Leibniz-Institut
für Präventionsforschung und
Epidemiologie – BIPS

Universität Bremen

Causal Model Selection in Epidemiology

Janine Lüschen geb. Witte

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

Universität Bremen
Fachbereich 03: Mathematik und Informatik

Dezember 2021

Erstgutachterin:	Prof. Dr. rer. nat. Vanessa Didelez
Zweitgutachterin:	Prof. Dr. rer. nat. Iris Pigeot-Kübler
Datum des Kolloquiums:	07.12.2021

Meinen Eltern

Danksagung

Zuallererst danke ich Vanessa Didelez für die ebenso fachkundige wie freundschaftliche Betreuung meiner Doktorarbeit. Du hattest immer Zeit für mich, warst in den richtigen Momenten geduldig, fordernd, hilfsbereit, kritisch und begeistert, und du wusstest manchmal besser, woran ich gerade arbeite, als ich selbst. Ich hätte mir keine bessere Doktormutter wünschen können.

Ein großes Dankeschön geht auch an meine Zweitgutachterin Iris Pigeot. Danke, dass du dich mit mir über meine Erfolge gefreut hast und mir Mut zugesprochen hast, wenn es mal nicht so gut lief. Unsere zahlreichen Doktorandenrunden werden mir in guter Erinnerung bleiben.

Ich danke außerdem Marloes Maathuis und Leonard Henckel für die schöne Zeit in Zürich und die produktive Zusammenarbeit am JMLR-Paper.

Für die Zusammenarbeit im DFG-Projekt danke ich vor allem Ronja Foraita, Ryan M. Andrews und Timon Hurink. Danke, dass ich mich immer auf euch verlassen konnte.

Und auch bei allen anderen BIPSler:innen möchte ich mich bedanken, für eure Hilfsbereitschaft, die Flurgespräche und die gemeinsamen Mittagessen. Ebenso bei allen Kolleg:innen außerhalb des BIPS, die mich auf die eine oder andere Weise inspiriert, motiviert oder unterstützt haben. Es war eine gute Zeit mit euch!

Zu guter Letzt danke ich meiner Familie, insbesondere meinen Eltern und meinem Mann, für eure Liebe und Unterstützung. Danke, dass ihr immer für mich da seid.

Contents

ABSTRACT

In this thesis, I investigate two related types of causal model selection: confounder selection and constraint-based causal discovery.

The aim of confounder selection is to determine a valid adjustment set when the interest lies in the causal effect of an exposure or treatment $\mathbf{X}$ on an outcome $\mathbf{Y}$. Ideally, this is based on a causal graph representing relevant domain knowledge. However, as sufficiently detailed knowledge is often not available, alternative strategies are common in practice. These range from simple knowledge-based rules to complex data-driven algorithms, and include e.g. the disjunctive cause criterion, univariate correlation testing and the change-in-estimate method. In this thesis, I investigate popular rules and algorithms from a graphical perspective. I point out implicit structural assumptions, propose a classification scheme and characterise the types of adjustment sets targeted by the different classes of methods. This is supplemented by an extensive simulation study. Main results are that structural assumptions cannot be avoided even if no causal graph is drawn, and that 'outcome-oriented' strategies often lead to more precise estimates than other methods.

The efficiency aspect in then further investigated for the case that the underlying causal graph is known or has been estimated, and the variables jointly follow a multivariate Gaussian distribution. I show that the 'optimal' adjustment set yielding the smallest asymptotic variance can be read off using graphical rules and does not depend on the parameters of the distribution. It has an intuitive interpretation in terms of a graphical projection I propose and can be viewed as the target set of backward regression selection.

Instead of focussing on a single treatment-outcome pair and its confounding factors, the aim of causal discovery is to infer the causal structure among several variables simultaneously. Constraint-based causal discovery algorithms such as the PC-algorithm are based on conditional independence testing and are thus related to backward regression selection and similar methods. I propose a modified version of PC that takes temporal background knowledge into account, and show that the new algorithm is sound and complete and has certain stability properties. Further, I formally investigate two recently suggested methods for handling missing values in causal discovery: test-wise deletion and multiple imputation. In particular, I present necessary and sufficient conditions for the recoverability of causal structures under test-wise deletion and argue that multiple imputation is more challenging in the context of causal discovery than for estimation. Using sim-

ulated and real data, I demonstrate that while both methods outperform list-wise deletion and single imputation, neither is uniformly best. Finally, I discuss chances and challenges of causal discovery and causal graphical modelling in general.

ZUSAMMENFASSUNG

In dieser Arbeit untersuche ich zwei verwandte Arten von kausaler Modellselektion: die Selektion von Confoundern und die constraintbasierte kausale Struktursuche.

Confounderselektion hat zum Ziel, eine gültige Adjustierungsmenge für den kausalen Effekt einer Exposition oder Behandlung $\mathbf{X}$ auf ein Outcome $\mathbf{Y}$ zu bestimmen. Im Idealfall wird dafür relevantes Fachwissen in Form eines kausalen Graphen dargestellt, aus dem dann gültige Adjustierungsmengen abgelesen werden können. Da vorhandenes Wissen jedoch oft nicht detailliert genug ist, wird in der statistischen Praxis häufig auf alternative Strategien zurückgegriffen. Diese reichen von einfachen, auf Hintergrundwissen basierenden Regeln bis hin zu komplexen datengetriebenen Algorithmen und umfassen beispielsweise das „Disjunctive-Cause"-Kriterium, univariate Korrelationstests und die „Change-in-Estimate"-Methode. In der vorliegenden Arbeit betrachte ich gängige Regeln und Algorithmen aus einer graphischen Perspektive. Ich zeige auf, dass viele Methoden implizite strukturelle Annahmen machen, schlage ein Klassifikationsschema vor und beschreibe, welche Arten von Adjustierungsmengen von den Methoden in den verschiedenen Klassen ausgewählt werden. Ergänzt wird dies durch eine umfangreiche Simulationsstudie. Zu den wichtigsten Ergebnissen gehört, dass strukturelle Annahmen auch dann unvermeidbar sind, wenn kein kausaler Graph spezifiziert wird, und dass sogenannte outcomeorientierte Methoden oft zu präziseren Schätzungen führen.

Die Frage nach der statistisch effizienten Adjustierung wird dann weiter untersucht für den Fall, dass der zugrunde liegende kausale Graph bekannt ist oder geschätzt wurde und die Variablen einer gemeinsamen multivariaten Normalverteilung folgen. Ich zeige, dass die „optimale" Adjustierungsmenge, die zu der kleinsten asymptotischen Varianz führt, direkt vom kausalen Graphen abgelesen werden kann und nicht von den Parametern der Verteilung abhängt. Sie lässt sich im Kontext einer von mir vorgeschlagenen graphischen Projektion anschaulich interpretieren und kann als Zielmenge der Rückwärtsselektion von Regressionsmodellen betrachtet werden.

Im Gegensatz zu Analysen, die einzelne Exposition-Outcome-Paare und ihre Störfaktoren betrachten, besteht das Ziel der kausalen Struktursuche darin, die kausalen Beziehungen zwischen mehreren Variablen gleichzeitig zu ermitteln. Constraintbasierte Algorithmen für die kausale Struktursuche wie der PC-Algorithmus basieren auf bedingten Unabhängigkeitstests und sind daher mit der Rückwärts-

selektion von Regressionsmodellen und ähnlichen Verfahren verwandt. Ich schlage eine modifizierte Variante des PC-Algorithmus vor, die zeitliches Hintergrundwissen berücksichtigt, und zeige, dass der neue Algorithmus korrekt und vollständig ist und bestimmte Stabilitätseigenschaften aufweist. Außerdem untersuche ich zwei kürzlich vorgeschlagene Methoden zum Umgang mit fehlenden Werten in der kausalen Struktursuche: das testweise Auslassen von Beobachtungen und die multiple Imputation. Dabei stelle ich notwendige und hinreichende Bedingungen für die Identifizierbarkeit kausaler Strukturen mittels testweiser Auslassung auf und lege dar, warum die multiple Imputation im Kontext der kausalen Struktursuche eine größere Herausforderung darstellt als beim Schätzen. Anhand von simulierten und echten Daten zeige ich, dass beide Methoden bessere Ergebnisse liefern als die analyseweite Auslassung von Beobachtungen und die Einfachimputation, aber keine der beiden der anderen durchweg überlegen ist. Abschließend diskutiere ich die Chancen und Herausforderungen der kausalen Struktursuche und der kausalen graphischen Modellierung im Allgemeinen.

1 Introduction

The aim of causal inference is to discover and quantify non-deterministic cause-effect relationships (Morgan and Winship, 2014; Pearl and Mackenzie, 2018; Cunningham, 2021). This is of central interest in various areas of science including epidemiology, economics, psychology and sociology. Example research questions within the scope of causal inference include:

How would the incidence rate of dementia or stroke change if everyone in the population stopped smoking? (Rojas-Saunero et al., 2021)

Did the job training intervention improve the employment rate, and how much of the effect was mediated by increased job search self-efficacy? (Imai et al., 2010)

What is the causal structure of academic achievement, and where are good points of intervention? (Quintana, 2020)

While these and similar questions could in theory be approached by experimentation, this is often not feasible in practice. Causal inference as a discipline is therefore concerned with the assumptions and methods that allow researchers to draw causal conclusions based on observational data (Hill and Stuart, 2015).

'Regular' statistical inference leverages data to draw conclusions about an underlying probability distribution (Fahrmeir et al., 2003). For example, a typical aim is to estimate the mean of a random variable and to quantify the uncertainty associated with the estimate. Causal inference differs from 'regular' inference in that at least two (joint) distributions are involved: the *observational* distribution underlying the observable data, and a *hypothetical* or *counterfactual* or *interventional* distribution describing the system after an intervention or decision of interest has taken place (Dawid and Didelez, 2010; Peters et al., 2017). A central notion in causal inference is therefore that of *identification*: An aspect of a hypothetical distribution is *identified* from observable data if it can be re-expressed in terms of an observational distribution. Identification is *non-parametric* if it does not rely on assumptions such as linearity, monotonicity or parallel trends.

Two major frameworks for formalising causal inference have evolved over the last decades: the *potential outcomes framework* (Neyman, 1990; Rubin, 1974) and the *causal graph framework* (Spirtes et al., 2000; Pearl, 2009). Both embrace the above idea of mapping hypothetical quantities to observable data, but they do so using different basic concepts and terminology. The potential outcomes framework mainly relies on algebraic equations and is particularly suited to express parametric assumptions about the causal relationships among a small number of variables (Imbens, 2020). Causal graphs represent causal structures among a potentially large number of variables in a non-parametric fashion and are useful especially for assessing non-parametric identification (Pearl, 2009). Each framework thus has its own strengths and weaknesses and main areas of application, but the two approaches do not in general contradict each other. The focus of this thesis lies on causal graphs.

Intuitively, the nodes in a causal graph represent random variables, and an edge $A \to B$ means that A has a direct causal effect on B relative to the other variables represented in the graph. A causal graph is ideally constructed based on subject-matter knowledge about the system of interest. It can then be used to read off whether an *estimand*, i.e. target of inference, is non-parametrically identified. For example, if the estimand is the average causal effect of X on Y and the graph contains a *valid adjustment set* $\mathbf{Z}$ relative to (X, Y), then this estimand is *identified by adjustment*, which is the most popular non-parametric identification strategy (Imbens, 2004; Morgan and Winship, 2014). However, as the causal graph is often only partially known in practice, a large number of alternative, non-graphical strategies for selecting an adjustment set are in use. Among them are methods developed specifically for causal inference, while others, such as backward regression selection, are multi-purpose variable selection tools popular also for predictive and descriptive model building (Heinze et al., 2018). In this work, I refer to the process of choosing an adjustment set for causal inference as 'confounder selection', and call any method used for this purpose a 'confounder selection strategy'. It should be noted, however, that the term *confounder* itself is difficult to define (VanderWeele and Shpitser, 2013); it loosely refers to a variable that is part of one or more valid adjustment set.

The process of estimating a causal graph from data is called *causal search*, *causal structure learning* or *causal discovery* (Spirtes et al., 2000; Heinze-Deml et al., 2018; Glymour et al., 2019). Even though it relies on strong (causal) assumptions, causal discovery can be a valuable addition to analyses focussing on individual treatment-outcome pairs. A famous example of a causal discovery algorithm is the PC-

algorithm (Spirtes et al., 2000), named after its inventors, Peter Spirtes and Clark Glymour.

1.1 Aim and structure of this work

In this cumulative dissertation thesis, I investigate aspects of model selection for causal inference, where 'model selection' includes both confounder selection and selection of a causal graph based on data, i.e. causal discovery. Throughout, I assume that the system under study can be represented by a causal graph, which may be known, partly known or unknown. Chapter 2 provides the necessary background on graph terminology, the probabilistic and causal interpretation of different types of causal graphs, and causal discovery.

Chapters 3 and 4 are concerned with confounder selection. Data-driven confounder selection strategies are popular, but most of them were not developed with causal inference in mind. The additional assumptions necessary for ensuring that a valid adjustment set is selected are rarely discussed. Further, it is not clear how these algorithmic strategies relate to alternative, knowledge-based rules, in terms of their assumptions and type of selected adjustment set. In Chapter 3, I formalise different knowledge-based and data-driven strategies using a common notation. I then analyse them from a graphical perspective and compare the assumptions under which each strategy selects a valid adjustment set. The publication associated with Chapter 3, *Witte and Didelez (2019), Biometrical Journal 61(5):1270–1289*, contributes a classification scheme shedding further light onto the differences and commonalities of the various strategies, and a simulation study in which different selection methods are combined with different adjustment methods.

One of the results in Chapter 3 is that different types of adjustment sets result in different variances of the causal effect estimators. This is investigated further in Chapter 4 for linear regression adjustment. I show that if the variables follow a multivariate normal distribution, there exists a unique 'optimal' adjustment set in the sense that adjustment for this set results in a smaller asymptotic variance than adjusting for any other valid adjustment set. The optimal adjustment set can be characterised graphically and, under certain assumptions, coincides with the target set of backward regression selection. Part of the results was published in *Witte, Henckel, Maathuis and Didelez (2020), Journal of Machine Learning Research 21(246):1–45*.

The topic of Chapter 5 is causal discovery with cohort data. In cohort studies, in-

dividuals are followed over time, and the number of measured variables is usually large, making this type of study a valuable resource for causal discovery analyses. I show in Chapter 5 how the PC-algorithm can be modified to efficiently exploit the partial temporal ordering while retaining its soundness and completeness. The modified algorithm is implemented in the `R` package `tpc` (Witte, 2021). One of the challenges posed by typical cohort data is missing values. In *Witte, Foraita and Didelez (2021), arXiv preprint arXiv:2108.13331*, test-wise deletion and multiple imputation are investigated as two possible solutions. The manuscript contains both theoretical and empirical results. Software implementing both methods is available in the `R` package `micd` (Foraita and Witte, 2021).

The thesis concludes with a discussion in Chapter 6, where I point out parallels between data-driven confounder selection and causal discovery, and critically discuss whether systems of interest in epidemiology can or should be represented by directed acyclic graphs.

2 Graphical models for causal inference

Graphs are used in numerous branches of science to depict systems of interconnected or interacting units. Examples include food webs in ecology, Feynman diagrams in theoretical physics and semantic networks in computational linguistics. In statistics, graphs represent dependence and independence relations in probability distributions (Lauritzen, 1996; Maathuis et al., 2018).

The first use of graphs with an explicit probabilistic as well as causal interpretation is usually attributed to the geneticist Sewall Wright (Denis and Legerski, 2006; Pearl and Mackenzie, 2018). Around 1930, he developed the method of *path analysis*, where *path diagrams* represent the association structure among a set of variables, and *path coefficients* quantify linear relations (Wright, 1921, 1934). While Wright deduced his path diagrams from subject-matter knowledge and gave them a causal interpretation, the method can also be used for purely associational analyses. Elements of path analysis were later combined with methods for handling latent variables, resulting in a suite of methods summarised under *structural equation modelling* (Denis and Legerski, 2006). The focus, however, remained on linear relations for several decades.

Non-parametric causal graphs, as they are popular today, had their breakthrough towards the end of the 20th century, when computer scientist Judea Pearl introduced the *do-notation* and developed the associated concept of *do-calculus* for the non-parametric identification of causal effects (see Section 2.3.1 of this thesis). Around the same time, Spirtes et al. (2000) laid the foundations of causal discovery. The new versatile and explicitly causal graphical framework was quickly endorsed by biometricians and epidemiologists. Following an influential introductory article by Greenland et al. (1999), a large number of tutorial-style papers were published in journals covering a wide variety of subdisciplines of medicine and epidemiology (e.g. Hernán et al., 2002; Shrier and Platt, 2008; Hardt et al., 2011; Williamson et al., 2014; Suttorp et al., 2015; Staplin et al., 2017). The first

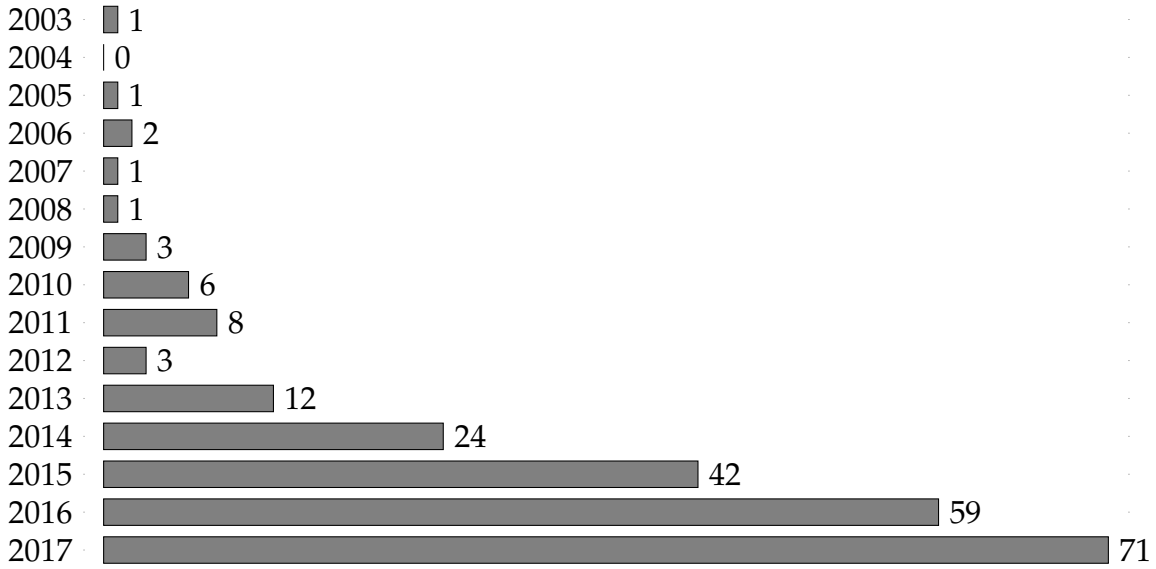


Figure 1: Number of publications in applied health research reporting the use of causal graphs, for years 2003–2017. The number of publications between 1990 and 2002 was zero. Modified from Tennant et al. (2020).

applications of causal graphs in health research that went beyond toy examples followed in the early 2000s, as investigated by Tennant et al. (2020). Their systematic review identified all published uses of causal graphs in applied health research during 1990–2017. The earliest of 234 articles they found was from 2003, and the numbers increased especially between 2013 and 2017, see Figure 1.

Other disciplines such as economics, psychology and sociology have been more hesitant in taking up the non-parametric causal graph framework. This is partly due to the fact that non-parametric identification by adjustment, which connects particularly well with graphical modelling, is less popular in these fields. Instead, identification methods e.g. in econometrics often rely on parametric and functional assumptions such as linearity, monotonicity and convexity, which can better be captured using potential outcomes (Imbens, 2020). When it comes to graphical modelling, linear structural equation models are often the method of choice in particular in psychology and sociology (Bollen and Pearl, 2013). Still, in all of the above disciplines, the interest in non-parametric causal graphs is growing, as can be witnessed from a large number of recently published textbooks (e.g. Morgan and Winship, 2014; Cunningham, 2021), introductory works (e.g. Rohrer, 2018; Hünermund and Bareinboim, 2019; Dablander, 2020) and works 'translating' assumptions and methods from other frameworks into graphical terms (e.g. Mansournia et al., 2013; Kim and Steiner, 2021a,b; Mohan and Pearl, 2021).

This chapter provides an overview about causal graphs and causal discovery. In Section 2.1, I introduce the basic graphical terminology used throughout the thesis. Additional terms will be added in later sections and chapters when needed. Sections 2.2 and 2.3 cover the probabilistic and causal interpretation, respectively, of *causal directed acyclic graphs* (*causal DAGs*). If two or more DAGs have the same probabilistic interpretation, they can collectively be represented by a *completed partially directed acyclic graph* (*CPDAG*). The class of CPDAGs is a subclass of the larger class of *maximally oriented partially directed acyclic graphs* (*MPDAGs*). Both classes are introduced in Section 2.4. Section 2.5 is about *acyclic directed mixed graphs* (*ADMGs*), which can be derived from DAGs by projection over unobserved nodes. Finally, Section 2.6 contains an introduction to causal discovery.

2.1 Basic terminology

Nodes and edges

A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ consists of a set of *nodes* $\mathbf{V}$ and a set of *edges* $\mathbf{E}$. An edge connects two distinct nodes, which are called its *endpoints*. Edges can be *undirected* ($-$), *directed* ($\rightarrow$) or *bi-directed* ($\leftrightarrow$). A directed edge $A \rightarrow B$ is said to be *directed from* or *out of* A and *directed to* or *into* B. A graph with at most one edge between a given pair of nodes is called a *simple graph*. An *undirected graph* is a simple graph where all edges are undirected, and a *directed graph* is a simple graph where all edges are directed.

Induced subgraphs

An induced subgraph of a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a graph $\mathcal{G}' = (\mathbf{V}', \mathbf{E}')$ such that $\mathbf{V}' \subseteq \mathbf{V}$ and $\mathbf{E}'$ is the subset of $\mathbf{E}$ containing all edges in $\mathcal{G}$ among nodes in $\mathbf{V}'$.

Paths

A path is an ordered sequence $(V_0, e_1, V_1, \ldots, e_K, V_K)$ of distinct nodes $V_0, V_1, \ldots, V_K$ and edges $e_1, \ldots, e_K$, $K \geq 1$, such that for $k = 1, \ldots, K$, e_k has endpoints V_{k-1} and V_k. The nodes and edges in the sequence are said to be *on* the path. In a simple graph, a path is uniquely determined by the sequence of its nodes. The path $(V_0, e_1, V_1, \ldots, e_K, V_K)$ has *endpoints* V_0 and V_K, and *non-endpoints* $V_1, \ldots, V_{K-1}$. Given a path $p = (V_0, e_1, V_1, \ldots, e_K, V_K)$, the subsequence $(V_i, e_{i+1}, V_{i+1}, \ldots, e_j, V_j)$, $0 \leq i, j \leq K$, is called the *subpath of p between V_i and V_j*.

