. For a more thorough discussion of how the ratios behave depending on the parameters, we refer to Section F.2 of the Supplement. The bulges at 1 are due to two reasons. First, cases in which the compared sets are similar or the same. Second, cases in which $Y \notin \text{possde}(X, \hat{G})$ occurs for a considerable number of the estimated graphs $\hat{G}$ (see Section F.3 of the Supplement).

The only true contender to $O(X, Y, G)$ in terms of performance is $\text{Adjust}(X, Y, G)$. It should be noted, however, that $\text{Adjust}(X, Y, G)$ is a superset of $O(X, Y, G)$ and hence will be more cumbersome to measure in practice (see also Figure 11 of the Supplement).

Another point worth noting is the bad performance of $pa(X, G) \setminus forb(X, Y, G)$. Even though this set is a valid adjustment set if $|X| = 1$, it only provides a small gain compared to the empty set, especially when the graphical structure has to be estimated. This aptly illustrates the importance of taking efficiency considerations into account when choosing a valid adjustment set.

In summary, these results indicate that there are clear benefits to using $O(X, Y, G)$. These benefits decrease when the underlying causal structure is not known in advance, but do remain respectable.

5 | REAL DATA EXAMPLE

Our result can easily be integrated into existing approaches to covariate adjustment. Using $O(X, Y, G)$ to estimate the total effect $\tau_{yx}$ instead of, for example, $pa(X, G)$ only requires the minimal additional effort of computing $O(X, Y, G)$ from the causal graph $G$. And yet, replacing $pa(X, G)$ with $O(X, Y, G)$ can only improve the asymptotic variance when the true causal graph $G$ is used. Of course, errors in the used graph $G$ and finite sample considerations might lead to cases where the use of $O(X, Y, G)$ actually leads to a loss of efficiency in practice, but our simulations (Section 4) indicate that this risk is manageable and that an overall efficiency gain, at essentially no cost, is the norm.

To investigate this further, we apply our results to the single-cell data of Sachs et al. (2005). This data set consists of flow cytometry measurements of 11 phosphorylated proteins and phospholipids in human T-cells, collected under 14 different experimental conditions. Each experimental condition corresponds to a different intervention on the abundance or activity of the proteins. We chose this data set due to the availability of a consensus graph (see figure 5a in Mooij & Heskes, 2013) and the large sample size.

Given that there is some uncertainty regarding the consensus graph, we apply our results using the following three different graphs: the consensus graph, the DAG estimated by Sachs et al. (2005) and the DAG estimated under the restriction to at most 17 edges by Mooij and Heskes (2013). These three DAGs are given in figure 5 of Mooij and Heskes (2013) and we denote them by $G_C$, $G_S$ and $G_M$, respectively.
Our data analysis is as follows. We first log transform the data as it is heavily right skewed. We then consider each of the three candidate DAGs in turn. Restricting ourselves to the 8 experimental conditions for which Mooij and Heskes (2013) provide a graphical interpretation of the condition’s effect, we adjust our starting DAG accordingly. For each such adjusted DAG $G$, we then compute all pairs $(X, Y)$ of nodes in $G$, such that $Y \in \text{de}(X, G)$, to ensure that there is a non-trivial total effect to estimate, and $O(X, Y, G) \neq \text{pa}(X, G)$, to ensure that we compare different estimators. For each such pair, we compute the least squares regressions of $Y$ on $X$ and $O$ as well as of $Y$ on $X$ and $P$. We note that $\hat{\beta}_{yx,o}$ and $\hat{\beta}_{yx,p}$ are estimators for the total effect of $X$ on $Y$ in the considered data regime; not necessarily in the observational regime.

As the true total effects are unknown, we compare the least squares coefficient variance estimates $\hat{\text{var}}(\hat{\beta}_{yx,o})$ and $\hat{\text{var}}(\hat{\beta}_{yx,p})$ by considering their ratio $\hat{\text{var}}(\hat{\beta}_{yx,o}) / \hat{\text{var}}(\hat{\beta}_{yx,p})$. Note that, differently from Section 4, where we are able to compute the empirical mean squared error with respect to the known true total effect, this approach raises some concerns regarding post-selection inference for the two estimated graphs $G_S$ and $G_M$, which we disregard here.

Figure 4 shows violin plots of these ratios, aggregated over all considered $(X, Y)$ pairs in the eight experimental settings, with one plot for each of the graphs $G_C$, $G_S$ and $G_M$. The plots in Figure 4 show that using $O(X, Y, G)$ instead of $\text{pa}(X, G)$ results in geometric means smaller than 1 for all three graphs. In particular, only few of the ratios are larger than 1, with only one larger than 1.2, showing that this gain is obtained at little risk of a potential downside. The gains are

![Figure 4 Violin plots of the ratios $\hat{\text{var}}(\hat{\beta}_{yx,o}) / \hat{\text{var}}(\hat{\beta}_{yx,p})$, with $O = O(X, Y, G)$ and $P = \text{pa}(X, G)$, for all pairs of nodes $(X, Y)$ such that $Y \in \text{de}(X, G)$ and $O(X, Y, G) \neq \text{pa}(X, G)$ in eight experimental conditions from Sachs et al. (2005), obtained under the assumption that $G_C$, $G_S$ or $G_M$, respectively, is the true underlying causal graph. Here, $G_C$ denotes the consensus graph, $G_S$ the acyclic graph estimated by Sachs et al. (2005) and $G_M$ the acyclic graph estimated by Mooij and Heskes (2013). The red squares show the geometric average of the ratios, the black squares the median. [Colour figure can be viewed at wileyonlinelibrary.com]
rather modest, however. This is likely due to the small size and sparsity of the considered graphs. In such settings, even when $O(X, Y, G) \neq \text{pa}(X, G)$, the two sets will often share nodes and only differ in minor ways. As a result, they provide similar asymptotic variances. In fact, this behavior can also be seen in our simulations where small graph sizes and small expected neighborhood sizes result in mean squared error ratios closer to 1 (see Figure 10 of the Supplement).