Consider the path $p = (V_0, e_1, V_1, \ldots, e_K, V_K)$ and two disjoint sets of nodes $\mathbf{A}$ and $\mathbf{B}$ such that $V_0 \in \mathbf{A}$ and $V_K \in \mathbf{B}$. Then p is said to be *between* $\mathbf{A}$ and $\mathbf{B}$. It is *directed*

from **A** *to* **B** if for $k = 1, \dots, K$, e_k is directed from V_{k-1} to V_k. It is *possibly directed from* **A** *to* **B** if for $k = 1, \dots, K$, e_k is either directed from V_{k-1} to V_k or undirected and there is no edge $V_i \leftarrow V_j$, $0 \leq i < j \leq K$, in $\mathcal{G}$. Note that the edge $V_i \leftarrow V_j$ is not required to be on p. This non-standard definition of a possibly directed path is required for the correct interpretation of causal MPDAGs, see Section 2.4. If p is not possibly directed from **A** to **B**, it is *non-directed from* **A** *to* **B**. A directed, possibly directed or non-directed path from **A** to **B** is *proper* if only one of its nodes is in **A**.

Ancestry

Two nodes A, B joined by at least one edge are *adjacent* to each other, otherwise they are *non-adjacent*. If $A - B$, then A and B are *siblings* of each other. If $A \rightarrow B$, then A is a *parent* of B, and B is a *child* of A. If there is a directed path from A to B, or if $A = B$, then A is an *ancestor* of B, and B is a *descendant* of A. If there is a possibly directed path from A to B, or if $A = B$, then A is a *possible ancestor* of B, and B is a *possible descendant* of A. A node that is not a possible descendant of A is called a *non-descendant* of A.

The sets of all siblings, parents, children, ancestors, possible ancestors, descendants, possible descendants and non-descendants of a node A are denoted as $\mathrm{sib}(A)$, $\mathrm{pa}(A)$, $\mathrm{ch}(A)$, $\mathrm{an}(A)$, $\mathrm{possan}(A)$, $\mathrm{de}(A)$, $\mathrm{possde}(A)$ and $\mathrm{nde}(A)$, respectively. For a set of nodes **A**, $\mathrm{sib}(\mathbf{A}) = \bigcup_{A \in \mathbf{A}} \mathrm{sib}(A)$, and analogously for $\mathrm{pa}(\mathbf{A})$, $\mathrm{ch}(\mathbf{A})$, $\mathrm{an}(\mathbf{A})$, $\mathrm{possan}(\mathbf{A})$, $\mathrm{de}(\mathbf{A})$, $\mathrm{possde}(\mathbf{A})$ and $\mathrm{nde}(\mathbf{A})$.

Colliders and definite-status paths

A *collider* on a path $p = (V_0, e_1, V_1, \dots, e_K, V_K)$ is a non-endpoint node V_i such that the subpath of p between V_{i-1} and V_{i+1} is one of $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$, $V_{i-1} \leftrightarrow V_i \leftarrow V_{i+1}$, $V_{i-1} \rightarrow V_i \leftrightarrow V_{i+1}$ and $V_{i-1} \leftrightarrow V_i \leftrightarrow V_{i+1}$. A *non-collider* on p is a non-endpoint node V_i such that either (i) $V_{i-1} \leftarrow V_i$ is on p, or (ii) $V_i \rightarrow V_{i+1}$ is on p, or (iii) the subpath of p between V_{i-1} and V_{i+1} is $V_{i-1} - V_i - V_{i+1}$, and V_{i-1} and V_{i+1} are not adjacent in the graph. A *definite-status* path is one on which every non-endpoint node is either a collider or a non-collider. A *v-structure* is a path $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$ such that V_{i-1} and V_{i+1} are non-adjacent in the graph.

DAGs

A *directed cycle* is a directed path from a node A to a node B together with an edge $B \rightarrow A$. A DAG is a directed graph without directed cycles. DAGs are called *acyclic digraphs* in some fields of the literature (e.g. Andersson et al., 1997).

Blocking and d-separation in DAGs (Pearl, 2009)

A path p in a DAG is *open given* a possibly empty set of nodes $\mathbf{Z}$ if (i) no non-collider on p is in $\mathbf{Z}$ and (ii) every collider on p has a descendant in $\mathbf{Z}$. Otherwise, p is *blocked given* $\mathbf{Z}$. Two disjoint sets of nodes $\mathbf{A}$ and $\mathbf{B}$ are *d-separated given* a possibly empty set $\mathbf{C}$ in a DAG $\mathcal{D}$ if all paths between $\mathbf{A}$ and $\mathbf{B}$ are blocked given $\mathbf{C}$ in $\mathcal{D}$. This is denoted as $\mathbf{A} \perp_{\mathcal{D}} \mathbf{B} \mid \mathbf{C}$.

2.2 Probabilistic modelling with DAGs

A DAG with node set $\mathbf{V}$ can be used to represent conditional independencies among a set of variables $\mathbf{V}$. Throughout the thesis, I use the convention that a node (or set of nodes) and its corresponding variable (or set of variables) are denoted with the same symbol. Further, in order to improve the readability of some of the proofs in later sections, I sometimes consider $\mathbf{V}$ to be a vector of random variables instead of a set. This affects the notation but does not change the meaning of the graphs or distributions.

Briefly, both DAGs and probability distributions induce *independence models*. A probability distribution is represented by a DAG if every conditional independence implied by the DAG is implied by the distribution as well. This is now explained in detail.

An independence model $\mathcal{I}$ over some set $\mathbf{V}$ is a set of triples $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle$, where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are disjoint subsets of $\mathbf{V}$, and $\mathbf{C}$ is possibly empty. The triple $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle$ is interpreted as a conditional independence statement: $\mathbf{A}$ *is conditionally independent of* $\mathbf{B}$ *given* $\mathbf{C}$.

A DAG $\mathcal{D}$ is associated with an independence model $\mathcal{I}(\mathcal{D})$ by letting every d-separation stand for a conditional independence statement between nodes in $\mathcal{D}$:

$$\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}(\mathcal{D}) \quad \Leftrightarrow \quad \mathbf{A} \perp_{\mathcal{D}} \mathbf{B} \mid \mathbf{C}.$$

As an example, consider a DAG $\mathcal{D}$ of the form $A \to B \to C$. Here $A \perp_{\mathcal{D}} C \mid B$, and no other d-separations hold. The independence model induced by $\mathcal{D}$ is thus $\mathcal{I}(\mathcal{D}) = \{\langle A, C \mid B \rangle\}$. The same independence model is induced by the DAGs $A \leftarrow B \leftarrow C$ and $A \leftarrow B \to C$. DAGs inducing the same independence model are said to be *Markov equivalent* and form a (*Markov*) *equivalence class* (Andersson et al., 1997). The term originated from the Markov property discussed below.

Independence models induced by DAGs are compositional graphoids, as defined next.

Definition 1 (Semi-graphoid, graphoid, compositional; Lauritzen and Sadeghi, 2018)
An independence model $\mathcal{I}$ over a set $\mathbf{V}$ is a semi-graphoid *if the following properties hold for disjoint subsets $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D} \subset \mathbf{V}$:*

(I1) *if $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}$, then $\langle \mathbf{B}, \mathbf{A} \mid \mathbf{C} \rangle \in \mathcal{I}$ (symmetry),*

(I2) *if $\langle \mathbf{A}, \mathbf{B} \cup \mathbf{D} \mid \mathbf{C} \rangle \in \mathcal{I}$, then $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}$ and $\langle \mathbf{A}, \mathbf{D} \mid \mathbf{C} \rangle \in \mathcal{I}$ (decomposition),*

(I3) *if $\langle \mathbf{A}, \mathbf{B} \cup \mathbf{D} \mid \mathbf{C} \rangle \in \mathcal{I}$, then $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \cup \mathbf{D} \rangle \in \mathcal{I}$ (weak union),*

(I4) *if $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}$ and $\langle \mathbf{A}, \mathbf{D} \mid \mathbf{B} \cup \mathbf{C} \rangle \in \mathcal{I}$, then $\langle \mathbf{A}, \mathbf{B} \cup \mathbf{D} \mid \mathbf{C} \rangle \in \mathcal{I}$ (contraction).*

It is a graphoid *if in addition,*

(I5) *if $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \cup \mathbf{D} \rangle \in \mathcal{I}$ and $\langle \mathbf{A}, \mathbf{C} \mid \mathbf{B} \cup \mathbf{D} \rangle \in \mathcal{I}$, then $\langle \mathbf{A}, \mathbf{B} \cup \mathbf{C} \mid \mathbf{D} \rangle \in \mathcal{I}$ (intersection).*

It is compositional *if it holds that*

(I6) *if $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}$ and $\langle \mathbf{A}, \mathbf{D} \mid \mathbf{C} \rangle \in \mathcal{I}$, then $\langle \mathbf{A}, \mathbf{B} \cup \mathbf{D} \mid \mathbf{C} \rangle \in \mathcal{I}$ (composition).*

The compositional graphoid properties do not form a complete set of axioms for DAG-induced independence models. In fact, it has been conjectured that no complete set of axioms exists (Geiger, 1987). One of the additional properties satisfied by all DAG-induced independence models, and that will be needed for two proofs in this thesis, is weak transitivity.

Definition 2 (Weak transitivity; Pearl, 1988)
An independence model $\mathcal{I}$ over a set $\mathbf{V}$ has the weak transitivity *property if, for disjoint subsets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subset \mathbf{V}$ and a singleton $D \in \mathbf{V} \setminus (\mathbf{A} \cup \mathbf{B} \cup \mathbf{C})$, it holds that:*

(I7) *if $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \cup D \rangle$ and $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle$, then either $\langle \mathbf{A}, D \mid \mathbf{C} \rangle$ or $\langle \mathbf{B}, D \mid \mathbf{C} \rangle$ (weak transitivity).*

Consider now a set of random variables $\mathbf{V}$ with joint distribution P. Following Lauritzen (1996), I use f as a generic symbol for the probability density of continuous, discrete or mixed variables. For disjoint subsets $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ of $\mathbf{V}$, $\mathbf{A}$ is said to be conditionally independent of $\mathbf{B}$ given $\mathbf{C}$ if $f(\mathbf{a} \mid \mathbf{b}, \mathbf{c}) = f(\mathbf{a} \mid \mathbf{c})$. This is denoted as $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$, where I suppress an explicit reference to the distribution P. The independence model induced by P is defined as the set of all conditional independencies implied by P:

$$\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}(P) \quad \Leftrightarrow \quad \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}.$$

Independence models induced by probability distributions are *semi-graphoids*. If the probability distribution has a strictly positive density, the corresponding independence model is also a *graphoid*.

A probability distribution is represented by a given DAG if it satisfies the Markov property with respect to that DAG.

Definition 3 (Markov property)
A distribution P over a set of random variables $\mathbf{V}$ is Markov *to a DAG $\mathcal{D}$ with node set $\mathbf{V}$ if for all disjoint subsets $\mathbf{A}, \mathbf{B}, \mathbf{C}, \subset \mathbf{V}$,*

$$\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}(\mathcal{D}) \quad \Rightarrow \quad \langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}(\mathcal{P})$$

or, equivalently,

$$\mathbf{A} \perp_{\mathcal{D}} \mathbf{B} \mid \mathbf{C} \quad \Rightarrow \quad \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}.$$

It can be shown that the Markov property holds if and only if the density $f(\mathbf{v})$ factorises according to

$$f(\mathbf{v}) = \prod_{V \in \mathbf{V}} f(v \mid \mathrm{pa}(V, \mathcal{D})) \tag{1}$$

(Lauritzen et al., 1990). The name 'Markov property' stems from this recursive factorisation expression.

A stronger condition than the Markov property is faithfulness.

Definition 4 (Faithfulness)
A distribution P over a set of random variables $\mathbf{V}$ is faithful *to a DAG $\mathcal{D}$ with node set $\mathbf{V}$*

if for all disjoint subsets $\mathbf{A}, \mathbf{B}, \mathbf{C}, \subset \mathbf{V}$,

$$\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}(\mathcal{D}) \quad \Leftrightarrow \quad \langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}(\mathcal{P})$$

or, equivalently,

$$\mathbf{A} \perp_{\mathcal{D}} \mathbf{B} \mid \mathbf{C} \quad \Leftrightarrow \quad \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}.$$

Faithfulness implies that the independence model induced by the distribution P is a compositional graphoid. Causal discovery algorithms often assume that the distribution of the variables of interest is faithful to an unknown underlying DAG (see Section 2.6).

2.3 Causal interpretation of DAGs

This section is concerned with the additional assumptions under which a probabilistic DAG model can be interpreted causally. Although the term 'causal DAG' is often used without further specification, there are in fact several (related) ways of giving a graphical model a causal interpretation. In the following, I first present the definitions and concepts used in this thesis, then briefly discuss alternatives.

2.3.1 Definition used in this thesis

Consider the joint density $f(\mathbf{v})$ over a set of variables $\mathbf{V}$. Spirtes et al. (2000) and Pearl (2009) introduced the notion of the (*atomic*) *manipulation* or *intervention*, which sets the value of a subset $\mathbf{X} \subset \mathbf{V}$ to $\mathbf{x}$, without directly influencing the values of the other variables. This is denoted by $do(\mathbf{X} = \mathbf{x})$, or $do(\mathbf{x})$ for short. Importantly, the atomic intervention is a (hypothetical) intervention on the system governing the distribution of the variables in $\mathbf{V}$, not a mathematical operation on $f(\mathbf{v})$. The *interventional density* of $\mathbf{V}$ after the intervention $do(\mathbf{x})$ has taken place is denoted by $f(\mathbf{v} \mid do(\mathbf{X} = \mathbf{x}))$ or $f(\mathbf{v} \mid do(\mathbf{x}))$ for short. I keep with this popular notation, even though it has been criticised for the danger of being confused with the conditional density $f(\mathbf{v} \mid \mathbf{X} = \mathbf{x})$. Using the do-notation, the *observational density* $f(\mathbf{v})$ can be expressed as $f(\mathbf{v} \mid do(\emptyset))$, where $do(\emptyset)$ stands for no intervention and is called the *observational regime*. To avoid notational clutter, I denote the observational density by just $f(\mathbf{v})$ in this thesis. The conditional interventional density of $\mathbf{V}$ given $\mathbf{W} = \mathbf{w}$ after the intervention $do(\mathbf{x})$, for $\mathbf{X} \cap \mathbf{W} = \emptyset$, is denoted as $f(\mathbf{v} \mid \mathbf{W} = \mathbf{w}; do(\mathbf{x}))$.

Definition 5 (Causal DAG; Spirtes et al., 2000; Pearl, 2009)
Let $\mathcal{D}$ be a DAG with node set $\mathbf{V}$. A joint density $f(\mathbf{v})$ is compatible with *$\mathcal{D}$ if for all $\mathbf{X} \subseteq \mathbf{V}$ and all $\mathbf{x}' \in \mathcal{X}$, the interventional density $f(\mathbf{v} \mid do(\mathbf{x}'))$ exists and can be written as*

$$f(\mathbf{v} \mid do(\mathbf{x}')) = \mathbf{1}(\mathbf{x} = \mathbf{x}') \prod_{V \in \mathbf{V} \setminus \mathbf{X}} f(v \mid \mathrm{pa}(V, \mathcal{D})), \tag{2}$$

where the indicator function $\mathbf{1}(\mathbf{x} = \mathbf{x}')$ equals 1 if $\mathbf{x} = \mathbf{x}'$, and 0 otherwise. The DAG $\mathcal{D}$ is then called a causal DAG.

Here $\mathcal{X}$ denotes the state space of $\mathbf{X}$. Equation (2) is called the *manipulation formula* (Spirtes et al., 2000) or the *truncated factorisation formula* (Pearl, 2009), due to its similarity to the 'untruncated' factorisation formula describing the Markov property in equation (1). The 'truncation term' $\mathbf{1}(\mathbf{x} = \mathbf{x}')$ sets the density to zero for all values of $\mathbf{x}$ not consistent with the intervention $do(\mathbf{x}')$. The conditional distributions of all non-manipulated variables, given their respective parents, are the same as in the observational regime. This assumes that the conditional distributions are *stable* (Dawid and Didelez, 2010), which is often motivated by the concept of autonomous physical mechanisms (Pearl, 2009): The idea is that each conditional density $f(v \mid \mathrm{pa}(V, \mathcal{D}))$ represents a physical mechanism taking as input the value of $\mathrm{pa}(V, \mathcal{D})$, and outputting a value v. The output hence depends on the input, but crucially, the output does not depend on whether the input arose 'naturally' or by intervention (Peters et al., 2017). Roughly said, the stable conditional distributions are what allows us to estimate causal effects from observational data (Dawid and Didelez, 2010).

In a causal DAG, directed paths are also called *causal* paths, and non-directed paths are called *non-causal* paths. An ancestor of a node A in a causal DAG is also called a *cause* of A.

2.3.2 Alternative definitions

Various alternative definitions of a causal DAG have been given in the literature, and the differences between them are often subtle (Didelez, 2018). Definition 5 of a causal DAG is sometimes called the 'Pearlian DAG' (Dawid, 2010). Two notable, though less popular, alternatives to the Pearlian DAG are the *influence diagram* (Dawid, 2002) and the *single world intervention graph* (*SWIG*; Richardson and Robins, 2013). They have in common that they represent a joint distribution under one intervention of interest. In influence diagrams, the intervention is depicted by

a special type of node without parents and with outgoing edges into all nodes representing variables directly affected by the intervention. SWIGs have split nodes where one half represents a variable in the observational regime and the other half represents the same variable under the intervention. Unlike Pearlian DAGs, intervention graphs and SWIGs do not rely on the assumption that every variable represented in the graph can be intervened on.

A Pearlian DAG $\mathcal{D}$ can be combined with a parametric or non-parametric structural equation model to represent a *functional causal model* (Pearl, 2009). For each node V in $\mathcal{D}$, it is then assumed that the variable V is generated as a function of its parents and an error term ε as $V \leftarrow h(\text{pa}(V, \mathcal{D}), \varepsilon)$. Functional causal models impose a joint distribution over all variables in the model under all possible interventions, including a joint distribution over different versions of the same variable under different interventions. The functional causal model has therefore been called the *multiple-world model*, in contrast to the Pearlian DAG aka the *single-world model*, which considers only one 'world' or intervention at a time (Shpitser and Tchetgen Tchetgen, 2016). Hence, the functional causal model relies on stronger assumption than the Pearlian DAG, even if no functional forms are specified. In this thesis, the additional assumptions of the functional causal model are not needed.

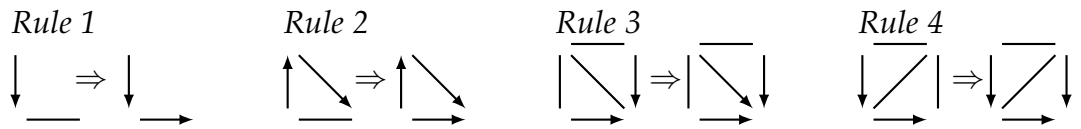
2.4 CPDAGs and MPDAGs — representing Markov equivalent DAGs

Recall that two DAGs inducing the same independence model are said to be Markov equivalent, and the set of all DAGs inducing the same model is called a (Markov) equivalence class. It has been shown that a given equivalence class of DAGs can compactly be represented by a CPDAG (Andersson et al., 1997). CPDAGs are simple graphs allowed to contain directed and undirected edges. A directed edge $A \to B$ means that A is a parent of B in every DAG in the equivalence class, and an undirected edge $A - B$ means that the equivalence class contains at least one DAG with $A \to B$, and at least one with $A \leftarrow B$. The set of all DAGs represented by a CPDAG $\mathcal{G}$ is denoted as $[\mathcal{G}]$. CPDAGs are also sometimes called *essential graphs* (Andersson et al., 1997), *d-separation-equivalence patterns* (Hoyer et al., 2008) or simply *patterns* (Spirtes et al., 2000).

Two DAGs are Markov equivalent if and only if they have the same adjacencies and v-structures (Verma and Pearl, 1990). Further, it has been shown that the CPDAG representing a given equivalence class can be constructed from any one

DAG in this class as follows (Meek, 1995a): In step 1, all edges in the DAG are turned into undirected edges, except for those forming v-structures. In step 2, as many edges as possible are oriented using the following rules:

Definition 6 (Meek's rules; Meek, 1995a)
Let $\mathcal{G}$ be a simple graph containing only directed and undirected edges, and no directed cycles. If the schematic on the left-hand side matches an induced subgraph of $\mathcal{G}$, then replace an undirected edge by a directed edge in $\mathcal{G}$ according to the schematic on the right-hand side.



The soundness of Meek's rules can be demonstrated by showing that they prevent new v-structures and directed cycles. Meek (1995a) showed that the rules are also complete in the sense that all edges remaining undirected after exhaustive application of the rules, correspond to edges for which both directions occur within the equivalence class. Figure 2 shows an example CPDAG together with all DAGs in the equivalence class it represents, and illustrates how the CPDAG can be constructed from one of those DAGs.

A given CPDAG $\mathcal{G}$ may be modified by orienting additional edges, and again applying Meek's rules. The resulting graph is then called an MPDAG and uniquely represents the subset of DAGs in the equivalence class that are compatible with the additional orientations (Meek, 1995a; Perković et al., 2017). The interpretation of the edges is the same as in a CPDAG. Other names for MPDAGs are *aggregated partially directed acyclic graph* (Eigenmann et al., 2017), *interventional essential graph* (Hauser and Bühlmann, 2012) and *distribution equivalence pattern* (Hoyer et al., 2008). Figure 3 illustrates the construction of an MPDAG from the CPDAG in Figure 2 under the restriction that $V_4 \to V_1$.

The class of DAGs and the class of CPDAGs both form subclasses of the class of MPDAGs. All results in this thesis that hold for MPDAGs thus hold for DAGs and CPDAGs as well. The set of DAGs represented by a given MPDAG $\mathcal{G}$ is denoted as $[\mathcal{G}]$.

Blocking and m-separation in MPDAGs (Maathuis and Colombo, 2015)

In MPDAGs, blocking is only defined for definite-status paths. A definite-status

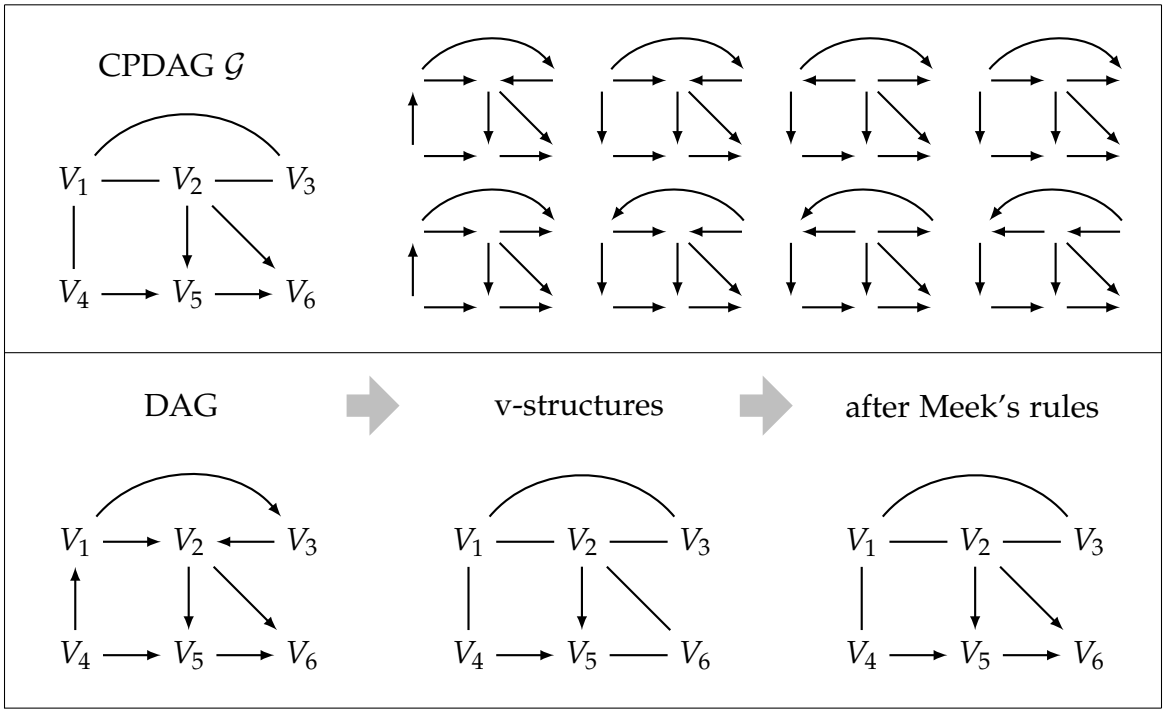


Figure 2: Upper panel: Example CPDAG $\mathcal{G}$ together with schematics of all DAGs in $[\mathcal{G}]$. Lower panel: Construction of $\mathcal{G}$ from one of the DAGs.

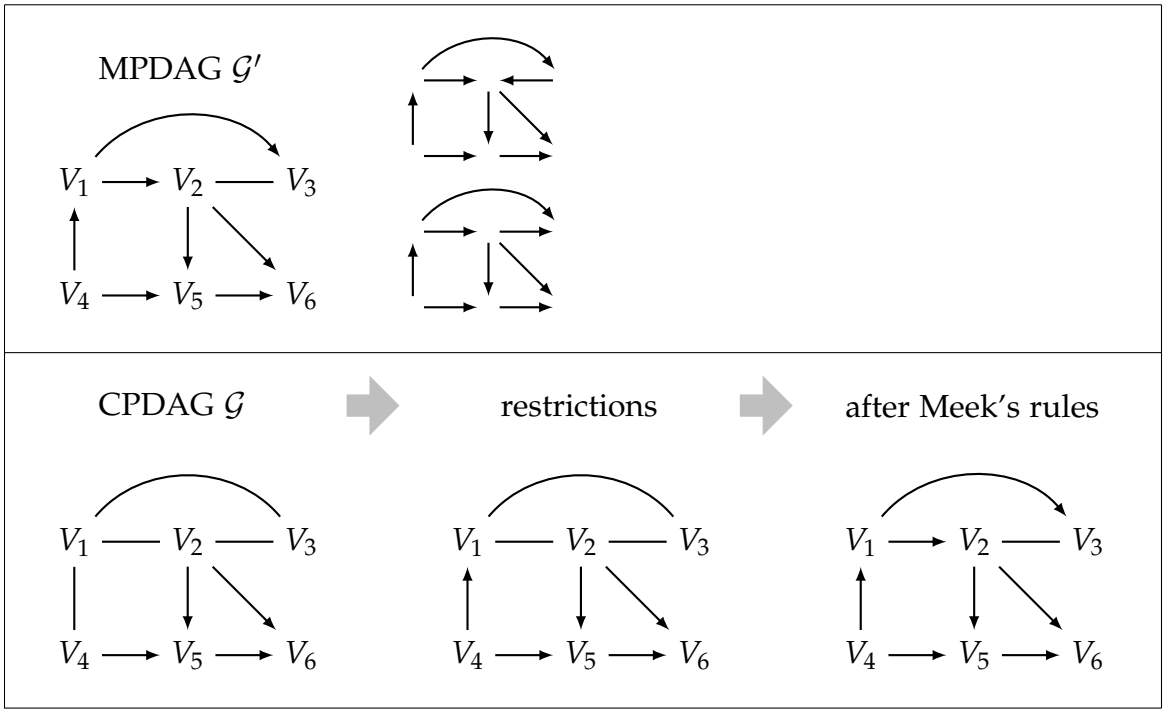


Figure 3: Upper panel: Example MPDAG $\mathcal{G}'$ together with schematics of all DAGs in $[\mathcal{G}']$. Lower panel: Construction of $\mathcal{G}'$ from the CPDAG $\mathcal{G}$ in Figure 2 under the edge restriction $V_4 \to V_1$.

path p in an MPDAG is *open given* a possibly empty set of nodes $\mathbf{Z}$ if (i) no non-collider on p is in $\mathbf{Z}$ and (ii) every collider on p has a descendant in $\mathbf{Z}$. Otherwise, p is *blocked given* $\mathbf{Z}$. Two disjoint sets of nodes $\mathbf{A}$ and $\mathbf{B}$ are *m-separated given* a possibly empty set $\mathbf{C}$ in an MPDAG $\mathcal{G}$ if all definite-status paths between $\mathbf{A}$ and $\mathbf{B}$ are blocked given $\mathbf{C}$ in $\mathcal{G}$. This is denoted as $\mathbf{A} \perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C}$. In a DAG, m-separation and d-separation coincide.

Maathuis and Colombo (2015) showed that for disjoint node sets $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ in an MPDAG $\mathcal{G}$, $\mathbf{A} \perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C}$ if and only if $\mathbf{A} \perp_{\mathcal{D}} \mathbf{B} \mid \mathbf{C}$ for every DAG $\mathcal{D} \in [\mathcal{G}]$. Thus, an MPDAG induces the same independence model as every DAG it represents.

Causal interpretation of MPDAGs (Perković et al., 2017)

An MPDAG is *causal* if it represents a causal DAG. When a causal MPDAG $\mathcal{G}$ is considered in practice, it is usually not known which of the DAGs in $[\mathcal{G}]$ has a causal interpretation. The DAGs in $[\mathcal{G}]$ are then called *possibly causal*.

In a causal MPDAG, directed paths are also called *causal* paths, possibly directed paths are called *possibly causal* paths and non-directed paths are called *non-causal* paths. This is analogous to DAGs. Perković et al. (2017) showed that a path in an MPDAG $\mathcal{G}$ is possibly causal if and only if $[\mathcal{G}]$ contains at least one DAG in which the corresponding path is causal, and at least one in which it is non-causal.

A density f is said to be *consistent with a causal MPDAG* $\mathcal{G}$ if it is consistent with one of the possibly causal DAGs in $[\mathcal{G}]$.

2.5 ADMGs — latent projection

If a (causal) DAG contains nodes representing latent, i.e. unobserved, variables, one may be interested in a graph representing the *marginal* independence model induced by the distribution over the observed nodes only. Such a graph can be constructed by *latent projection*. In the following definition, the sets of observed and latent variables are denoted as $\mathbf{W}$ and $\mathbf{L}$, respectively.

Definition 7 (Latent projection; Verma and Pearl, 1990; Shpitser et al., 2014)
Let $\mathcal{D}$ be a DAG with node set $\mathbf{W} \cup \mathbf{L}$ and $\mathbf{W} \cap \mathbf{L} = \emptyset$. The latent projection $\mathcal{D}(\mathbf{W})$ over $\mathbf{L}$ on $\mathbf{W}$ *is a graph with node set $\mathbf{W}$ and edges as follows: For distinct nodes $W_i, W_j \in \mathbf{W}$,*

(i) $\mathcal{D}(\mathbf{W})$ contains a directed edge $W_i \to W_j$ if and only if $\mathcal{D}$ contains a directed path $W_i \to \cdots \to W_j$ on which all non-endpoint nodes are in $\mathbf{L}$,

(ii) $\mathcal{D}(\mathbf{W})$ *contains a bi-directed edge* $W_i \leftrightarrow W_j$ *if and only if* $\mathcal{D}$ *contains a path, with at least one non-endpoint node, of the form* $W_i \leftarrow \cdots \rightarrow W_j$ *on which all non-endpoint nodes are non-colliders and in* $\mathbf{L}$.

The definition is quoted from Witte et al. (2020) for consistency. Latent projections are allowed to have directed and bi-directed edges, and may have two edges between a given pair of nodes. They belong to the class of ADMGs. Figure 4 shows example DAGs containing unobserved nodes, and their latent projections.

Blocking and m-separation in latent projections

The definitions of blocking and d-separation I gave in Section 2.1 may also be applied to ADMGs; d-separation is then called m-separation (Richardson, 2003), as in MPDAGs.

The latent projection $\mathcal{D}(\mathbf{W})$ derived from a DAG $\mathcal{D}$ faithfully represents the independence relations in $\mathcal{D}$ among the nodes in $\mathbf{W}$ in the following sense: For disjoint node sets $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbf{W}$, $\mathbf{A} \perp_{\mathcal{D}(\mathbf{W})} \mathbf{B} \mid \mathbf{C}$ if and only if $\mathbf{A} \perp_{\mathcal{D}} \mathbf{B} \mid \mathbf{C}$ (Shpitser et al., 2014).

Causal interpretation of latent projections

In a latent projection of a causal DAG, a directed edge $A \rightarrow B$ means that A is a cause of B in the DAG. A bi-directed edge means $A \leftrightarrow B$ means that A and B share a latent common cause in the DAG (see Section 2.6.1 for a formal definition of a common cause).

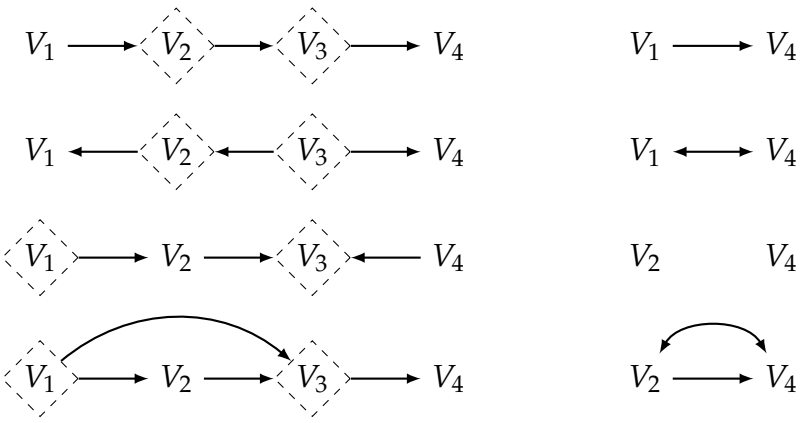


Figure 4: Example DAGs with latent nodes shown as dashed diamonds (left), and the corresponding latent projections (right).

2.6 Causal discovery

In the previous sections, it was described how DAGs and other types of graphs can compactly represent causal structures, and what information they contain about conditional independencies in observable and hypothetical distributions. Chapters 3 and 4 will illustrate how to read off adjustment sets from graphs when the aim is to estimate a causal effect. All this assumes that the graph is known or can be constructed based on subject-matter knowledge; the graph is then used in order to make inference about aspects of the probability distribution.

The aim of *causal discovery* is to achieve the opposite, i.e. to infer aspects of the underlying graph based on data. In this section, I first describe the PC-algorithm by Spirtes et al. (2000), which is one of the most popular causal discovery algorithms, and then give a brief overview about alternative strategies.

2.6.1 The PC-algorithm

The PC-algorithm was first proposed in Spirtes and Glymour (1991) and later improved by adding Meek's rules (Meek, 1995a; see Definition 6). The improved version has three phases: I. skeleton search, II. v-structure phase, III. application of Meek's rules.

I. Skeleton phase

The skeleton of a (partially) directed graph is an undirected graph with the same nodes and adjacencies. In phase I of the PC-algorithm, the skeleton is estimated based on the fact that two nodes V_i and V_j in a DAG with nodes set $\mathbf{V}$ are *not* adjacent if and only if they are separated given a set $\mathbf{S} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$ (Lauritzen, 1996, pp. 47). The separations are determined by testing for conditional independencies. Algorithm 1 shows pseudocode for this phase. It starts with a fully connected undirected graph in line 1. Edges are then gradually removed between pairs of nodes for which separating sets are found (see lines 11 and 12). The conditional independence tests are organised such that marginal tests are performed first, followed by conditional tests of order $|\mathbf{S}| = 1, 2, \ldots$, where the current order is determined by the parameter ℓ (see lines 3 and 5). In order to reduce the number of tests performed, only those pairs of nodes (V_i, V_j) are considered that are still adjacent in the current intermediate skeleton estimate (see line 7), and the candidate separating sets are chosen from among the current siblings of V_i (see line 10). Together, these restrictions have the effect that for reasonably sparse graphs, the majority of the tests performed are low-order tests ($|\mathbf{S}| = 0$ or $|\mathbf{S}| = 1$), which tend to have a

higher power than higher-order tests and are thus expected to be more reliable.

Algorithm 1 Skeleton phase of PC

INPUT: i.i.d. data on a set of variables $\mathbf{V}$; a procedure for testing conditional independencies

form the complete undirected graph $\mathcal{C}$ on node set $\mathbf{V}$
$\mathcal{C}' = \mathcal{C}$
$\ell = -1$
repeat
  $\ell = \ell + 1$
  repeat
    select new ordered pair of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ and $|\text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}| \geq \ell$
    repeat
      choose new $\mathbf{S} \subseteq \text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$
      if $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$ **then**
        delete edge $V_i - V_j$ from $\mathcal{C}'$
        record $\mathbf{S}$ as **Sepset**(V_i, V_j)
      end if
    until edge $V_i - V_j$ is deleted or all $\mathbf{S} \subseteq \text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$ have been chosen
  until all ordered pairs of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ have been chosen
until all ordered pairs of nodes (V_i, V_j) adjacent in $\mathcal{C}'$ satisfy $|\text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}| < \ell$

OUTPUT: estimated skeleton $\mathcal{C}'$; list of separating sets

II. V-structure phase

In the v-structure phase, PC goes through all triples of variables (V_i, V_j, V_k) in the estimated skeleton $\mathcal{C}'$ such that V_i and V_j are adjacent, V_j and V_k are adjacent, and V_i and V_k are not adjacent. There are four possible constellations of edge orientations, of which three are Markov equivalent: (i) $V_i \to V_j \to V_k$, (ii) $V_i \leftarrow V_j \to V_k$ and (iii) $V_i \leftarrow V_j \leftarrow V_k$ all imply that any set d-separating V_i and V_k must contain V_j, while (iv) $V_i \to V_j \leftarrow V_k$ implies that any set d-separating V_i and V_k must *not* contain V_j. Hence, the PC-algorithm checks whether V_j is included in the separating set of V_i and V_k found in the skeleton phase, and if this is not the case, orients the edges as $V_i \to V_j \leftarrow V_k$.

III. Meek's rules

In the last phase of the PC-algorithm, further edges are oriented by exhaustively applying Meek's rules (Meek, 1995b; Definition 6).

Pseudocode for all three phases is included in Algorithm 2 in Appendix A.1.

The theoretical properties of constraint-based causal discovery algorithms are often assessed by analysing the 'oracle version' of the respective algorithm, i.e. what the algorithm would do and return if the input was not data, but the list of true conditional independencies implied by the true DAG. Spirtes et al. (2000) and Meek (1995a) showed that oracle PC is *sound and complete*, i.e. it returns the true CPDAG, under the assumptions of faithfulness (Definition 4) and *causal sufficiency*.

Definition 8 (Causal sufficiency)
Let $\mathcal{D}$ be a DAG with node set $\mathbf{V}$. Then a subset $\mathbf{V}' \subseteq \mathbf{V}$ is causally sufficient *relative to $\mathcal{D}$ if for every pair (V_i, V_j), $V_i, V_j \in \mathbf{V}'$, $V_i \neq V_j$, the set of common causes of (V_i, V_j) is a subset of $\mathbf{V}'$.*

A *common cause* of a pair (V_i, V_j) of distinct nodes is a node V_k such that there is a directed path from V_k to V_i that does not include V_j, and a directed path from V_k to V_j that does not include V_i (Spirtes et al., 2000). A simple example is the causal graph $V_i \leftarrow V_k \rightarrow V_j$, where V_k is a common cause of (V_i, V_j).

Deriving theoretical guarantees for the PC-algorithm applied to data is more difficult. Assuming faithfulness and causal sufficiency, the PC-algorithm using any standard procedure for testing conditional independencies (e.g. Fisher's z-test or the G^2-test) is *pointwise consistent* (Robins et al., 2003). This means that for a given set of variables $\{X, Y\} \cup \mathbf{Z}$ in a given DAG, for every faithful distribution in which the null hypothesis $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ holds, there exists a sample size at which the probability of falsely rejecting the null is controlled, i.e. smaller than an arbitrary $\varepsilon > 0$; and for every faithful distribution in which the null does not hold, there exists a sample size at which the probability of falsely accepting the null is controlled. However, PC is not *uniformly consistent*, which would mean that there exists a sample size at which the total error probability (i.e. either falsely rejecting or falsely accepting the null hypothesis) if the distribution is unknown is controlled (Robins et al., 2003; Zhang and Spirtes, 2003). Intuitively, this is because without restricting the parameter space, it is possible to construct a faithful distribution that is arbitrarily close to an unfaithful distribution (Uhler et al., 2013). Uniform consistency does hold under assumptions stronger than faithfulness, e.g. the *λ-strong faithfulness* assumption for Gaussian distributions in Zhang and Spirtes (2003), where the parameters are restricted such that the conditional correlations among variables are always larger than a constant $\lambda > 0$.

While the output of oracle PC is always a CPDAG, the output of PC applied to data may be a partially directed graph not representing an equivalence class of DAGs, which makes the interpretation more difficult. Further, PC is column index-sensitive in the sense that shuffling the data column indices and re-running the algorithm can change the output, as certain decisions made during the algorithm depend on the sequence in which the variables are considered. Colombo and Maathuis (2014) demonstrated how shuffling the column indices and re-running the algorithm can lead to very different conclusions, and suggested an index-stable[1] version of PC, called *LMPC-stable* ('L' and 'M' stand for 'lists' and 'majority rule', respectively; see Appendix A.2 for more details). Besides solving the column index-sensitivity problem, LMPC-stable has the additional advantage that its skeleton phase can easily be parallelised on multi-core machines. However, a drawback is that it is even more likely with LMPC-stable to obtain an estimated graph that is not a CPDAG. Pseudocode for LMPC-stable is included in Algorithm 3 in Appendix A.2.

PC and LMPC-stable can be modified to take background knowledge in the form of required or forbidden edges or orientations into account. In Section 5.1, a version of LMPC-stable accounting for a partial node ordering is considered. The output of oracle PC or oracle LMPC-stable using background knowledge is an MPDAG (Meek, 1995a; Perković et al., 2017).

2.6.2 Other algorithms

Based on the main assumptions they rely on, causal discovery algorithms can roughly be divided into two classes (Glymour et al., 2019), which I call 'traditional' and 'functional'. The traditional methods rely on the faithfulness assumption. They can further be divided into *constraint-based*, *score-based* and *hybrid* algorithms. The PC-algorithm belongs to the constraint-based algorithms, which exploit the faithfulness assumptions rather directly by testing for conditional independencies and reconstructing the graph based on the constraints the test results imply. The *fast causal inference* (*FCI*) algorithm is a variant of PC that does not assume causal sufficiency (Spirtes et al., 2000; Zhang, 2008a). It uses different orientation rules than PC and outputs a so-called *partial ancestral graph* (*PAG*), which is less informative, in terms of the causal directions, than a CPDAG. An advantage of all constraint-based algorithms is that they are, in principle, non-parametric, even

[1] Note that Colombo and Maathuis (2014) used the terms 'order-dependent' and 'order-independent' to characterise PC and LMPC-stable, respectively. In this thesis, I decided to reserve the term 'order' in the context of causal discovery to refer to a time ordering or topological node ordering, see Chapter 5.

though in practice, conditional independencies are often tested using parametric tests.

In contrast to constraint-based algorithms, score-based algorithms always rely on parametric assumptions. They assign a parametric score to each DAG or CPDAG in the search space, for example a penalised Gaussian or multinomial likelihood. The search space is then usually traversed in a heuristic manner in order to maximise the score function. As an example, *greedy equivalence search* (*GES*) first adds and then removes edges in a greedy manner until an optimum is reached (Chickering, 2002). It assumes faithfulness and causal sufficiency and under these assumptions has been shown to be uniformly consistent (Chickering, 2002). Recently, score-based search has been reformulated as a continuous optimisation problem, which can be solved using deep learning (Zheng et al., 2018; Vowels et al., 2021). While the algorithm proposed by Zheng et al. (2018) in particular has been criticised for making unrealistic assumptions (Reisach et al., 2021), the general strategy of using continuous optimisation might prove to be a fruitful direction of development.

Hybrid algorithms combine constraint-based and score-based learning. An example is *adaptively restricted greedy equivalence search* (*ARGES*; Nandy et al., 2018), combining elements of PC with GES.

For the functional approach, a functional causal model is assumed, where each variable can be written as a function of its parents and an independent error term (see Section 2.3.2). In addition, assumptions are made about either the functional form or the distribution of the error terms or both. The DAG is then reconstructed by searching for asymmetries in the joint distribution entailed by these assumptions, for example by regressing a variable on its potential parent variables and checking whether the residuals follow the distribution assumed for the error terms. A popular example is the Linear Non-Gaussian Acyclic Model (LiNGAM) algorithm, which assumes non-linear functional forms and Gaussian errors (Shimizu et al., 2006).

3 A graphical perspective on confounder selection

The idea of adjusting for confounding factors in order to isolate a causal contrast of interest, is much older than the formal frameworks of causal inference. As a result, the 'traditional' literature is abound with vague, informal descriptions of adjustment in general and confounder selection in particular. For example, the aim of confounder selection in causal regression modelling has been described as selecting variables 'essential to the regression on the basis of theory' (Studenmund, 2014, p. 177). Without further explanation, such descriptions are not very useful.