It is also interesting to consider how the gain in efficiency differs between the three graphs. It is smallest for $G_C$ and largest for $G_S$. As discussed by Mooij and Heskes (2013), several strong faithfulness violations seem to be present in the considered data set. While Theorem 3 does not require faithfulness to hold, faithfulness violations can lead to cases where $O(X, Y, G)$ and $\text{pa}(X, G)$ differ and yet provide the same asymptotic variance, making them both asymptotically optimal. For an example consider the DAG $G$ in Figure 5 and the corresponding causal linear model with the edge weights indicated on the edges (and arbitrary error variances). Here, $O(X, Y, G) = \{P, O\}$ and $\text{pa}(X, G) = \{P\}$ but due to the non-faithfulness of the causal linear model, $O \perp Y | \{X, P\}$, even though $O \in \text{pa}(Y, G)$. Since $O \perp_C X | P$, we can thus conclude with Lemma C.2 that for the considered causal linear model

$$a. \text{var}(\hat{\alpha}_{xy}) = a. \text{var}(\hat{\beta}_{yx}).$$

It appears that this issue is most prevalent for $G_C$, as it is the densest of the three considered graphs and faithfulness violations require multiple connecting paths between nodes. In $G_S$, the least dense of the three considered graphs, this issue appears to be less prominent, leading to a larger gain.

6 DISCUSSION

In this paper, we derived three graphical tools for efficient total effect estimation via adjustment in causal linear models: (i) a graphical criterion to qualitatively compare the asymptotic variance that certain pairs of valid adjustment sets provide, (ii) a pruning procedure to decrease the variance and (iii) a graphical characterization of an asymptotically optimal valid adjustment set. These easy-to-use tools can guide researchers in selecting adjustment sets that estimate the total effect of interest both consistently and efficiently.

One limitation of our results is that the theoretical efficiency guarantees crucially rely on the assumption that we have information of the underlying causal graph in form of a DAG, CPDAG or maximal PDAG. This can be problematic in practice. However, such knowledge is also needed to identify valid adjustment sets in the first place, irrespective of any efficiency considerations. It is also possible, under some assumptions, to learn the underlying DAG, CPDAG or maximal PDAG. Our simulations indicate that in settings where we estimate the graph, applying our results still reduces the mean squared errors of our total effect estimates respectively.

Another limitation is that we only consider total effect estimation via covariate adjustment. As a result, our efficiency guarantees for $O(X, Y, G)$ do not hold with respect to other estimators, such as, for example, ensemble estimators or the front-door criterion (Hayashi & Kuroki, 2014).
Lastly, we would like to discuss possible extensions of our results. First, our results partially extend to cases with latent variables. Theorem 1 extends to settings with latent variables and without selection bias, by simply changing d-separation to (definite) m-separation (Richardson & Spirtes, 2002; Zhang, 2008) in the latent variable graph (MAG or PAG). Theorem 3, however, does not extend to latent variable models. To see this, we can look back at Example 5. If we suppose that C is latent and therefore only consider valid adjustment sets that do not contain C, then the valid adjustment set providing the optimal asymptotic variance depends on the edge coefficients and error variances.

Independent follow-up research has already expanded our work in various other directions. Rotnitzky and Smucler (2020) show that our results on the asymptotic optimality of $\mathbf{O}(X, Y, \mathcal{G})$ extend to a broad class of non-parametric estimators. Building on this, Smucler et al. (2020) consider even more general settings and construct adjustment sets with efficiency guarantees other than asymptotic optimality. Van der Zander and Liskiewicz (2019) provide a polynomial-time algorithm to compute $\mathbf{O}(X, Y, \mathcal{G})$. Witte et al. (2020) provide an alternative characterization of $\mathbf{O}(X, Y, \mathcal{G})$ and also integrate it into the IDA algorithm (Maathuis et al., 2009, 2010; Perković et al., 2017). Kuipers and Moffa (2020) investigate the theoretical finite sample performance of $\mathbf{O}(X, Y, \mathcal{G})$ in a specific non-linear example and discuss how $\mathbf{O}(X, Y, \mathcal{G})$'s performance relates to causal discovery considerations.

Finally, another possible generalization is to consider settings with selection bias. Correa et al. (2018) give a necessary and sufficient graphical criterion for causal effect estimation under confounding and selection bias. It remains to be investigated whether the results presented in this paper can be adapted to this setting.

ACKNOWLEDGEMENTS

Open Access Funding provided by Eidgenossische Technische Hochschule Zurich.

[Correction added on 24 May 2022, after first online publication: CSAL funding statement has been added.]

ORCID

Leonard Henckel http://orcid.org/0000-0002-1000-3622
Emilija Perković http://orcid.org/0000-0003-3198-9213

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.