Moreover, regression modelling in particular can serve very different purposes. Shmueli (2010) and Hernán et al. (2019) distinguished between descriptive, predictive and causal modelling. Accordingly, model selection will sometimes aim primarily at dimension reduction, sometimes at optimising the predictive performance in unseen data, and only in certain cases at confounder selection. Although the categories are not clear-cut and may overlap, it is certainly important to know what the aim of model selection is in any real-data application. Many textbooks, however, do not distinguish between the different purposes that regression modelling can have. There is a tendency to call every regression coefficient the 'effect' of an explanatory variable on the outcome, but the assumptions under which the coefficient can actually be given a causal interpretation are often not discussed in sufficient detail.

Fortunately, a shift in paradigm is visible. In newer textbooks, adjustment and confounder selection are approached from an explicitly causal, often graphical point of view (Morgan and Winship, 2014; Westreich, 2019; Hernán and Robins, 2020). Further, new methods for confounder selection have been proposed that do not require complete knowledge of the causal graph, but are still based on causal theory (de Luna et al., 2011; VanderWeele and Shpitser, 2011; Shortreed and Ertefaie, 2017). At the same time, 'traditional' data-driven confounder selection strategies such as stepwise regression selection and the change-in-estimate method

are still very popular (Weitzen et al., 2004; Walter and Tiemeier, 2009; Ali et al., 2015; Talbot and Massamba, 2019; Pressat-Laffouilhère et al., 2021), but they do not seem to fit into the graphical framework.

The aim of this chapter is to clarify the causal assumptions underlying 'traditional' and newer strategies for confounder selection, and to compare them with regard to the type of adjustment set they select. The chapter is organised as follows: In Section 3.1, I define a valid adjustment set and provide an overview over common classes of adjustment methods, including regression adjustment and propensity score methods. In Section 3.2, graphical criteria for confounder selection using a causal DAG are given. Section 3.3 summarises review articles investigating how adjustment sets are selected in the epidemiological practice. The most common methods identified in the reviews are investigated in detail in Section 3.4. In particular, I show under what causal assumptions the different methods select valid adjustment sets, given that the underlying causal structure can be represented by an unknown causal DAG. Section 3.5 contains the publication associated with this chapter, *Witte and Didelez (2019)*, which can be read as a complement to Section 3.4. It contributes a classification scheme and a simulation study providing further insight into the differences and commonalities between the confounder selection methods considered in this chapter. A special focus in the paper lies on estimation efficiency, i.e. the variance of the estimator, when adjusting for different types of selected sets.

3.1 Identification by adjustment and adjustment methods

Consider a set of *exposures* or *treatments* $\mathbf{X}$ with state space $\mathcal{X}$ and a set of *outcomes* $\mathbf{Y}$. If $\mathbf{X}$ and/or $\mathbf{Y}$ are singletons, which is a common case in practice, I denote them as X and Y, respectively. I call all other measured variables *covariates* and denote them as $\mathbf{W}$.

The (*total*) *causal effect* of $\mathbf{X}$ on $\mathbf{Y}$ is said to be *non-parametrically identified* if the interventional density $f(\mathbf{y} \mid do(\mathbf{x}))$, which is a function of both $\mathbf{y}$ and $\mathbf{x}$, can be re-expressed in terms of observational ('do-free') terms without making parametric assumptions (but structural assumptions are usually required). Note that the term 'causal effect' does not refer to a particular estimand, but rather to the collection of all estimands that can be written as functions of $f(\mathbf{y} \mid do(\mathbf{x}))$, $\mathbf{x} \in \mathcal{X}$.

Adjustment is a commonly used non-parametric identification strategy. The causal effect of $\mathbf{X}$ on $\mathbf{Y}$ is said to be *identified by adjustment* if there exists a set of variables $\mathbf{Z} \subseteq \mathbf{W}$ satisfying the following definition of a valid adjustment set:

Definition 9 (Valid adjustment set; Perković et al., 2018)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint sets of random variables, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *is a* valid adjustment set *relative to* $(\mathbf{X}, \mathbf{Y})$ *if for every* $\mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{y} \mid do(\mathbf{x})) = \begin{cases} f(\mathbf{y} \mid \mathbf{x}) & \text{if } \mathbf{Z} = \emptyset, \\ \int_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) f(\mathbf{z}) d\mathbf{z} & \text{otherwise.} \end{cases} \tag{3}$$

Equation (3) is called *adjustment formula*, *standardisation formula* or *G-formula* (Hernán and Robins, 2020). It holds if the following three conditions are satisfied: First, for all $\mathbf{z}$ such that $f(\mathbf{z}) > 0$, it must hold that $f(\mathbf{x} \mid \mathbf{z}) > 0$, as otherwise $f(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ is not well-defined. This condition is called *positivity*. Second, $f(\mathbf{y} \mid \mathbf{z}; do(\mathbf{x})) = f(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$; this is the *conditional exchangeability* or *no unobserved confounding* condition. Third, $f(\mathbf{z} \mid do(\mathbf{x})) = f(\mathbf{z})$, which is sometimes called the *pre-treatment* condition, expressing that $\mathbf{Z}$ 'happens before $\mathbf{X}$', in the sense that $\mathbf{Z}$ is not causally influenced by $\mathbf{X}$. Under these three conditions, it follows immediately that $f(\mathbf{y} \mid do(\mathbf{x})) = \int_{\mathbf{z}} f(\mathbf{y}, \mathbf{z} \mid do(\mathbf{x})) d\mathbf{z} = \int_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{z}; do(\mathbf{x})) f(\mathbf{z} \mid do(\mathbf{x})) d\mathbf{z} = \int_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) f(\mathbf{z}) d\mathbf{z}$.

Analyses in practice usually focus on specific aspects of $f(\mathbf{y} \mid do(\mathbf{x}))$. For binary X and binary or continuous Y, a commonly targeted estimand is the *average causal effect*

$$\tau = \mathrm{E}(Y \mid do(X = 1)) - \mathrm{E}(Y \mid do(X = 0)),$$

which, by equation (3), is identified from observational data on X, Y and a valid adjustment set $\mathbf{Z}$ as

$$\tau = \mathrm{E}_{\mathbf{Z}} \left[\mathrm{E}_Y(Y \mid X = 1, \mathbf{Z}) - \mathrm{E}_Y(Y \mid X = 0, \mathbf{Z}) \right] \tag{4}$$

For binary X and binary Y, an alternative estimand is the *marginal causal odds ratio*

$$\delta = \frac{P(Y = 1 \mid do(X = 1)) \ / \ P(Y = 0 \mid do(X = 1))}{P(Y = 1 \mid do(X = 0)) \ / \ P(Y = 0 \mid do(X = 0))},$$

identified as

$$\delta = \frac{\mathrm{E}_{\mathbf{Z}}[P(Y = 1 \mid X = 1, \mathbf{Z})] \ / \ \mathrm{E}_{\mathbf{Z}}[P(Y = 0 \mid X = 1, \mathbf{Z})]}{\mathrm{E}_{\mathbf{Z}}[P(Y = 1 \mid X = 0, \mathbf{Z})] \ / \ \mathrm{E}_{\mathbf{Z}}[P(Y = 0 \mid X = 0, \mathbf{Z})]}.$$

There are several options for estimating τ and δ. In the following, I give an overview about estimators for τ; the analogous estimators for δ are omitted.

Regression adjustment

Equation (4) suggests fitting a regression model for $\mathrm{E}(Y \mid X, \mathbf{Z})$ and estimating τ via *regression standardisation* as

$$\hat{\tau}^{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_1(\mathbf{Z}_i) - \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_0(\mathbf{Z}_i),$$

where N is the sample size, and $\hat{\mu}_1(\mathbf{Z}_i) = \hat{\mathrm{E}}(Y \mid X = 1, \mathbf{Z}_i)$ and $\hat{\mu}_0(\mathbf{Z}_i) = \hat{\mathrm{E}}(Y \mid X = 0, \mathbf{Z}_i)$ are predictions obtained from the fitted regression model. Zhang (2008b) showed that $\hat{\tau}^{\text{reg}}$ consistently estimates τ if a correctly specified logistic regression model is used. If a linear main effects model $\mathrm{E}(Y \mid X, \mathbf{Z}) = \beta_{y.x\mathbf{z}} + \beta_{yx.\mathbf{z}}X + \boldsymbol{\beta}_{y\mathbf{z}.x}^T\mathbf{Z}$ is assumed, then $\hat{\tau}^{\text{reg}} = 1/N\sum_{i=1}^{N}(\hat{\beta}_{y.x\mathbf{z}} + \hat{\beta}_{yx.\mathbf{z}} + \hat{\boldsymbol{\beta}}_{y\mathbf{z}.x}^T\mathbf{Z}_i) - 1/N\sum_{i=1}^{N}(\hat{\beta}_{y.x\mathbf{z}} + \hat{\boldsymbol{\beta}}_{y\mathbf{z}.x}^T\mathbf{Z}_i) = \hat{\beta}_{xy.\mathbf{z}}$, hence the standardisation step can be skipped and the estimator is the ordinary least squares estimator.

Inverse probability weighting

The inverse probability weighted estimator is defined as

$$\hat{\tau}^{\text{weight}} = \frac{1}{N}\sum_{i=1}^{N} \frac{X_i Y_i}{\hat{e}(\mathbf{Z}_i)} - \frac{1}{N}\sum_{i=1}^{N} \frac{(1 - X_i)Y_i}{1 - \hat{e}(\mathbf{Z}_i)},$$

where $\hat{e}(\mathbf{Z}_i) = \hat{P}(X = 1 \mid \mathbf{Z}_i)$ is the estimated *propensity score*. The intuition behind inverse probability weighting is to create a pseudo-population in which the distribution of the covariates $\mathbf{Z}$ is balanced between the two treatment groups, i.e. the group of individuals with $X = 1$ (treatment group) and the group of individuals with $X = 0$ (control group). Consider $\frac{1}{N}\sum_{i=1}^{N} \frac{X_i Y_i}{\hat{e}(\mathbf{Z}_i)}$, which estimates the weighted mean outcome in the treatment group. Here individuals with a covariate profile that is 'typical' for the treatment group are downweighted, as they have large propensity score terms in the denominator. Individuals that are 'typical' for the control group are upweighted, as their denominator terms are small. The converse is the case for $\frac{1}{N}\sum_{i=1}^{N} \frac{(1-X_i)Y_i}{1-\hat{e}(\mathbf{Z}_i)}$, which estimates the weighted mean outcome in the control group. As a consequence, in the weighted population, the distribution of $\mathbf{Z}$ is the same in both groups, and hence independent of X, just as it would be in a randomised experiment.

The inverse probability weighted estimator is a consistent estimator of τ if the propensity score model is correctly specified (Robins et al., 1994; Lunceford and

Davidian, 2004). In practice, the propensity score model is often fitted by logistic regression (Granger et al., 2020; Webster-Clark et al., 2021).

Doubly robust estimation

Doubly robust estimation combines regression adjustment and inverse probability weighting. The average treatment effect is estimated as

$$\hat{\tau}^{\text{dr}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i Y_i}{\hat{e}(\mathbf{Z}_i)} - \frac{X_i - \hat{e}(\mathbf{Z}_i)}{\hat{e}(\mathbf{Z}_i)}\hat{\mu}_1(\mathbf{Z}_i)\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(1-X_i)Y_i}{1-\hat{e}(\mathbf{Z}_i)} + \frac{X_i - \hat{e}(\mathbf{Z}_i)}{1-\hat{e}(\mathbf{Z}_i)}\hat{\mu}_0(\mathbf{Z}_i)\right).$$

This estimator is doubly robust in the sense that it is a consistent estimator of τ as long as either the outcome model ($\hat{\mu}_1(\mathbf{Z})$ and $\hat{\mu}_0(\mathbf{Z})$) or the treatment model $\hat{e}(\mathbf{Z}_i)$ is correctly specified (Scharfstein et al., 1999; Lunceford and Davidian, 2004). If both models are correct, then $\hat{\tau}^{\text{dr}}$ has a smaller asymptotic variance than $\hat{\tau}^{\text{weight}}$ (Robins et al., 1994; Lunceford and Davidian, 2004).

Matching

Similar to inverse probability weighting, the idea behind matching is to create a pseudo-population in which the distribution of the covariates is the same in the two treatment groups. However, in contrast to inverse probability weighting, matching is a non-parametric method. In 1:1 matching, each individual i in the treatment group is matched with an individual j from the control group such that $\mathbf{Z}_j$ is 'as close as possible' to $\mathbf{Z}_i$, and vice versa. Matching methods exist in many variants, using different metrics to measure the distance between $\mathbf{Z}_i$ and $\mathbf{Z}_j$, and different ways of dealing with individuals for whom no good match can be found. Further, matching can be done with and without replacement of already matched individuals to the pool of possible matches, and a ratio of 1:k may be used instead of matching 1:1 (Morgan and Winship, 2014). All matching procedures have in common that the average treatment effect is estimated from the matched population by simply subtracting the average outcome in the control group from the average outcome in the treatment group.

Finding good matches is harder when the number of variables in $\mathbf{Z}$ is large. An alternative is to match on the estimated propensity score, which has been shown to have the same balancing effect, in expectation, as matching directly on the covariates (Rosenbaum and Rubin, 1983). The propensity score model need not be correctly specified as long as balance is achieved, which can be checked empirically

for the observed covariates (Granger et al., 2020).

3.2 Adjustment criteria for DAGs

Valid adjustment sets can be read off causal DAGs, as explained next. An adjustment set is valid with respect to a causal graph if it is valid for every joint density compatible with the graph, as formalised in Maathuis and Colombo (2015):

Definition 10 (Valid adjustment set in a causal DAG)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$*, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *is a* valid adjustment set *relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$ *if* $\mathbf{Z}$ *is a valid adjustment set (according to Definition 9) relative to* $(\mathbf{X}, \mathbf{Y})$ *in every density compatible with* $\mathcal{D}$*.*

Pearl (1993) introduced the *back-door criterion* for checking whether a given set of variables is a valid adjustment set in a DAG. A *back-door path* from $\mathbf{X}$ to $\mathbf{Y}$ in a DAG $\mathcal{D}$ is a path between $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{D}$ that starts with a directed edge into $\mathbf{X}$.

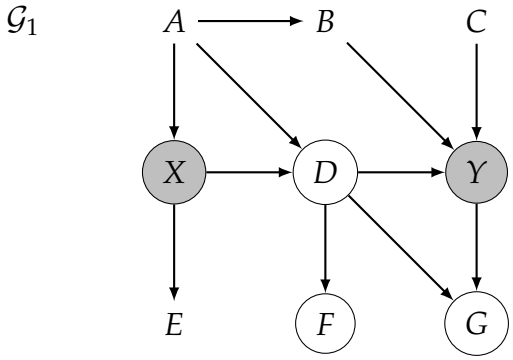
Definition 11 (Back-door criterion; Pearl, 2009)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$*, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *satisfies the* back-door criterion *relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$ *if*

(i) $\mathbf{Z} \cap \mathrm{de}(\mathbf{X}, \mathcal{D}) = \emptyset$ *and*

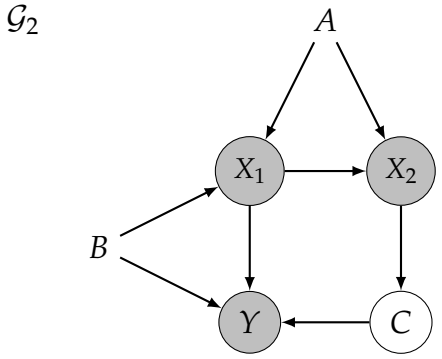
(ii) all back-door paths from $\mathbf{X}$ *to* $\mathbf{Y}$ *in* $\mathcal{D}$ *are blocked given* $\mathbf{Z}$*.*

Proposition 12 (Pearl, 2009)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$*, where* $\mathbf{Z}$ *is possibly empty. If* $\mathbf{Z}$ *satisfies the back-door criterion relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$*, then* $\mathbf{Z}$ *is a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$*.*

The back-door criterion is sufficient, but not necessary. An example for an adjustment set that does not satisfy the back-door criterion, but can be shown to be a valid adjustment set in the sense of Definition 10, is the set $\{Z\}$ in the causal DAG $Z \leftarrow X \rightarrow Y$, when X is the treatment and Y is the outcome of interest. Here $\{Z\}$ satisfies condition (ii) of the back-door criterion, but not condition (i), as Z is a descendant of X. A sufficient and necessary criterion, the *adjustment*



Non-causal paths from X to Y:
$X \leftarrow A \rightarrow B \rightarrow Y$
$X \leftarrow A \rightarrow D \rightarrow Y$
$X \leftarrow A \rightarrow D \rightarrow G \leftarrow Y$
$X \rightarrow D \leftarrow A \rightarrow B \rightarrow Y$
$X \rightarrow D \rightarrow G \leftarrow Y$



Proper non-causal paths from X to Y:
$X_1 \leftarrow B \rightarrow Y$

Figure 5: Example causal DAGs. $\mathcal{G}_1$: Shown in circles are the forbidden nodes with respect to treatment X and outcome Y (highlighted in grey). The sets $\{A\}$, $\{A, B\}$, $\{A, C\}$, $\{A, E\}$, $\{A, B, C\}$, $\{A, B, E\}$, $\{A, C, E\}$ and $\{A, B, C, E\}$ are valid adjustment sets relative to (X, Y) in $\mathcal{G}_1$. $\mathcal{G}_2$: Shown in circles are the forbidden nodes with respect to treatment $\mathbf{X} = \{X_1, X_2\}$ and outcome Y (highlighted in grey). The sets $\{B\}$ and $\{A, B\}$ are valid adjustment sets relative to $(\mathbf{X}, Y)$ in $\mathcal{G}_2$.

criterion, was proposed by Shpitser et al. (2010). Consider the following additional terminology for disjoint sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ in a causal DAG $\mathcal{D}$: The *causal nodes* $\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$ are all nodes on causal paths from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}$, excluding the nodes in $\mathbf{X}$. The *forbidden set* with respect to $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{D}$ is defined as $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \text{de}(\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D}), \mathcal{D}) \cup \mathbf{X}$. It contains the 'mediators' (the term is used informally here) of the causal effect of interest, together with their descendants. The nodes in the forbidden set are called *forbidden nodes*. Figure 5 shows two example DAGs illustrating the concept of the forbidden set and the adjustment criterion.

Definition 13 (Adjustment criterion[1]; Shpitser et al., 2010; van der Zander et al., 2014)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$*, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *satisfies the* adjustment criterion *relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$ *if*

(i) $\mathbf{Z} \cap \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \emptyset$ *and*

(ii) all proper non-causal paths from $\mathbf{X}$ *to* $\mathbf{Y}$ *in* $\mathcal{D}$ *are blocked given* $\mathbf{Z}$*.*

Proposition 14 (Shpitser et al., 2010; Perković et al., 2018)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$*, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *satisfies the adjustment criterion relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$ *if and only if* $\mathbf{Z}$ *is a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$*.*

Although the sufficient and necessary adjustment criterion in Definition 13 makes the sufficient, but not necessary back-door criterion in Definition 11 obsolete, the latter remains popular in textbooks (Hernán and Robins, 2020; Lash et al., 2020; Cunningham, 2021).

The causal effect of a singleton treatment X on a singleton outcome Y in a DAG $\mathcal{D}$ is always identified: If $Y \in \text{pa}(X, \mathcal{D})$, then X has not causal effect on Y due to the acyclicity of DAGs; if $Y \notin \text{pa}(X, \mathcal{D})$, then the effect is identified by adjustment for $\text{pa}(X, \mathcal{D})$ (Pearl, 2009, p. 72f.). For sets $\mathbf{X}$ and $\mathbf{Y}$, the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ is always non-parametrically identified in $\mathcal{D}$, but not necessarily identified by adjustment (Perković, 2020). As an example, consider the causal graph in Figure 6 with treatment $\mathbf{X} = \{X_1, X_2\}$ and outcome Y. Here the set $\{V\}$ is not a valid adjustment set relative to $(\mathbf{X}, Y)$, as $V \in \text{forb}(\mathbf{X}, Y, \mathcal{G})$. The empty set is not valid either, as the non-causal path $X_2 \leftarrow V \rightarrow Y$ is not blocked given the empty set. Hence, the effect of $\mathbf{X}$ on Y is not identified by adjustment, but it is identified e.g. by the G-formula for sequential treatments (Robins, 1986; Dawid and Didelez, 2010).

[1] The adjustment criterion was first published by Shpitser et al. (2010) in a slightly different formulation than presented here, and was later revised by the authors in an unpublished addendum that also contains a corrected proof (see Perković et al., 2018). An alternative proof was published in Perković et al. (2018). The formulation presented here follows van der Zander et al. (2014) and is equivalent to the revised formulation.

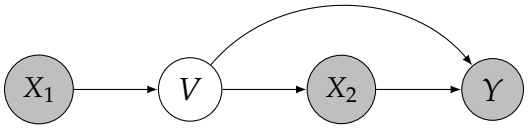


Figure 6: Example causal DAG in which the causal effect of $\{X_1, X_2\}$ on Y is not identified by adjustment. The forbidden nodes are shown as circles, treatment and outcome are additionally highlighted in grey.

3.3 Confounder selection in practice

Tennant et al. (2020) demonstrated in their systematic review that the number of health-related applied publications mentioning DAGs has been increasing dramatically from 26 in total between 1999 and 2012, to 71 in 2017 alone. It is therefore likely that a growing number of authors use the graphical adjustment criteria of Section 3.2 for confounder selection. However, 71 is still small compared to the total number of health-related studies published each year. So how were adjustment variables selected in the remaining studies?

A partial answer for the year 2008 was provided by Walter and Tiemeier (2009). They considered 300 applied studies aiming at causal inference that were published in *American Journal of Epidemiology*, *Epidemiology*, *European Journal of Epidemiology* or *International Journal of Epidemiology*. From each article, they extracted information on how adjustment variables were selected. They found that 105 articles provided no information or at most a vague description (e.g. 'based on prior knowledge'). Among those that described the selection process in more detail, 1 % provided a DAG, 43 % stated the background knowledge on which they relied in words, 30 % reported the use of stepwise regression, 23 % used the change-in-estimate criterion and 4 % other methods (percentages not adding up due to rounding). Talbot and Massamba (2019) presented a similar analysis of articles published in 2015 and noted that data-driven methods remained popular, although their relative share was smaller than in Walter and Tiemeier (2009): 9 % stepwise regression, 18 % change-in-estimate and 14 % univariate regression selection, which had not been assessed in Walter and Tiemeier (2009). In an even more recent literature review, Pressat-Laffouilhère et al. (2021) considered 488 articles published in *New England Journal of Medicine*, *The Lancet*, *Journal of American Medical Association*, *British Medical Journal* or *Annals of Internal Medicine* between 2017 and 2019. The variable selection method was unclear in 234 articles. Among the remaining articles, 4 % reported a DAG, 66 % other background knowledge (including vague hints such as 'based on existing literature', 6 % stepwise regression, 3 % change-in-estimate and 9 % univariate regression selection (possibly among

other methods). The three data-driven methods — stepwise regression, change-in-estimate and univariate regression selection — are explained in more detail in Section 3.4.

Similar results were obtained in systematic reviews focussing on propensity score analyses. Weitzen et al. (2004) considered applied studies published in 2001 that mentioned propensity score-related terms in their title or abstract, or that were published in 2001 and cited at least one of a list of seminal propensity score methods papers. They identified 47 articles, 24 of which provided no information on how the variables in the (initial) propensity score model were selected. Among the remaining articles, 26 % included 'all available variables', 17 % relied on background knowledge, 17 % applied stepwise regression to the propensity score model, 30 % used univariate regression selection, 4 % combined background knowledge and stepwise regression and 4 % selected variables based on the goodness-of-fit of the propensity score model (percentages not adding up due to rounding). Another systematic review identified 296 health-related applied studies published between December 2011 and May 2012 that mentioned propensity score-related terms (Ali et al., 2015). Of these, 194 did not describe how the adjustment variables were selected. Among those that did, 14 % mentioned background knowledge (possibly among other methods), at least one (number not provided) reported the use of stepwise regression, and 56 % used univariate regression selection (possibly among other methods). Both review articles also investigated the use of balance checking; for a recent review regarding this topic, see Granger et al. (2020).

I will ignore the worryingly large numbers of publications failing to describe how adjustment variables were selected. Instead, my aim is to provide a graphical perspective on the strategies mentioned in the literature reviews above. Assuming that an underlying DAG exists, can each of those strategies succeed in selecting a valid adjustment set? If yes, under what causal assumptions? Answers to these questions are given in the next section.

3.4 Non-graphical confounder selection from a graphical point of view

Throughout this section, I assume that the causal system under study can be represented by an unknown causal DAG $\mathcal{D} = (\mathbf{V}, \mathbf{E})$, where potentially only a subset of the variables in $\mathbf{V}$ have been measured. The measured variables include the treatment $X \in \mathbf{V}$, the outcome $Y \in \mathbf{V} \setminus \{X\}$ and a set of covariates $\mathbf{W} \subseteq \mathbf{V} \setminus \{X, Y\}$.

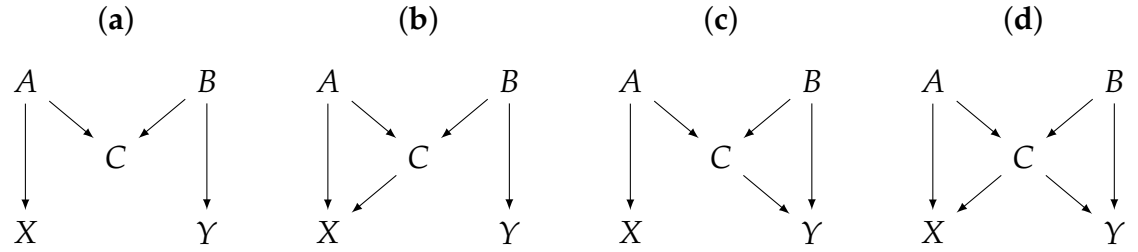


Figure 7: (**a**) M-graph; (**b**,**c**) one-winged butterfly graphs; (**d**) butterfly graph.

I consider seven (groups of) strategies for confounder selection. Among them are the popular data-driven methods identified in Section 3.3, i.e. change-in-estimate, univariate regression selection and stepwise regression. The reviews further identified knowledge-based selection as a common strategy, but no details were given. I thus consider three possible knowledge-based rules for confounder selection, namely the pre-treatment criterion, the common cause criterion and the disjunctive cause criterion (VanderWeele and Shpitser, 2011). Lastly, I include the CovSel algorithm (de Luna et al., 2011) as a less well-known data-driven approach. For each strategy, I state a set of assumptions under which the selected set $\mathbf{W}^*$ is a valid adjustment set, or give examples of situations in which $\mathbf{W}^*$ is not valid.

All considered methods considered have in common that they always output a selected adjustment set (which may be the empty set), but do not attempt to detect situations in which no valid adjustment set exists among the variables to choose from (but see Entner et al., 2013, for an alternative approach). It is therefore clear that a minimal assumption for the validity of each of these procedures must be that the set $\mathbf{W}$ of covariates to choose from contains a valid adjustment set. In fact, most of the considered strategies require the stronger assumption that $\mathbf{W}$ is itself a valid adjustment set. The simplest example where $\mathbf{W}$ contains a valid adjustment set, but is not valid itself, is $\mathbf{W} = \{C\}$ in the so-called 'M-graph' in Figure 7(**a**). Here the empty set is valid, but adjusting for the collider C opens the non-causal path. The bias thus introduced is sometimes called 'M-bias' (Greenland, 2003). The M-graph will be used as a recurring example to demonstrate how different strategies fail when the input set $\mathbf{W}$ is not a valid adjustment set.

3.4.1 Pre-treatment criterion

It is sometimes argued that in the absence of detailed causal knowledge, the safest strategy is to adjust for all measured pre-treatment covariates, i.e. all measured covariates that are non-descendants of the treatment in the graph. This argument was put forward, for example, by Rosenbaum (2002) and Rubin (2009). Formally,

the 'pre-treatment criterion' can be expressed as follows:

Procedure 1 (Pre-treatment criterion)

Input:	knowledge about which nodes in the set $\mathbf{W}$ of observed covariates are descendants of the treatment X in the causal DAG $\mathcal{D}$
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \in \text{nde}(X, \mathcal{D})$.
Output:	selected covariates $\mathbf{W}^*$

The incorrect intuition behind the pre-treatment criterion is that if the set of pre-treatment covariates contains a valid adjustment set, then by adjusting for all covariates in the set, no relevant covariate can be missed. As has been pointed out by numerous authors, this intuition is flawed (e.g. Greenland et al., 1999; Shrier, 2008; Pearl, 2009; Sjölander, 2009; Elwert and Winship, 2014). Consider the M-graph in Figure 7(**a**) and assume that only X, Y and C have been measured. Then the set of measured covariates contains a valid adjustment set — the empty set — but the set of measured pre-treatment covariates — $\{C\}$ — is not valid. The next proposition shows that if the starting set is a valid adjustment set, the pre-treatment criterion will succeed in selecting a valid subset.

Proposition 15
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ with node set $\mathbf{V}$. Let $\mathbf{W}$ be a subset of $\mathbf{V} \setminus \{X, Y\}$ such that $\mathbf{W}$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. If $\mathbf{W}$ is subjected to Procedure 1, then the output $\mathbf{W}^*$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$.

Proof. By construction, $\mathbf{W}^* \subseteq \text{nde}(X, \mathcal{D})$, hence $\mathbf{W}^*$ satisfies condition (i) of the adjustment criterion (Definition 13) relative to (X, Y) in $\mathcal{D}$. In order to see that $\mathbf{W}^*$ satisfies condition (ii) as well, pick a non-causal path p from X to Y in $\mathcal{D}$. As p is blocked given $\mathbf{W}$, it contains (1) a collider A such that $\text{de}(A, \mathcal{D}) \cap \mathbf{W} = \emptyset$, or (2) a non-collider B such that $B \in \mathbf{W}$. In case (1), since $\mathbf{W}^* \subseteq \mathbf{W}$, $\text{de}(A, \mathcal{D}) \cap \mathbf{W} = \emptyset$ implies $\text{de}(A, \mathcal{D}) \cap \mathbf{W}^* = \emptyset$, hence p is blocked given $\mathbf{W}^*$. In case (2), if $B \in \mathbf{W}^*$, then p is blocked given $\mathbf{W}^*$. Hence, for the rest of the proof, consider the case that $B \notin \mathbf{W}^*$, which implies $B \in \text{de}(X, \mathcal{D})$. There are now three cases; I show that p is blocked given $\mathbf{W}^*$ in all of them.

(a) The node B is a collider on p. Then p is blocked given $\mathbf{W}^*$.

(b) The subpath p' of p between X and B contains an edge out of B. The subpath cannot be directed all the way to X, as this would imply a directed cycle. Hence, p' must contain a collider in $\mathrm{de}(B, \mathcal{D}) \subseteq \mathrm{de}(X, \mathcal{D})$, which implies that p is blocked given $\mathbf{W}^*$.

(c) The subpath p'' between B and Y contains an edge out of B. The subpath cannot be directed all the way to Y, as this would imply that $B \in \mathrm{forb}(X, Y, \mathcal{D})$ and hence $B \notin \mathbf{W}$. Hence, p'' must contain a collider in $\mathrm{de}(B, \mathcal{D}) \subseteq \mathrm{de}(X, \mathcal{D})$, which implies that p is blocked given $\mathbf{W}^*$. □

The question whether M-bias is relevant in practice continues to be a topic of debate. Intuitively, the associations corresponding to all four edges in the M-structure need to be sufficiently large in order to induce a non-negligible association between treatment and outcome when conditioning on the collider. Several analytical and empirical studies have been undertaken to quantify M-bias in realistic scenarios (Greenland, 2003; Liu et al., 2012; Ding and Miratrix, 2015; Pearl, 2015; Thoemmes, 2015). While the authors come to different conclusions, it seems fair to say that taken together, their analyses illustrate two points: First, while M-bias tends to be small in practice, there are realistic scenarios in which erroneously conditioning on a pre-treatment collider alters the conclusions drawn from an analysis. Second, the situation becomes considerably more complicated when the underlying structure is not the M-structure shown in panel (**a**) of Figure 7, but the so-called 'butterfly structure' in panel (**d**) or one of the 'one-winged butterflies' in panels (**b**) and (**c**). Neither the empty set nor $\{C\}$ are valid adjustment sets in these two graphs. The results in the papers cited above suggest that the M-bias resulting from adjusting for C is then usually smaller than the confounding bias resulting from not adjusting for C, assuming that A and B are unmeasured and hence not adjusted for. Thus, if there is uncertainty about the exact causal relationship between the pre-treatment covariates and treatment and outcome, adjusting for all pre-treatment covariates appears to be a good option regarding bias reduction.

3.4.2 Common cause criterion

The aim of confounder selection is often informally described as identifying *common causes* of treatment and outcome (e.g. Glymour et al., 2008; Schomaker et al., 2016, p. 288). Recall that a node V_3 is a common cause of two nodes V_1 and V_2 if there is a directed path from V_3 to V_1 that does not include V_2, and a directed path from V_3 to V_2 that does not include V_1. Based on this definition, the 'common cause criterion' can be formalised as follows (see VanderWeele and Shpitser, 2011,

for a similar definition):

Procedure 2 (Common cause criterion)

Input:	knowledge about which variables in the set $\mathbf{W}$ of observed covariates are ancestors of the treatment X, and which are ancestors of the outcome Y through paths that do not contain X, in the causal DAG $\mathcal{D}$
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \in \text{an}(X, \mathcal{D})$ and there is a directed path from W to Y in $\mathcal{D}$ that does not contain X.
Output:	selected covariates $\mathbf{W}^*$

Note that Procedure 2 does not require that W is an ancestor of X through a path that does not contain Y. This is because it is implicitly assumed that $X \notin \text{de}(Y, \mathcal{D})$. This assumption is made also in the next proposition.

Proposition 16
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ with node set $\mathbf{V}$ such that $X \notin \text{de}(Y, \mathcal{D})$. Let $\mathbf{W}$ be a subset of $\mathbf{V} \setminus \{X, Y\}$ such that $\mathbf{W}$ contains all common causes of X and Y in $\mathcal{D}$. If $\mathbf{W}$ is subjected to Procedure 2, then the output $\mathbf{W}^*$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$.

Proof. By construction, $\mathbf{W}^*$ is the set of common causes of X and Y in $\mathcal{D}$. As $(\text{an}(X, \mathcal{D}) \setminus \{X\}) \cap \text{forb}(X, Y, \mathcal{D}) = \emptyset$, $\mathbf{W}^*$ satisfies condition (i) of the adjustment criterion (Definition 13). In order to see that $\mathbf{W}^*$ satisfies condition (ii) as well, pick a non-causal path p from X to Y in $\mathcal{D}$. There are two cases:

Consider first the case that p does not contain a collider. Then p must be of the form $X \leftarrow \cdots \leftarrow A \rightarrow \cdots \rightarrow Y$, which implies $A \in \mathbf{W}^*$. Hence, p is blocked given $\mathbf{W}^*$ in this case.

Consider now the case that p contains a collider. Denote the collider closest to Y on p by C and denote the subpath of p between C and Y by p'. If p' is a directed path from Y to C, then neither C nor any descendant of C is in $\mathbf{W}^*$, as a node cannot be in $\text{de}(X, \mathcal{D})$ and $\text{an}(X, \mathcal{D})$ at the same time. Hence, p is blocked given $\mathbf{W}^*$. The only other possibility is that p' is of the form $C \leftarrow \cdots \leftarrow B \rightarrow \cdots \rightarrow Y$, where B is a common cause of C and Y. If $B \in \text{an}(X, \mathcal{D})$, it follows immediately that $B \in \mathbf{W}^*$

and p is blocked given $\mathbf{W}^*$. If $B \notin \text{an}(X, \mathcal{D})$, then no node in $\text{de}(B, \mathcal{D})$ can be in $\text{an}(X, \mathcal{D})$ either, hence $\text{de}(C, \mathcal{D}) \cap \mathbf{W}^* = \emptyset$ and p is blocked given $\mathbf{W}^*$. □

However, the common cause criterion does not necessarily select a valid adjustment set when some common causes of X and Y are not observed. Consider the graph $X \leftarrow A \rightarrow B \rightarrow Y$ and assume that $\mathbf{W} = \{B\}$. Then $\{B\}$ is a valid adjustment set, but not a common cause of X and Y. Hence, the set $\mathbf{W}^*$ selected by the common cause criterion is the empty set, which is not a valid adjustment set in this example. VanderWeele and Shpitser (2011) provided further examples where the common cause criterion fails to select a valid adjustment set.

In conclusion, thinking about common causes of treatment and outcome, irrespective of the available data, can be a useful first step in the confounder selection process. However, the common cause criterion does not provide guidance when some common causes are unmeasured or even unmeasurable. Further, a valid adjustment set does not necessarily include one or more common causes.

3.4.3 Disjunctive cause criterion

The 'disjunctive cause criterion' was suggested by VanderWeele and Shpitser (2011) with the explicit goal of overcoming the shortcomings of the pre-treatment criterion and the common cause criterion. It selects a valid adjustment set under the relatively mild assumption that the set of covariates to select from *contains* a valid adjustment set (without necessarily being valid itself) and does not contain post-treatment covariates.

Procedure 3 (Disjunctive cause criterion)

Input:	knowledge about which of the variables in the set $\mathbf{W}$ of observed covariates are ancestors of the treatment X or the outcome Y in the causal DAG $\mathcal{D}$
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \in \text{an}(X, \mathcal{D})$ or $W \in \text{an}(Y, \mathcal{D})$ or both.
Output:	selected covariates $\mathbf{W}^*$

Proposition 17

Let X and Y be two nodes in a causal DAG $\mathcal{D}$ with node set $\mathbf{V}$. Let $\mathbf{W}$ be a subset of

$\mathbf{V} \setminus \{X, Y\}$ *such that (i) all nodes in* $\mathbf{W}$ *are non-descendants of* X *and (ii)* $\mathbf{W}$ *contains a valid adjustment set relative to* (X, Y) *in* $\mathcal{D}$*. If* $\mathbf{W}$ *is subjected to Procedure 3, then the output* $\mathbf{W}^*$ *is a valid adjustment set relative to* (X, Y) *in* $\mathcal{D}$*.*

Proof. See the Online Appendix to VanderWeele and Shpitser (2011). □

The proof is not reprinted here as it is somewhat tedious. The main challenge is to show that all non-causal paths are blocked given $\mathbf{W}^*$ despite the fact that some of the paths may contain colliders that are in $\mathbf{W}^*$.

Compared to e.g. the common cause criterion, the disjunctive cause criterion tends to select a much larger number of variables into the adjustment set, including so-called *instrumental variables* or *instruments*, i.e. variables that are causes of treatment, but unrelated to the outcome other than through treatment. VanderWeele and Shpitser (2011) acknowledged that the inclusion of instruments can increase the variance of common estimators, and can amplify bias due to residual confounding (see e.g. Bhattacharya and Vogt, 2007; Myers et al., 2011; Pearl, 2011; Ding et al., 2017). As a solution, they proposed to combine the disjunctive cause criterion with data-driven backward or forward selection, as discussed below in Section 3.4.5.

More recently, VanderWeele (2019) proposed the 'modified disjunctive cause criterion': Select an adjustment set using the disjunctive cause criterion first, then exclude covariates known to be instrumental variables, and add proxies, if available, for unmeasured covariates that are common causes of treatment and outcome. The aim of the latter heuristic is to reduce bias due to residual confounding.

3.4.4 Univariate regression selection

The review papers cited in the introduction to this chapter identified univariate regression selection as a popular method for (pre-)selecting adjustment variables. Univariate regression selection involves testing for marginal independencies between each covariate and the outcome Y and/or the treatment X in the available data. For example, a commonly used variant is to regress Y on each covariate in turn, and select into the adjustment set all covariates whose coefficients turn out to be significantly different from zero (Sun et al., 1996). Alternatively, the conditional independence between each covariate and Y given X is tested by regressing Y on the covariate and X. (Strictly speaking, this is a multivariate regression procedure, but will be covered in this section due to its similarity to the genuinely univariate

procedures.) In the context of confounder selection for a propensity score model, univariate selection can also mean that X, not Y, is regressed on each covariate. Moreover, outcome regression and treatment regression are sometimes combined, and a covariate is selected for adjustment if its coefficient reaches statistical significance (or meets some other criterion) in both regressions, or, alternatively, in at least one of them (Miettinen and Cook, 1981).

Univariate regression selection is most commonly implemented within a linear or logistic model framework, but in principle any (conditional) independence test can be used. In the following, I formalise the above procedures without specifying how the independence information is obtained. One further procedure is added that corresponds to an often stated 'traditional definition of a confounder': According to that definition, a confounder is a cause of the outcome that is associated with the treatment and does not lie on the causal path from the treatment to the outcome (Morabia, 2011). Although known to be misleading (VanderWeele and Shpitser, 2013), the 'traditional definition' is popular in applied epidemiology.

Procedure 4A (Univariate outcome regression selection)

Input:	list of conditional independencies given treatment X between outcome Y and each variable in the set $\mathbf{W}$ of observed covariates
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \not\perp\!\!\!\perp Y \mid X$.
Output:	selected covariates $\mathbf{W}^*$

Procedure 4B (Univariate treatment regression selection)

Input:	list of marginal independencies between treatment X and each variable in the set $\mathbf{W}$ of observed covariates
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \not\perp\!\!\!\perp X$.
Output:	selected covariates $\mathbf{W}^*$

Procedure 4C (Univariate outcome AND treatment regression selection)

Input:	list of conditional independencies given treatment X between outcome Y and each variable in the set $\mathbf{W}$ of observed covariates, and list of marginal independencies between X and each variable in $\mathbf{W}$
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \not\perp\!\!\!\perp Y \mid X$ and $W \not\perp\!\!\!\perp X$.
Output:	selected covariates $\mathbf{W}^*$

Procedure 4D (Univariate outcome OR treatment regression selection)

Input:	list of conditional independencies given treatment X between outcome Y and each variable in the set $\mathbf{W}$ of observed covariates, and list of marginal independencies between X and each variable in $\mathbf{W}$
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \not\perp\!\!\!\perp Y \mid X$ or $W \not\perp\!\!\!\perp X$ or both.
Output:	selected covariates $\mathbf{W}^*$

Procedure 4E (Traditional confounder selection)

Input:	knowledge about which of the variables in the set $\mathbf{W}$ of observed covariates are descendants of the treatment X or ancestors of the outcome Y in the DAG $\mathcal{D}$, and a list of marginal independencies between X and each variable in $\mathbf{W}$
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \in \mathrm{an}(Y, \mathcal{D})$, $W \in \mathrm{nde}(X, \mathcal{D})$ and $W \not\perp\!\!\!\perp X$.
Output:	selected covariates $\mathbf{W}^*$

Variants that consider the marginal dependence $Y \not\perp\!\!\!\perp W$ instead of the conditional dependence $Y \not\perp\!\!\!\perp W \mid X$ could be formulated for Procedures 4A, 4C, 4D and 4E, but are omitted here, as they do not exhibit any additional interesting behaviours.

The next examples demonstrate that none of the five variants of univariate regression selection has the potential to select a valid adjustment set under reasonably

general assumptions that allow for some covariates in the underlying causal DAG $\mathcal{D}$ to be unobserved. In all examples, assume that the joint distribution of $\mathbf{V}$ is faithful to $\mathcal{D}$.

Consider first the M-graph in Figure 7(**a**) and suppose that the set of measured covariates is $\mathbf{W} = \{A, C\}$, which is a valid adjustment set relative to (X, Y). Then Procedures 4A and 4C select $\mathbf{W}^* = \{C\}$, which is not valid. Similarly, if $\mathbf{W} = \{B, C\}$, then Procedures 4B and 4C select $\mathbf{W}^* = \{C\}$. If $\mathbf{W} = \{A, B, C\}$, which is valid, then Procedures 4A and 4B select the valid subsets $\mathbf{W}^* = \{B, C\}$ and $\mathbf{W}^* = \{A, C\}$, respectively, but Procedure 4C still selects $\mathbf{W}^* = \{C\}$. Note that the high-dimensional propensity score method (Schneeweiss et al., 2009; Wyss et al., 2018) is based on the same principle as Procedure 4C, and may thus be susceptible to the same kind of problematic behaviour.

Consider next the causal DAG $X \leftarrow A \rightarrow B \leftarrow C \rightarrow D \leftarrow E \rightarrow Y$ and suppose that the set of measured covariates is $\mathbf{W} = \{B, C, D\}$, which is a valid adjustment set relative to (X, Y). Then Procedure 4D selects $\mathbf{W}^* = \{B, D\}$, which is not valid. This example in particular shows that it is not advisable in general to exclude covariates from the adjustment set based solely on their lack of association with treatment and/or outcome. Even the exclusion of a covariate that is marginally independent of both X and Y, such as C in this example, can invalidate an adjustment set.

Finally, consider the causal DAG $X \leftarrow A \leftarrow B \rightarrow Y$ and suppose that $\mathbf{W} = \{A\}$, which is a valid adjustment set relative to (X, Y). Here Procedure 4E selects the empty set, which is not valid.

In summary, from a theoretical perspective, univariate regression selection has little to recommend itself. For all Procedures 4A to 4E, there exist situations in which the procedure fails to select a valid adjustment set even if the set to select from is itself a valid adjustment set. Procedure 4C even fails in a situation where all variables in the graph are observed. In practice, additional problems are false test decisions due to an insufficient sample size, and misspecification of the function form of continuous covariates. A major advantage, on the other hand, is that univariate regression selection considers one covariate at a time. This can be relevant especially when the number of covariates is larger than the sample size.

3.4.5 Stepwise regression

Stepwise procedures are among the most popular and most widely implemented methods for variable selection in regression models (Heinze et al., 2018). The

review papers cited in the introduction to this chapter confirmed that stepwise procedures are popular also for causal inference. They can be used either for outcome model selection, or for treatment model selection when the propensity score is estimated. Two common variants of stepwise regression selection are backward selection and forward selection, both of which are discussed in this section.

Stepwise regression procedures are usually implemented within a parametric framework, although non-parametric variants have been proposed (Li et al., 2005). Variables are selected into or out of the adjustment set based on the p-values of their estimated coefficients, or based on information criteria such as Akaike's Information Criterion (AIC; Akaike, 1974) or the Bayesian Information Criterion (BIC; Schwarz, 1978). It has been shown that selection of a linear model using AIC or BIC can be re-formulated as p-value selection with AIC- or BIC-specific significance levels (e.g. Murtaugh, 2014; Derryberry et al., 2018). Hence, all three strategies are essentially based on conditional independence testing.

In the following, the selection procedures are formalised in terms of conditional independencies, without specifying how the information is obtained. For the proofs, I assume that the information is correct ('independence oracle'). In practice, of course, choices need to be made regarding functional forms and interactions to include. Further practical issues with stepwise regression include instable selection decisions due to wrong test decisions, under-selection of covariates with small true coefficients, and under-coverage of naive confidence intervals.

I first consider outcome model selection.

Procedure 5A (Backward regression selection, outcome model)

Input:	list of conditional independencies in the joint distribution of treatment X, outcome Y and the set $\mathbf{W}$ of observed covariates
Procedure:	Define $\mathbf{W}' = \mathbf{W}$. Consider all variables in $\mathbf{W}'$ in turn. If for $W \in \mathbf{W}'$, $W \perp\!\!\!\perp Y \mid (\mathbf{W}' \setminus \{W\}) \cup \{X\}$, then update $\mathbf{W}'$ to $\mathbf{W}' \setminus \{W\}$ before proceeding.
Output:	selected covariates $\mathbf{W}^* = \mathbf{W}'$

Proposition 18
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ and let $\mathbf{W}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Then the output $\mathbf{W}^*$ of Procedure 5A with input $(X, Y, \mathbf{W})$ and oracle independence information is a valid adjustment set relative to (X, Y).

Proof. If $\mathbf{W} = \mathbf{W}^*$, the proposition follows trivially. Consider therefore the case where at least one variable is removed during Procedure 5A. Denote the first variable being removed as R and define $\mathbf{W}' = \mathbf{W} \setminus \{R\}$. By Definition 10 of a valid adjustment set, for any density f compatible with $\mathcal{D}$, $\mathbf{W}$ satisfies

$$f(y \mid do(x)) = \int_{\mathbf{w}} f(y \mid x, \mathbf{w}) f(\mathbf{w}) d\mathbf{w}.$$

As $R \perp\!\!\!\perp Y \mid \{X\} \cup \mathbf{W}'$,

$$f(y \mid do(x)) = \int_{\mathbf{w}'} f(y \mid \mathbf{x}, \mathbf{w}') f(\mathbf{w}') d\mathbf{w}',$$

hence $\mathbf{W}'$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. By induction, it follows that $\mathbf{W}^*$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. □

Similar proofs using the potential outcome notation are given in VanderWeele and Shpitser (2011) and in de Luna et al. (2011). Note that the proof does not actually require the distribution of $\{X, Y\} \cup \mathbf{W}$, possibly together with unmeasured variables, to be compatible with a DAG; it suffices to assume that the induced independence model is a semi-graphoid (see Definition 1). The graphical view will be maintained, however, for the purpose of consistency within this chapter.

The next lemma shows that the set selected by Procedure 5A is unique, i.e. that the order in which the covariates are considered does not matter. Here the proof requires that the induced independence model is a graphoid, which is guaranteed if there is an underlying DAG, but also under the weaker assumption that the joint density of $\{X, Y\} \cup \mathbf{W}$ is strictly positive (see Section 2.2).

Lemma 19
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ and let $\mathbf{W}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Then the output $\mathbf{W}^*$ of Procedure 5A with input $(X, Y, \mathbf{W})$ and oracle independence information is the same regardless of the order in which the variables in $\mathbf{W}$ are considered during the procedure.

Proof. I show that a variable $R_1 \in \mathbf{W}$ is removed in the course of Procedure 5A

if and only if $R_1 \perp\!\!\!\perp Y \mid (\mathbf{W}' \setminus \{R_1\}) \cup \{X\}$ at some point during the procedure, where $\mathbf{W}'$ is the current reduced set of covariates.

One direction follows immediately from the definition of Procedure 5A: If variable R_1 is removed, then $R_1 \perp\!\!\!\perp Y \mid (\mathbf{W}' \setminus \{R_1\}) \cup \{X\}$ at the point of its removal.

Thus, assume that $R_1 \perp\!\!\!\perp Y \mid (\mathbf{W}' \setminus \{R_1\}) \cup \{X\}$, but that the variable that is removed in the current step is $R_2 \neq R_1$. Then, by the definition of Procedure 5A, $R_2 \perp\!\!\!\perp Y \mid (\mathbf{W}' \setminus \{R_2\}) \cup \{X\}$. By the property of 'intersection' (see Definition 1), $\{R_1, R_2\} \perp\!\!\!\perp Y \mid (\mathbf{W}' \setminus \{R_1, R_2\}) \cup \{X\}$, and by 'decomposition' (see Definition 1), $R_1 \perp\!\!\!\perp Y \mid (\mathbf{W}' \setminus \{R_1, R_2\}) \cup \{X\}$. It follows by induction that R_1 will be removed at a later stage of the procedure. □

Lemma 19 suggests that it may be possible to select the output $\mathbf{W}^*$ of Procedure 5A in one step without iteratively updating the adjustment set. Indeed, given oracle independence information, the following Procedure 5A′ selects the same set as Procedure 5A if an independence oracle is available. The proof requires again that the independence model is a graphoid.

Procedure 5A′ (Backward regression selection, outcome model)

Input:	list of conditional independencies in the joint distribution of treatment X, outcome Y and the set $\mathbf{W}$ of observed covariates
Procedure:	Consider all variables in $\mathbf{W}$ in turn. Include variable $W \in \mathbf{W}$ in $\mathbf{W}^*$ if $W \not\perp\!\!\!\perp Y \mid (\mathbf{W} \setminus \{W\}) \cup \{X\}$.
Output:	selected covariates $\mathbf{W}^*$

Proposition 20
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ and let $\mathbf{W}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Denote the output of Procedure 5A with input $(X, Y, \mathbf{W})$ and oracle independence information by $\mathbf{W}^*$, and the output of Procedure 5A′ with the same input by $\mathbf{W}'$. Then $\mathbf{W}^* = \mathbf{W}'$.

Proof. Lemma 19 implies that a variable $W \in \mathbf{W}$ is removed during Procedure 5A if and only if $W \perp\!\!\!\perp Y \mid (\mathbf{W} \setminus \{W\}) \cup \{X\}$. Hence, the remaining variables are exactly those not removed in Procedure 5A′. □

Thus, the adjustment set selected by backward regression selection using oracle independence information can alternatively be characterised as the set of all covariates that are independent of Y given X and all other covariates. In a finite data setting, Procedure 5A' could be implemented by regressing Y on X and all covariates, and removing all covariates with estimated coefficients close to zero. Whether this has advantages over conventional backward regression, e.g. in terms of selection stability or coverage of naive confidence intervals, is an interesting question that goes beyond the scope of this thesis.

Proposition 20 further implies that in a causal DAG with node set $\{X, Y\} \cup \mathbf{W}$ where the joint distribution of $\{X, Y\} \cup \mathbf{W}$ is faithful to the DAG and $\mathbf{W}$ is a valid adjustment set, the set $\mathbf{W}^*$ selected by backward regression selection using oracle independence information is the set of parents of Y in the graph. For causal DAGs including unobserved nodes, it is not obvious how the selected set could be characterised graphically. This if further investigated in Chapter 4.

Yet another way of characterising the output of backward regression selection is to say that the selected set $\mathbf{W}^*$ is the subset of $\mathbf{W}$ with the smallest cardinality such that $\mathbf{W} \setminus \mathbf{W}^* \perp\!\!\!\perp Y \mid \mathbf{W}^* \cup \{X\}$. This property is proven next, again relying on the graphoid properties. It will connect backward selection to the CovSel method in Section 3.4.6.

Procedure 5A" (Backward regression selection, outcome model)

Input:	list of conditional independencies in the joint distribution of treatment X, outcome Y and the set $\mathbf{W}$ of observed covariates
Procedure:	Select $\mathbf{W}^* \subseteq \mathbf{W}$ such that $\mathbf{W} \setminus \mathbf{W}^* \perp\!\!\!\perp Y \mid \mathbf{W}^* \cup \{X\}$ and there is no proper subset $\mathbf{U} \subset \mathbf{W}^*$ such that $\mathbf{W} \setminus \mathbf{U} \perp\!\!\!\perp Y \mid \mathbf{U} \cup \{X\}$.
Output:	selected covariates $\mathbf{W}^*$

Proposition 21
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ and let $\mathbf{W}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Denote the output of Procedure 5A with input $(X, Y, \mathbf{W})$ and oracle independence information by $\mathbf{W}^*$, and the output of Procedure 5A" with the same input by $\mathbf{W}''$. Then $\mathbf{W}^* = \mathbf{W}''$.

Proof. Assume for contradiction that $\mathbf{W}^* \neq \mathbf{W}''$. By Proposition 20, for all $R \in \mathbf{W} \setminus \mathbf{W}^*$, $R \perp\!\!\!\perp Y \mid (\mathbf{W} \setminus \{R\}) \cup \{X\}$, and by the 'composition' property (see Defini-

tion 1), $\mathbf{W} \setminus \mathbf{W}^* \perp\!\!\!\perp Y \mid \mathbf{W}^* \cup \{X\}$. It follows that $\mathbf{W}^*$ cannot be a proper subset of $\mathbf{W}''$, as $\mathbf{W}''$ is chosen by Procedure 5A'' such that for all proper subsets $\mathbf{U} \subset \mathbf{W}''$, $\mathbf{W} \setminus \mathbf{U} \not\perp\!\!\!\perp Y \mid \mathbf{U} \cup \{X\}$. Hence, there exists a variable W such that $W \in \mathbf{W}^*$, but $W \notin \mathbf{W}''$. By construction, $\mathbf{W} \setminus \mathbf{W}'' \perp\!\!\!\perp Y \mid \mathbf{W}'' \cup \{X\}$. By the 'weak union' property (see Definition 1), $W \perp\!\!\!\perp Y \mid (\mathbf{W} \setminus \{W\}) \cup \{X\}$. But then by Proposition 20, $W \notin \mathbf{W}^*$, which is a contradiction. □

To sum up, backward regression selection on the outcome model reduces a valid adjustment set to a smaller unique valid adjustment set under assumptions that are weaker than requiring the joint distribution of the variables to be compatible with a DAG. If the starting set $\mathbf{W}$ contains a valid adjustment set but is not itself valid, the selected set is not in general valid, as the example with the M-graph in Figure 7(**a**) and $\mathbf{W} = \{C\}$ shows: Here the empty set is a valid adjustment set, but, under the assumptions of Proposition 18 plus faithfulness, Procedure 5A selects $\{C\}$.

Next, forward selection on the outcome model is considered.

Procedure 5B (Forward regression selection, outcome model)

Input:	list of conditional independencies in the joint distribution of treatment X, outcome Y and the set $\mathbf{W}$ of observed covariates
Procedure:	Define $\mathbf{W}' = \emptyset$. Consider all variables in $\mathbf{W} \setminus \mathbf{W}'$ in turn. If for $W \in \mathbf{W} \setminus \mathbf{W}'$, $W \not\perp\!\!\!\perp Y \mid \mathbf{W}' \cup \{X\}$, then update $\mathbf{W}'$ to $\mathbf{W}' \cup \{W\}$ before proceeding.
Output:	selected covariates $\mathbf{W}^* = \mathbf{W}'$

Proposition 22
Let $\mathcal{D}$ be a causal DAG with node set $\mathbf{V}$ such that the joint distribution of $\mathbf{V}$ is faithful to $\mathcal{D}$. Let X and Y be two nodes in $\mathbf{V}$ and let $\mathbf{W} \subseteq \mathbf{V} \setminus \{X, Y\}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Then the output $\mathbf{W}^*$ of Procedure 5B with input $(X, Y, \mathbf{W})$ and oracle independence information is a valid adjustment set relative to (X, Y).

Proof. Define $\mathbf{R} = \mathbf{W} \setminus \mathbf{W}^*$. By construction, for all $R \in \mathbf{R}$, $R \perp\!\!\!\perp Y \mid \mathbf{W}^* \cup \{X\}$. Faithfulness implies that the independence model induced by the joint distribution of the variables in $\mathbf{V}$ is a compositional graphoid, see Definition 1. It follows from the 'composition' property that $\mathbf{R} \perp\!\!\!\perp Y \mid \mathbf{W}^* \cup \{X\}$.

By Definition 10 of a valid adjustment set, for any density f compatible with $\mathcal{D}$, $\mathbf{W}$ satisfies

$$f(y \mid do(x)) = \int_{\mathbf{w}} f(y \mid x, \mathbf{w}) f(\mathbf{w}) d\mathbf{w}.$$

As $\mathbf{R} \perp\!\!\!\perp Y \mid \mathbf{W}^* \cup \{X\}$,

$$f(y \mid do(x)) = \int_{\mathbf{w}^*} f(y \mid x, \mathbf{w}^*) f(\mathbf{w}^*) d\mathbf{w}^*,$$

hence $\mathbf{W}^*$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. □

Procedure 5B does not select a valid adjustment set in general if the starting set $\mathbf{W}$ contains a valid adjustment set but is not itself valid. This can again be seen from the M-graph example in Figure 7(**a**) with $\mathbf{W} = \{C\}$. Under the assumptions of Proposition 22, Procedure 5B selects $\{C\}$, which is not valid.

The set of covariates selected by Procedure 5B is not in general unique, but depends on the order in which the covariates are considered. As an example, suppose the underlying DAG is $X \leftarrow A \rightarrow B \rightarrow Y$ and all variables are observed. Then, under the assumptions of Proposition 22, Procedure 5B selects $\mathbf{W}^* = \{A, B\}$ if A is considered first, but $\mathbf{W}^* = \{B\}$ if B is considered first. Consequently, Procedures 5A (backward) and 5B (forward) do not in general select the same set of covariates. As shown next, however, the set selected by Procedure 5B is always a superset of the set selected by Procedure 5A under faithfulness.

Lemma 23
Let $\mathcal{D}$ be a causal DAG with node set $\mathbf{V}$ such that the joint distribution of $\mathbf{V}$ is faithful to $\mathcal{D}$. Let X and Y be two nodes in $\mathbf{V}$ and let $\mathbf{W} \subseteq \mathbf{V} \setminus \{X, Y\}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Denote the output of Procedure 5A with input $(X, Y, \mathbf{W})$ and oracle independence information by $\mathbf{W}_A^*$, and the output of Procedure 5B with the same input by $\mathbf{W}_B^*$. Then $\mathbf{W}_A^* \subseteq \mathbf{W}_B^*$.

Proof. I show that all variables *not* selected by Procedure 5B are also *not* selected by Procedure 5A. Define $\mathbf{R} := \mathbf{W} \setminus \mathbf{W}_B^*$. By construction, for all $R \in \mathbf{R}$, $R \perp\!\!\!\perp Y \mid \mathbf{W}_B^* \cup \{X\}$, and by the 'composition' property implied by faithfulness, $\mathbf{R} \perp\!\!\!\perp Y \mid \mathbf{W}_B^* \cup \{X\}$. By the properties of 'decomposition' and 'weak union' (see Definition 1), for all $R \in \mathbf{R}$, $R \perp\!\!\!\perp Y \mid (\mathbf{W} \setminus R) \cup \{X\}$. It follows from Propositions 20 and 18 that R is removed during Procedure 5A and thus $R \notin \mathbf{W}_A^*$. □

Next, I consider backward and forward selection on the treatment model, as form-

alised in Procedures 5C and 5D. Their properties are very similar to those of Procedures 5A and 5B. For completeness, I prove the validity of the selected adjustment sets, but omit analogues of Lemma 19, Procedure 5A', Proposition 20, Procedure 5A'', Proposition 21 and Lemma 23, as they follow from the same arguments as used before.

Procedure 5C (Backward regression selection, treatment model)

Input:	list of conditional independencies in the joint distribution of treatment X and the set $\mathbf{W}$ of observed covariates
Procedure:	Define $\mathbf{W}' = \mathbf{W}$. Consider all variables in $\mathbf{W}'$ in turn. If for $W \in \mathbf{W}'$, $W \perp\!\!\!\perp X \mid \mathbf{W}' \setminus \{W\}$, then update $\mathbf{W}'$ to $\mathbf{W}' \setminus \{W\}$ before proceeding.
Output:	selected covariates $\mathbf{W}^* = \mathbf{W}'$

Proposition 24
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ and let $\mathbf{W}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Then the output $\mathbf{W}^*$ of Procedure 5C with input $(X, Y, \mathbf{W})$ and oracle independence information is a valid adjustment set relative to (X, Y).

Proof. If $\mathbf{W} = \mathbf{W}^*$, the proposition follows trivially. Consider therefore the case where at least one variable is removed during Procedure 5C. Denote the first variable being removed as R and define $\mathbf{W}' = \mathbf{W} \setminus \{R\}$. By Definition 10 of a valid adjustment set, for any density f compatible with $\mathcal{D}$, $\mathbf{W}$ satisfies

$$f(y \mid do(x)) = \int_{\mathbf{w}} f(y \mid x, \mathbf{w}) f(\mathbf{w}) d\mathbf{w} = \int_{\mathbf{w}} \frac{f(y, x, \mathbf{w}', r)}{f(x \mid \mathbf{w}', r)} d\mathbf{w}.$$

As $R \perp\!\!\!\perp X \mid \mathbf{W}'$,

$$f(y \mid do(x)) = \int_{\mathbf{w}} \frac{f(y, x, \mathbf{w}', r)}{f(x \mid \mathbf{w}')} d\mathbf{w} = \int_{\mathbf{w}} f(y, r \mid \mathbf{w}', x) f(\mathbf{w}') d\mathbf{w},$$

and since $\int_{\mathbf{w}} f(r \mid \mathbf{w}', x) f(\mathbf{w}') d\mathbf{w} = \int_{\mathbf{w}} f(r \mid \mathbf{w}') f(\mathbf{w}') d\mathbf{w} = \int_{\mathbf{w}} f(\mathbf{w}) d\mathbf{w} = 1$, we have

$$f(y \mid do(x)) = \int_{\mathbf{w}'} f(y \mid \mathbf{w}', x) f(\mathbf{w}') d\mathbf{w}'.$$

Hence, $\mathbf{W}'$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. By induction, it follows that $\mathbf{W}^*$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. □

Procedure 5D (Forward regression selection, treatment model)

Input:	list of conditional independencies in the joint distribution of treatment X and a set $\mathbf{W}$ of observed covariates
Procedure:	Define $\mathbf{W}' = \emptyset$. Consider all variables in $\mathbf{W} \setminus \mathbf{W}'$ in turn. If for $W \in \mathbf{W} \setminus \mathbf{W}'$, $W \not\perp\!\!\!\perp X \mid \mathbf{W}'$, then update $\mathbf{W}'$ to $\mathbf{W}' \cup \{W\}$ before proceeding.
Output:	selected covariates $\mathbf{W}^* = \mathbf{W}'$

Proposition 25
Let $\mathcal{D}$ be a causal DAG with node set $\mathbf{V}$ such that the joint distribution of $\mathbf{V}$ is faithful to $\mathcal{D}$. Let X and Y be two nodes in $\mathbf{V}$ and let $\mathbf{W} \subseteq \mathbf{V} \setminus \{X, Y\}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Then the output $\mathbf{W}^*$ of Procedure 5D with input $(X, Y, \mathbf{W})$ and oracle independence information is a valid adjustment set relative to (X, Y) in $\mathcal{D}$.

Proof. Define $\mathbf{R} := \mathbf{W} \setminus \mathbf{W}^*$. By construction, for all $R \in \mathbf{R}$, $R \perp\!\!\!\perp X \mid \mathbf{W}^*$. Faithfulness implies that the independence model induced by the joint distribution of the variables in $\mathbf{V}$ is a compositional graphoid. It follows from the 'composition' property in Definition 1 that $\mathbf{R} \perp\!\!\!\perp X \mid \mathbf{W}^*$.

By Definition 10 of a valid adjustment set, for any density f compatible with $\mathcal{D}$, $\mathbf{W}$ satisfies

$$f(y \mid do(x)) = \int_{\mathbf{w}} f(y \mid x, \mathbf{w}) f(\mathbf{w}) d\mathbf{w} = \int_{\mathbf{w}} \frac{f(y, x, \mathbf{w}^*, \mathbf{r})}{f(x \mid \mathbf{w}^*, \mathbf{r})} d\mathbf{w}.$$

As $\mathbf{R} \perp\!\!\!\perp X \mid \mathbf{W}^*$,

$$f(y \mid do(x)) = \int_{\mathbf{w}} \frac{f(y, x, \mathbf{w}^*, \mathbf{r})}{f(x \mid \mathbf{w}^*)} d\mathbf{w} = \int_{\mathbf{w}} f(y, \mathbf{r} \mid \mathbf{w}^*, x) f(\mathbf{w}^*) d\mathbf{w},$$

and since $\int_{\mathbf{w}} f(\mathbf{r} \mid \mathbf{w}^*, x) f(\mathbf{w}^*) d\mathbf{w} = \int_{\mathbf{w}} f(\mathbf{r} \mid \mathbf{w}^*) f(\mathbf{w}^*) d\mathbf{w} = \int_{\mathbf{w}} f(\mathbf{w}) d\mathbf{w} = 1$, we have

$$f(y \mid do(x)) = \int_{\mathbf{w}^*} f(y \mid \mathbf{w}^*, x) f(\mathbf{w}^*) d\mathbf{w}^*.$$

Hence, $\mathbf{W}^*$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. □

3.4.6 CovSel

De Luna et al. (2011) proposed a confounder selection algorithm that I call CovSel, after the `R` package in which the algorithm is implemented (Häggström et al., 2015). Their motivation was to devise a selection procedure that combines well with a matching analysis. Desirable properties in that context are that the method should be non-parametric in nature, and that the selected set of covariates should be small, in order to simplify matching. CovSel is thus based on non-parametric conditional independence testing, and aims at identifying a *minimal valid adjustment set*, i.e. a valid adjustment set such that no proper subset is a valid adjustment set. The algorithm exists in two variants, which are presented here in slightly simplified versions (de Luna et al., 2011, allowed the conditional independence relations to differ between the subgroup where $X = 0$ and the subgroup where $X = 1$). As before, it is not specified in the procedures how the conditional independence information is obtained.

Procedure 6A (CovSel variant A)

Input:	list of conditional independencies in the joint distribution of treatment X, outcome Y and a set $\mathbf{W}$ of observed covariates
Procedure:	(0) Define $\mathbf{W}_0 = \mathbf{W}$. (1) Choose $\mathbf{W}_1 \subseteq \mathbf{W}_0$ such that $\mathbf{W}_0 \setminus \mathbf{W}_1 \perp\!\!\!\perp X \mid \mathbf{W}_1$ and there is no subset $\mathbf{U} \subset \mathbf{W}_1$ such that $\mathbf{W}_0 \setminus \mathbf{U} \perp\!\!\!\perp X \mid \mathbf{U}$. (2) Choose $\mathbf{W}_2 \subseteq \mathbf{W}_1$ such that $\mathbf{W}_1 \setminus \mathbf{W}_2 \perp\!\!\!\perp Y \mid \mathbf{W}_2 \cup \{X\}$ and there is no subset $\mathbf{U}' \subset \mathbf{W}_2$ such that $\mathbf{W}_1 \setminus \mathbf{U}' \perp\!\!\!\perp Y \mid \mathbf{U}' \cup \{X\}$. (3) If $\mathbf{W}_2 \neq \mathbf{W}_1$, then update $\mathbf{W}_0$ to $\mathbf{W}_2$ and repeat steps (1) to (3).
Output:	selected covariates $\mathbf{W}^* = \mathbf{W}_2$

Procedure 6B (CovSel variant B)

Input:	list of conditional independencies in the joint distribution of treatment X, outcome Y and a set $\mathbf{W}$ of observed covariates
Procedure:	(0) Define $\mathbf{W}_0 = \mathbf{W}$. (1) Choose $\mathbf{W}_1 \subseteq \mathbf{W}_0$ such that $\mathbf{W}_0 \setminus \mathbf{W}_1 \perp\!\!\!\perp Y \mid \mathbf{W}_1 \cup \{X\}$ and there is no subset $\mathbf{U} \subset \mathbf{W}_1$ such that $\mathbf{W}_0 \setminus \mathbf{U} \perp\!\!\!\perp Y \mid \mathbf{U} \cup \{X\}$. (2) Choose $\mathbf{W}_2 \subseteq \mathbf{W}_1$ such that $\mathbf{W}_1 \setminus \mathbf{W}_2 \perp\!\!\!\perp X \mid \mathbf{W}_2$ and there is no subset $\mathbf{U}' \subset \mathbf{W}_2$ such that $\mathbf{W}_1 \setminus \mathbf{U}' \perp\!\!\!\perp X \mid \mathbf{U}'$. (3) If $\mathbf{W}_2 \neq \mathbf{W}_1$, then update $\mathbf{W}_0$ to $\mathbf{W}_2$ and repeat steps (1) to (3).
Output:	selected covariates $\mathbf{W}^* = \mathbf{W}_2$

The CovSel algorithms thus iterate between backward outcome selection (Procedure 5A″) and backward treatment model selection (Procedure 5C″). CovSel variant A starts with treatment model selection, while CovSel variant B starts with outcome model selection. Therefore, the selected sets are not in general equal, as demonstrated in the example in Figure 8, where CovSel A selects $\{B\}$ and CovSel B selects $\{C\}$.

It follow immediately from the results in Section 3.4.5 that the output of each CovSel version using oracle independence information is valid and unique (see de Luna et al., 2011, for an alternative proof):

Proposition 26
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ and let $\mathbf{W}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Then the output $\mathbf{W}^*$ of Procedure 6A with input $(X, Y, \mathbf{W})$ and oracle independence information is a valid adjustment set relative to (X, Y).

Proof. The validity follows directly from Proposition 21 and the analogous result for the backward treatment model selection. The uniqueness follows from Proposition 21 together with Lemma 19 and the analogous results for backward treatment model selection. □

Proposition 27
Let X and Y be two nodes in a causal DAG $\mathcal{D}$ and let $\mathbf{W}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Then the output $\mathbf{W}^*$ of Procedure 6B with input $(X, Y, \mathbf{W})$ and oracle independence information is a valid adjustment set relative to (X, Y).

Proof. See proof of Proposition 26. □

In order to show that the adjustment sets selected by CovSel A and CovSel B are minimal valid adjustment sets, de Luna et al. (2011) assumed that the joint distribution of $\mathbf{W} \cup \mathbf{U}$, where $\mathbf{U}$ is a (possibly empty) set of unobserved variables, is faithful to an unknown DAG. The minimality then follows from the 'weak transitivity' property (Definition 2) of DAG-induced independence models. The DAG does not need to be given a causal interpretation in order for the result to hold. In de Luna et al. (2011), a proof sketch is presented using the potential outcome notation. For completeness, I provide a completely graphical proof in Lemma 29, making use of the following result from Tian et al. (1998).

Lemma 28 (Theorem 1 in Tian et al., 1998)
Let X and Y be two nodes in a DAG $\mathcal{D}$ and let $\mathbf{Z}$ be a set of nodes in $\mathcal{D}$ such that $X \perp_{\mathcal{D}} Y \mid \mathbf{Z}$. If there is a set $\mathbf{Z} \cup \{Z_1, \ldots, Z_M\}$ with $M \geq 2$ and $\mathbf{Z} \cap \{Z_1, \ldots, Z_M\} = \emptyset$ such that $X \perp_{\mathcal{D}} Y \mid \mathbf{Z} \cup \{Z_1, \ldots, Z_M\}$, then $X \perp_{\mathcal{D}} Y \mid \mathbf{Z} \cup \{Z_m\}$ holds for at least one $m \in M$.

Lemma 29
Let $\mathcal{D}$ be a causal DAG with node set $\mathbf{V}$ such that the joint distribution of $\mathbf{V}$ is faithful to $\mathcal{D}$. Let X and Y be two nodes in $\mathbf{V}$ and let $\mathbf{W} \subseteq \mathbf{V} \setminus \{X, Y\}$ be a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Let $\mathbf{W}^*$ be the output of Procedure 6A or Procedure 6B with input $(X, Y, \mathbf{W})$ and oracle independence information. Then no proper subset of $\mathbf{W}^*$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$.

Proof. Assume for contradiction that there exists a subset $\mathbf{U} \subset \mathbf{W}^*$ such that $\mathbf{U}$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$, and define $\mathbf{R} := \mathbf{W}^* \setminus \mathbf{U}$. As the variables in $\mathbf{R}$ were not removed during Procedure 6A/6B, it must hold that $X \not\perp\!\!\!\perp \mathbf{R} \mid \mathbf{U}$ and $Y \not\perp\!\!\!\perp \mathbf{R} \mid \mathbf{U} \cup \{X\}$, hence $X \not\perp_{\mathcal{D}} \mathbf{R} \mid \mathbf{U}$ and $Y \not\perp_{\mathcal{D}} \mathbf{R} \mid \mathbf{U} \cup \{X\}$.

Denote by $\mathcal{D}'$ the DAG that can be constructed by removing from $\mathcal{D}$ all edges from X to $\text{forb}(X, Y, \mathcal{D})$. It follows from the adjustment criterion in Definition 13 together with Proposition 14 that if a set $\mathbf{Z}$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$, then $X \perp_{\mathcal{D}'} Y \mid \mathbf{Z}$. Hence, $X \perp_{\mathcal{D}'} Y \mid \mathbf{W}^*$ and $X \perp_{\mathcal{D}'} Y \mid \mathbf{U}$. By Lemma 28, $X \perp_{\mathcal{D}'} Y \mid \mathbf{U} \cup \{R\}$ for at least one $R \in \mathbf{R}$, and by the 'weak transitivity' property (Definition 2), $X \perp_{\mathcal{D}'} R \mid \mathbf{U}$ or $Y \perp_{\mathcal{D}'} R \mid \mathbf{U}$ or both. By induction and the 'contraction' property (Definition 1), it then follows that $X \perp_{\mathcal{D}'} \mathbf{R} \mid \mathbf{U}$ or $Y \perp_{\mathcal{D}'} \mathbf{R} \mid \mathbf{U}$ or both. I show next that either leads to a contradiction.

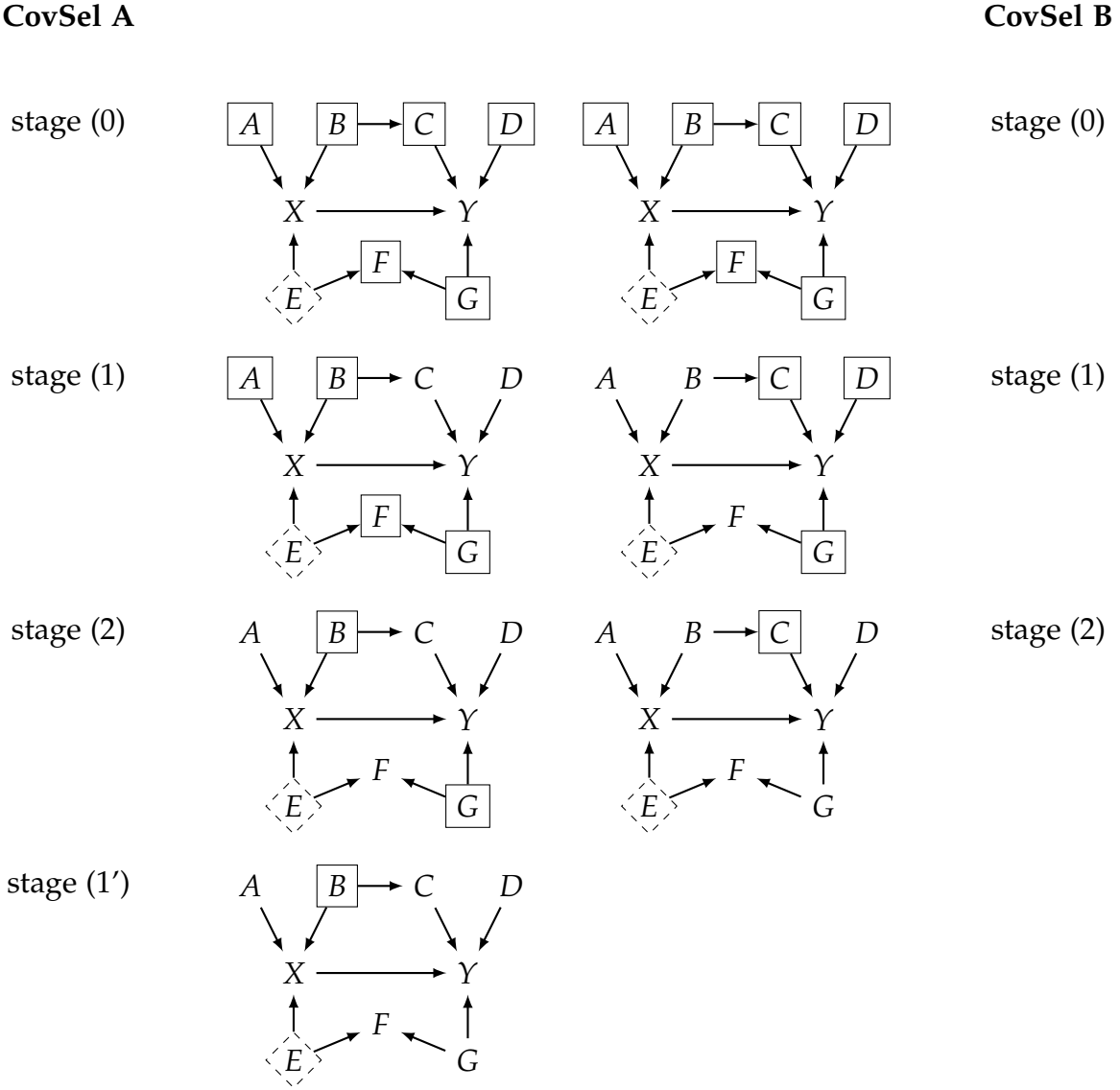


Figure 8: Illustration of CovSel A (Procedure 6A) and CovSel B (Procedure 6B). Node E, shown as a dashed diamond, is unobserved and thus not available for adjustment. The nodes in the rectangles are those selected at the different stages of the algorithms.

Consider first $Y \perp_{\mathcal{D}'} \mathbf{R} \mid \mathbf{U}$. As $Y \not\perp_{\mathcal{D}} \mathbf{R} \mid \mathbf{U} \cup \{X\}$, there exists a path p between Y and $\mathbf{R}$ in $\mathcal{D}$. There are three cases: (i) X is not on p, (ii) X is a non-collider on p, (iii) X is a collider on p. In case (i), p is also in $\mathcal{D}'$. As $Y \perp_{\mathcal{D}'} \mathbf{R} \mid \mathbf{U}$, p is blocked given $\mathbf{U}$ in $\mathcal{D}'$. As X is not on p, p is also blocked given $\mathbf{U} \cup \{X\}$ in both $\mathcal{D}'$ and $\mathcal{D}$. In case (ii), p is blocked given $\mathbf{U} \cup \{X\}$ in $\mathcal{D}$ because X is a non-collider on p. In case (iii), denote the subpath of p between Y and X, which is a non-causal path from X to Y, as p'. As $\mathbf{U}$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$, p' is bocked given $\mathbf{U}$ in $\mathcal{D}$, hence p is blocked given $\mathbf{U} \cup \{X\}$ in $\mathcal{D}$. But then p is blocked given $\mathbf{U} \cup \{X\}$ in $\mathcal{D}$, and the same is true for all other paths between Y and $\mathbf{R}$ in $\mathcal{D}$, which contradicts $Y \not\perp_{\mathcal{D}} \mathbf{R} \mid \mathbf{U} \cup \{X\}$.

Consider now $X \perp_{\mathcal{D}'} \mathbf{R} \mid \mathbf{U}$. As $X \not\perp_{\mathcal{D}} \mathbf{R} \mid \mathbf{U}$, there exists a path p' between X and $\mathbf{R}$ in $\mathcal{D}$. There are three cases: (i) p contains an edge into X, (ii) p contains an edge $X \to T$ such that $T \notin \text{forb}(X, Y, \mathcal{D})$, (iii) p contains an edge $X \to T$ such that $T \in \text{forb}(X, Y, \mathcal{D})$. In cases (i) and (ii), p is also in $\mathcal{D}'$. As $X \perp_{\mathcal{D}'} \mathbf{R} \mid \mathbf{U}$, p is blocked given $\mathbf{U}$ in $\mathcal{D}'$, and hence also in $\mathcal{D}$. In case (iii), p cannot be a directed path from X to $\mathbf{R}$, as $\mathbf{R} \cap \text{forb}(X, Y, \mathcal{D}) = \emptyset$ due to $\mathbf{R}$ being a subset $\mathbf{W}^*$, which is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. Hence, p contains a collider in $\text{forb}(X, Y, \mathcal{D})$. But this means that p is blocked given $\mathbf{U}$ in $\mathcal{D}$, and the same is true for all other paths between X and $\mathbf{R}$ in $\mathcal{D}$, which contradicts $X \perp_{\mathcal{D}} \mathbf{R} \mid \mathbf{U}$. □

3.4.7 Change-in-estimate

Change-in-estimate was among the most frequently mentioned confounder selection methods in the literature reviews cited in Section 3.3. It is a heuristic method and cannot be characterised graphically. The starting point of the procedure is an estimate adjusted for the 'full set' of covariates; covariates are then gradually removed until the estimated effect deviates from the initial estimate by more than a pre-specified amount (usually 10 %). The intuition is that if the starting set is a valid adjustment set, the initial estimate will be close to the true effect. By removing covariates whose removal does not substantially change the estimate, it is hoped that the remaining covariates more accurately represent the relevant confounding factors (Greenland and Pearce, 2015). Change-in-estimate is usually used within a parametric regression framework, although it is not inherently restricted to parametric estimation nor regression. It has been recommended over stepwise regression selection due to its focus on the estimate of interest, and because it does not rely on statistical testing (Sonis, 1998).

However, it is obvious that change-in-estimate does not in general select a valid

adjustment set. Even under the assumption that the initial estimate is equal to the true effect, change-in-estimate has an inbuilt selection bias, as it allows the final estimate to deviate from the initial estimate by a non-negligible amount. The final estimator may have a smaller variance than the initial one, but there is no guarantee. An additional issue of relevance e.g. in logistic regression models is non-collapsibility (Daniel et al., 2021): In naive implementations of change-in-estimate, estimated partial regression coefficients from different models are compared with each other. However, these are estimates of conditional dependencies, and the estimands depend on the conditioning variables. A better alternative is to choose a marginal estimand, e.g. the marginal causal odds ratio, which can be estimated using regression standardisation. A more sophisticated procedure comparing mean squared errors instead of point estimates was proposed by Vansteelandt et al. (2012).

3.5 Paper 1: *Witte and Didelez (2019)*

The paper associated with this chapter is Witte and Didelez (2019). It makes two main contributions. First, a classification system for covariate selection strategies is suggested, covering all methods analysed above as well as additional ones. The second contribution is a simulation study illustrating the properties of the different approaches and their targeted adjustment sets when estimating an average treatment effect, where the focus is on estimation efficiency.

The notation and terminology in Witte and Didelez (2019) mostly agree with those used in this frame text, with the following deviations: In the paper, the treatment is denoted as T, the set of measured covariates as $\mathbf{X}^*$ and the selected adjustment set as $\mathbf{X}$ (instead of X, $\mathbf{W}$ and $\mathbf{W}^*$); valid adjustment sets are called *sufficient* adjustment sets; univariate regression selection is referred to as *univariate screening*; and causal DAGs are called *causal diagrams*. Further, in the paper we use a sum notation where in this thesis I use integrals.

Own contributions

As the first author of this publication, I took the lead in identifying the different confounder selection strategies from the literature and constructed the classification scheme. I designed and programmed the simulation study and created all plots and figures. I wrote the first draft of the manuscript and led the revision process.

Received: 1 December 2017 | Revised: 6 July 2018 | Accepted: 25 July 2018

DOI: 10.1002/bimj.201700294

RESEARCH PAPER

Biometrical Journal

Covariate selection strategies for causal inference: Classification and comparison

Janine Witte[1,2] | Vanessa Didelez[1,2]

[1]Department Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

[2]Faculty 03: Mathematics/Computer Science, University of Bremen, Germany

Correspondence
Vanessa Didelez, Department Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Achterstr. 30, 28359 Bremen, Germany.
Email: didelez@leibniz-bips.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: DI 2372/1-1

Abstract

When causal effects are to be estimated from observational data, we have to adjust for confounding. A central aim of covariate selection for causal inference is therefore to determine a set that is sufficient for confounding adjustment, but other aims such as efficiency or robustness can be important as well. In this paper, we review six general approaches to covariate selection that differ in the targeted type of adjustment set. We discuss and illustrate their advantages and disadvantages using causal diagrams. Moreover, the approaches and different ways of implementing them are compared empirically in an extensive simulation study. We conclude that there are considerable differences between the approaches but none of them is uniformly best, with performance depending on the chosen adjustment method as well as the true confounding structure. Any prior structural knowledge on the causal relations is helpful to choose the most appropriate method.

KEYWORDS

average causal effect, causal diagram, confounder selection, propensity score matching, variable selection

1 | INTRODUCTION

In epidemiology, a common aim is to estimate causal effects from observational data. This typically requires adjustment for confounding to avoid bias. Hence, a central aim of covariate selection for causal inference is to provide a set of covariates sufficient for confounding adjustment. This is in contrast to covariate selection for prediction, where prediction accuracy is of main concern, or for descriptive modeling, which aims at a sparse representation of the association structure (Shmueli, 2010).

In addition to finding a sufficient adjustment set, covariate selection for causal inference can have other aims such as efficiency. In linear regression, for example, the estimates are most precise when adjusted for predictors of the outcome even when these are not necessary for confounding adjustment. For matching, on the other hand, a central requirement is that the adjustment set be small (see de Luna, Waernbaum, & Richardson, 2011, and references therein). A further aim can be robustness against misspecification of the functional form, or more generally non- or semiparametric estimation. Selecting a small set requires fewer functional forms to be determined. Another idea is to combine outcome- and treatment-oriented selection so that a confounder that is missed out in one selection process has a second chance to be selected in the other one (Belloni, Chernozhukov, & Hansen, 2014). This is related to the double-robust property of adjustment methods that combine treatment and outcome modeling (Bang & Robins, 2005). Finally, in situations where the number of covariates is large compared to the number of observations, the mere reduction of the number of covariates may be an aim in itself.

Considering these different situations and the variety of implied adjustment sets, it becomes clear why so many different approaches to covariate selection have been suggested and also why it can be difficult to decide which method is best for a particular problem. In this paper, we offer some orientation by sketching a classification scheme useful for reasoning about and

TABLE 1 Classification of covariate selection strategies for causal inference

	Preadjustment			Wrapper
	Knowledge-based	**Nonparametric**	**Parametric**	
Minimal approach	• DAGitty algorithm (Textor & Liśkiewicz, 2011) • Common cause criterion (cf. Glymour, Weuve, & Chen, 2008)	• Univariate confounder screening (uniTandY) • CovSel (de Luna et al., 2011)		
Outcome approach		• Univariate outcome screening (uniY) • Model-free variable selection (Li, Cook, & Nachtsheim, 2005) • Random forest variable selection, e.g., Genuer, Poggi, and Tuleau-Malot (2010) and Kursa and Rudnicki (2010)	• Outcome-adaptive lasso (Shortreed & Ertefaie, 2017)	• Optimize outcome model, e.g., AIC, BIC, *p*-value method, validation data
Treatment approach		• Univariate treatment screening	• Optimize treatment model, e.g., AIC, BIC, *p*-value method, validation data	
Union set approach	• Disjunctive cause criterion (VanderWeele & Shpitser, 2011)	• Univariate double screening (uniTorY)	• Double selection (Belloni et al., 2014) • Penalized credible regions (Wilson & Reich, 2014)	
Causal search approach		• EHS algorithm (Entner et al., 2013)		
Estimation approach				• Change in estimate (CIE) • Change in MSE (CI-MSE; Greenland, Daniel, & Pearce, 2016) • Focused confounder selection (FCS; Vansteelandt, Bekaert, & Claeskens, 2012)

Note. Methods in the left-most column assume prior knowledge, the EHS algorithm assumes that all measured covariates are pretreatment, all remaining methods assume that the full set of measured covariates is sufficient.

comparing the different properties of covariate selection strategies when used for causal inference. We begin in Section 2 with key assumptions. Importantly, all selection methods rely on assumptions that can only be justified with subject-matter knowledge. Section 3 outlines our classification: we describe six types of target sets and two selection mechanisms, see also Table 1. In the spirit of Boulesteix, Binder, Abrahamowicz, and Sauerbrei (2018), we conduct an extensive simulation study in Section 4 to compare the principles and methods of selection, carefully separating general approaches from specific implementations by considering the following aspects in turn: First, we investigate how each of the target adjustment sets (minimal, outcome-oriented, etc.) performs in combination with typical adjustment methods (regression, matching, etc.). As the target adjustment sets are only known exactly when (most of) the causal structure is known a priori, we evaluate in a second step the performance of standard implementations (univariate selection, change in estimate [CIE], etc.), for different sample sizes, with regard to their precision and ability to select the desired adjustment set. The results, presented in Section 5, illustrate quantitatively the strengths and weaknesses expected based on theoretical properties; in particular they reveal that there is not a uniformly best selection method. Rather, the target set should be determined based on the method used for adjustment. Not surprisingly, the ability of different methods to find the target sets depends, among other things, on the specific causal structure and the sample size. We conclude the paper with a discussion in Section 6.

2 | SUFFICIENT ADJUSTMENT SETS

The underlying principles of most covariate selection approaches are valid for general treatment types, but for simplicity we consider a binary indicator T of treatment ($T = 1$ when treated, $T = 0$ otherwise). Let further Y be the outcome of interest and $\mathbf{X}^*$ the set of measured covariates. We denote by $\mathbf{X}$ a subset of $\mathbf{X}^*$ with realizations $\mathbf{x} \in \mathcal{X}$. Where appropriate, we use subscripts to $\mathbf{X}$ to indicate the selection method by which $\mathbf{X}$ has been selected from $\mathbf{X}^*$, for example, $\mathbf{X}_{\text{EHS}}$ for a set selected by the Entner–Hoyer–Spirtes algorithm described in Section 3.5. We write $\mathcal{P}(\cdot)$ for distribution functions and $P(\cdot)$ for probability and assume that both are defined on one common population of interest throughout the paper.

The central problem of causal inference from observational data is that the distribution of variables we see as passive observers is different from the distribution we would see if we were able to intervene in the treatment. Therefore, we need a terminology that differentiates between observational and interventional regimes. We use in this paper Pearl's do-notation (Pearl, 2009), where distributions are indexed by $do(Z = z)$ to indicate that Z is set to value z by intervention. For example, $\mathcal{P}(Y; do(T = 1))$ describes the distribution of the outcome Y given treatment is enforced for everyone in the population. In contrast, $\mathcal{P}(Y \mid T = 1)$ describes the distribution of Y in the subpopulation that is observed to receive treatment.

Causal treatment effects can be defined as contrasts between (summaries of) $\mathcal{P}(Y; do(T = 1))$ and $\mathcal{P}(Y; do(T = 0))$. A popular estimand is the average causal effect (ACE),

$$\text{ACE} = \text{E}(Y; do(T = 1)) - \text{E}(Y; do(T = 0))$$

(Imbens, 2004; Lunceford & Davidian, 2004; Schafer & Kang, 2008). An alternative estimand for binary Y is the marginal causal odds ratio (MCOR),

$$\text{MCOR} = \frac{P(Y = 1; do(T = 1))/P(Y = 0; do(T = 1))}{P(Y = 1; do(T = 0))/P(Y = 0; do(T = 0))}$$

(Zhang, 2008). Obviously, in observational studies neither $\mathcal{P}(Y; do(T = t))$ nor any summaries thereof are measured. Identification from observational data requires that the effect can be expressed in $do(\cdot)$-free terms. This is possible when a set $\mathbf{X} \subseteq \mathbf{X}^*$ is available satisfying the following three assumptions:

Assumption 1 (**Pretreatment covariates**).

$$\mathcal{P}(\mathbf{X}; do(T = 0)) = \mathcal{P}(\mathbf{X}; do(T = 1)) = \mathcal{P}(\mathbf{X}).$$

Assumption 1 says that the distribution of the covariates $\mathbf{X}$ is not affected by interventions in the treatment. We call this the pretreatment assumption and say that $X \in \mathbf{X}$ is a pretreatment covariate. Assumption 1 automatically holds if X has been measured prior to T, but it suffices to know that X cannot be affected by T.

Assumption 2 (**Conditional exchangeability**).

$$\mathcal{P}(Y \mid \mathbf{X}; do(T = t)) = \mathcal{P}(Y \mid \mathbf{X}, T = t) \quad \text{for } t = 0, 1.$$

Assumption 2 is the key to identifying causal effects from observational data. Intuitively, it guarantees that conditional on covariates $\mathbf{X}$, association is causation. Assumption 2 is also called "no unobserved confounding" (Robins, 1992). Unless additional experimental data are available or the data happen to contain a structure similar to an instrument (de Luna & Johansson, 2014; Entner, Hoyer, & Spirtes, 2013), it is untestable and therefore needs to be justified by subject-matter knowledge.

Assumption 3 (**Positivity**).

$$P(T = t \mid \mathbf{X} = \mathbf{x}) > 0, \quad \text{for } t = 0, 1 \text{ and all } \mathbf{x} \in \mathcal{X}.$$

Assumption 3, positivity, implies that in sufficiently large samples, both treated and untreated individuals are observed for any given value of the covariates.

It is now shown for $T = 1$ how the interventional distribution $\mathcal{P}(Y; do(T = 1))$ can be expressed by observable terms, using Assumptions 1 and 2 in the second equation:

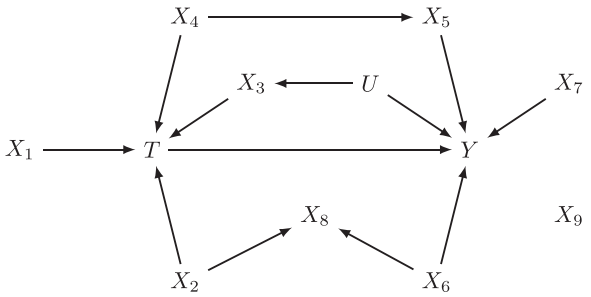


FIGURE 1 Example of causal diagram. T, treatment; Y, outcome; $X_1, \dots, X_9$, observed covariates; U, unobserved covariate. Data generation for setup 1 in Section 4.1 is according to this causal diagram

$$\begin{aligned}
&\mathcal{P}(Y; do(T = 1)) \\
&= \sum_{\mathcal{X}} \mathcal{P}(Y \mid \mathbf{X} = \mathbf{x}; do(T = 1)) P(\mathbf{X} = \mathbf{x}; do(T = 1)) \\
&= \sum_{\mathcal{X}} \mathcal{P}(Y \mid \mathbf{X} = \mathbf{x}, T = 1) P(\mathbf{X} = \mathbf{x}).
\end{aligned}$$

Assumption 3 ensures that $\mathcal{P}(Y \mid \mathbf{X} = \mathbf{x}, T = 1)$ is defined over the whole range of $\mathcal{X}$. The terms in the resulting expression are estimable from observational data. Note that the last equality shows that the intervention distribution can be regarded as a weighted average that illustrates the principle of standardization. This is used in Section 4.3 to estimate the MCOR by standardizing a logistic regression.

We call a set $\mathbf{X} \subseteq \mathbf{X}^*$ satisfying Assumptions 1–3 a *sufficient* adjustment set (Greenland, Pearl, & Robins, 1999). As addressed later, many selection methods assume that $\mathbf{X}^*$ itself is a sufficient adjustment set, and then attempt to reduce this set. We call a sufficient adjustment set *globally minimal* if it has the smallest cardinality among all such sets. We call $\mathbf{X}$ *locally minimal* if it is sufficient and no proper subset of $\mathbf{X}$ is sufficient. Every globally minimal adjustment set is also locally minimal, but not vice versa. Note that none of these sets are necessarily unique.

2.1 | Sufficient adjustment sets and causal diagrams

Causal diagrams are formal but intuitive representations of the underlying causal structure of a problem (Pearl, 2009). Even though it is rare, in practice, that a causal diagram can be fully specified in all detail just based on background knowledge, they are still useful to reason about and illustrate types of adjustment sets. For a short introduction to causal diagrams see Appendix A.1. In the following, we use the terms "ancestor"/"cause" and "parent"/"direct cause" interchangeably. The notion of "direct" cause has to be understood as relative to the variables in the graph, for example, one can often think of unobserved intermediates between variables not shown in the causal diagram. Note that the strong assumptions of a causal diagram can to some extent be relaxed while still allowing identification of sufficient adjustment sets (see Dawid, 2002, for further details).

Assumption 1 (pretreatment covariates) and Assumption 2 (conditional exchangeability) have the following graphical counterparts that can be checked on a given causal diagram (Pearl, 2009):

Assumption 1g (Pretreatment covariates).

Every $X \in \mathbf{X}$ is a nondescendant of T.

Assumption 2g (Backdoor criterion).

All backdoor paths from T to Y are blocked by $\mathbf{X}$.

Thus, $\mathbf{X}$ is a sufficient adjustment set when Assumptions 1–3 or, graphically, Assumptions 1g, 2g, and 3 hold. As an example, consider Figure 1, where $\mathbf{X}^* = \{X_1, \dots, X_9\}$ and U is an unobserved covariate. There are three backdoor paths from T to Y: $T \leftarrow X_4 \rightarrow X_5 \rightarrow Y$, $T \leftarrow X_3 \leftarrow U \rightarrow Y$, and $T \leftarrow X_2 \rightarrow X_8 \leftarrow X_6 \rightarrow Y$. The latter is blocked by the empty set because X_8 is a collider on the path. Examples of sufficient adjustment sets are $\{X_3, X_4\}$, $\{X_1, X_3, X_5, X_9\}$, and $\{X_3, X_5, X_6, X_8\}$. Moreover, the sets $\{X_3, X_4\}$ and $\{X_3, X_5\}$ are both globally and locally minimal.

2.2 | Assumptions underlying covariate selection for causal inference

Assumptions 1–3 define a sufficient adjustment set. We now give assumptions relating such a set and the observed covariates. The majority of selection strategies rely on the following key assumption:

Assumption 4. $\mathbf{X}^*$ itself is a sufficient adjustment set.

In other words, it is assumed that Assumptions 1–3 hold for the full set of measured covariates $\mathbf{X}^*$ (some methods require Assumption 3 only for the actually selected set). Under Assumption 4, selection aims at reducing the number of covariates, for one of the reasons mentioned in Section 1.

A weaker assumption is the following:

Assumption 5.

(i) Every $X \in \mathbf{X}^*$ is a pretreatment covariate.

(ii) $\mathbf{X}^*$ contains a sufficient adjustment set.

Approaches requiring only Assumption 5 but not Assumption 4 are evidently desirable, but we are only aware of one (VanderWeele & Shpitser, 2011). To see the difference between Assumptions 4 and 5, consider the following example: In Figure 1, the set $\{X_1, \dots, X_9\}$ satisfies Assumption 4. However, assume X_2 and X_6 are unobserved, then the set $\{X_1, X_3, X_4, X_5, X_7, X_8, X_9\}$ does not satisfy Assumption 4 (conditioning on X_8 opens a backdoor path) but satisfies Assumption 5 as it contains a sufficient adjustment set.

A further selection strategy, the algorithm proposed by Entner et al. (2013), is sometimes able to infer conditional exchangeability under an even weaker assumption:

Assumption 6.

(i) Every $X \in \mathbf{X}^*$ is a pretreatment covariate.

(ii) Positivity holds for every selected set $\mathbf{X} \subseteq \mathbf{X}^*$.

Note that Assumptions 4–6 are specific to causal inference. When selection is for predictive modeling, confounding is not an issue.

3 | CLASSIFICATION OF COVARIATE SELECTION STRATEGIES

We now describe six different approaches to covariate selection (Sections 3.1–3.6), each corresponding to a different type of target adjustment set. For each approach, there are several proposed methods to implement the approach. They differ mainly in how much prior structural knowledge they assume, and in their validity under different structures. The six approaches correspond to the rows in Table 1. The columns correspond to a second classification criterion, the mechanism of selection, which we describe in Section 3.7.

We illustrate the approaches with the causal diagram of Figure 1, but note that for the majority of selection methods, it is not required that the data-generating mechanism can in fact be represented by a causal diagram. However, even if only incomplete prior knowledge is available, we still recommend to consider a set of plausible diagrams specifically to rule out problematic unobserved quantities, and hence to help justify Assumption 4 or 5. The problem of unobserved covariates is, in fact, a general one but affects the different approaches in different ways; this will be addressed individually below.

3.1 | Minimal approach

3.1.1 | Motivation

Small adjustment sets are advantageous for nonparametric adjustment methods in terms of both bias and variance (de Luna et al., 2011). Especially in the context of matching, small adjustment sets appear desirable as it is then easier to find suitable matches. They are also favorable for regression procedures with continuous covariates because fewer functional forms need to be specified. This first approach can therefore be described as aiming at small, ideally locally or even globally minimal adjustment sets.

3.1.2 | Examples

Given a causal diagram, globally and locally minimal adjustment sets can in principle be read off using the backdoor criterion. For large and complex diagrams, there exist algorithms that list all minimal sets, for example the algorithm given by Textor and Liśkiewicz (2011), implemented in DAGitty (Textor, Hardt, & Knüppel, 2011; *DAGitty algorithm*). In the more realistic situation that the causal structure is not fully known, proposals exist that aim at approximations to such minimal sets. A popular rule requiring partial causal knowledge recommends to adjust for "all common causes of T and Y" (*common cause criterion*; cf. Glymour et al., 2008). Another rule selects all covariates that are associated with the treatment and with the outcome conditional on treatment and do not lie on the causal pathway between treatment and outcome. We call this method *univariate confounder screening (uniTandY)*. It is univariate in the sense that the associations with treatment and outcome are assessed for each covariate separately, not conditionally on other covariates. In contrast, de Luna et al. (2011) suggested to base selection on conditional associations/dependencies and independencies (see also Robins, 1997; VanderWeele & Shpitser, 2011). They describe two algorithms. Roughly speaking, starting from the full set, Algorithm 1 first removes covariates conditionally independent of T given the remaining covariates, then further removes covariates conditionally independent of Y given T and the rest. The alternative Algorithm 2 reverses the role of T and Y. At each stage, the covariates to be removed are chosen so that the number of remaining covariates is as small as possible. The target sets of Algorithms 1 and 2 are each unique but can differ from each other (de Luna et al., 2011). In the example in Figure 1, Algorithm 1 selects the minimal set $\{X_3, X_4\}$ and Algorithm 2 selects the minimal set $\{X_3, X_5\}$. We refer to the general idea as the *CovSel* method, after the associated `R` package. It can be shown that CovSel selects locally minimal adjustment sets under Assumption 4 and additional mild assumptions, given the dependence structure is correctly inferred, which requires a sufficiently large sample size (de Luna et al., 2011).

3.1.3 | Caveats

Selecting minimal sets without knowing the causal diagram is a difficult task in practice. Even the intuitively appealing common cause criterion cannot guarantee under Assumption 4 that a sufficient adjustment set is found, for example, due to unobserved common causes (cf. VanderWeele & Shpitser, 2011). In Figure 1, even though Assumption 4 holds, it will only select $\{X_4\}$, which is insufficient, as U is unobserved. The data-driven methods are of course affected by sampling variation. In particular, univariate confounder screening on the one hand tends to select unnecessarily large sets when the sample size allows many significances. In the example in Figure 1, univariate confounder selection selects X_3, the collider X_8, the unnecessary covariates X_1 and X_2, and both X_4 and X_5 where one of them would suffice. On the other hand, because two tests have to be significant for selection, univariate confounder screening tends to miss important covariates when the sample size is small. Selection of colliders and redundant selection are avoided by CovSel by considering *conditional* (in)dependencies. Except when the true causal diagram is known, none of the aforementioned methods can guarantee globally minimal sets.

3.2 | Outcome approach

3.2.1 | Motivation

The idea of outcome-oriented selection strategies is to determine a sufficient adjustment set that includes strong predictors of the outcome. As a major advantage, adjusting for outcome predictors reduces the standard errors in a variety of settings, including linear outcome regression and propensity score (PS) weighting (Lunceford & Davidian, 2004). Excluding covariates that are conditionally independent of the outcome given treatment and the remaining covariates is a valid strategy that leads to a sufficient adjustment set, provided that Assumption 4 and possibly additional parametric assumptions hold (formally, the conditional independencies relate to potential outcomes as in de Luna et al., 2011, or interventions as in Guo and Dawid, 2010). In terms of a causal diagram, the desired set is sufficient for adjustment and additionally includes all direct causes of the outcome. For the example in Figure 1, the ideal outcome-oriented set would be $\{U, X_5, X_6, X_7\}$. As U is unobserved, the next best set is $\{X_3, X_5, X_6, X_7\}$.

3.2.2 | Examples

Examples are general strategies aiming to select all nonredundant predictors of the outcome. These methods include univariate screening for covariates associated with the outcome either marginally or conditionally on treatment (we refer to the latter as *univariate outcome screening, uniY*). They further include selection methods for parametric regression that aim to *optimize the outcome model* regarding prediction performance or model fit, for example, model selection based on Akaike's information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978), the p-value method (cf. Greenland & Pearce, 2015, for description and criticism), or evaluation of the prediction accuracy using a validation data set. Another

example is *model-free variable selection* as described by Li et al. (2005). Starting from the full set, covariates are removed if conditionally independent of the outcome given treatment and the remaining covariates, based on nonparametric tests. This is the same principle as the first part of CovSel Algorithm 2. Other nonparametric methods include *random forest variable selection* (e.g., Genuer et al., 2010; Kursa & Rudnicki, 2010). An example of how outcome-oriented covariate selection can be combined with treatment modeling is the *outcome-adaptive lasso* where the coefficients of a treatment regression model are penalized inversely proportional to the association of the respective covariates with the outcome in a separate outcome model (Shortreed & Ertefaie, 2017).

3.2.3 | Caveats

Methods focusing on the outcome alone share the drawback that, when used with a small sample size, they might miss covariates that are only weakly associated with the outcome but strongly associated with the treatment and hence still induce confounding bias (Wilson & Reich, 2014). This reflects the fact that covariates important for causal inference are not necessarily as important for outcome prediction, and vice versa. Further, some of the methods have conceptual shortcomings. For example, univariate outcome screening selects redundant covariates and covariates that will not contribute to an improved precision. In the example in Figure 1, the set selected by univariate outcome screening contains X_1 because X_1 is associated with Y conditional on T. However, X_1 is not on a backdoor path and would reduce rather than increase the precision. Multivariate methods avoid this to a certain extent. As an example where even multivariate methods select more covariates than necessary, consider some backdoor path of the form $T \leftarrow X_a \rightarrow X_b \leftarrow U \rightarrow Y$. Although the empty set is sufficient to block this particular path, multivariate methods select $\{X_a, X_b\}$, which is also sufficient but larger than necessary.

3.3 | Treatment approach

3.3.1 | Motivation

As every backdoor path necessarily begins with a direct cause of treatment, the set of all direct causes of T is a sufficient adjustment set (Pearl, 2009), given Assumption 4. Selecting the direct causes of T appears natural when adjustment involves the PS, that is, typically a regression model for treatment given the selected covariates. A main advantage is that selection is clearly separated from any outcome modeling or treatment effect estimation (Rubin, 2001).

3.3.2 | Examples

In principle, all variable selection methods mentioned for the outcome approach can also be used to select predictors of treatment by replacing outcome regression with treatment regression. Basic methods include, for instance, univariate screening for association with the treatment (*univariate treatment screening*) and stepwise regression to *optimize the treatment model* (Weitzen, Lapane, Toledano, Hume, & Mor, 2004).

3.3.3 | Caveats

The treatment model used to estimate the PS should not include causes (or more generally, predictors) of treatment that are not required to block a backdoor path, such as X_1 in Figure 1. It has been shown that adjusting for such unnecessary covariates can lead to a higher variance of causal effect estimates (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Lunceford & Davidian, 2004) and to bias amplification in case of residual unobserved confounding (Bhattacharya & Vogt, 2007; Wooldridge, 2009). Note that treatment-oriented selection is not required to be combined with PS-based adjustment methods but can, for example, be used with regression adjustment. However, this is known to be equally inefficient or biased (Bhattacharya & Vogt, 2007; Myers et al., 2011; Pearl, 2011). Finally, when the sample size is not large enough, association-based methods might miss covariates that are only weakly associated with the treatment but strongly associated with the outcome (Wilson & Reich, 2014).

3.4 | Union set approach

3.4.1 | Motivation

If there is a sufficient adjustment set among the measured pretreatment covariates (Assumption 5), the union of all causes of treatment or outcome is sufficient as well (VanderWeele & Shpitser, 2011). An advantage of this approach is that in the absence of detailed prior knowledge on the causal structure, it is easier for subject-matter experts to justify which covariates are either causes of treatment or outcome. Moreover, when selection is data-driven, the outcome or treatment approaches may miss covariates

that are only weakly associated with the outcome but strongly associated with the treatment, and vice versa. By considering the relationship to the treatment and to the outcome in turn, the selection process gains robustness (Belloni et al., 2014).

3.4.2 | Examples

VanderWeele and Shpitser (2011) suggested to select all causes of treatment or outcome or both (*disjunctive cause criterion*). A data-driven, univariate method pursuing essentially this principle is *univariate double screening* (*uniTorY*), where a covariate is selected if it is associated with treatment or outcome or both (Schafer & Kang, 2008). For high-dimensional problems, Belloni et al. (2014) described *double selection*, where two penalized "nuisance" models are fitted, one for treatment and one for outcome. The union of covariates selected by either penalized model is then used for the causal effect estimation. Another example is the *penalized credible regions* method by Wilson and Reich (2014). Here, Bayesian regression models for treatment and outcome are fitted, and all models within a specified posterior region are defined as "feasible"; within the constrained "feasible" parameter space, the set of covariates with the smallest cardinality is targeted.

3.4.3 | Caveats

The union set approach results in the largest adjustment sets of all approaches considered. This may lead to problems for model fitting and robustness toward misspecification. Replacing the union of causes set by the union set of nonredundant predictors of treatment or outcome, as all methods not based on prior knowledge do, requires again Assumption 4 in order to yield a sufficient adjustment set. Also, the union set might contain strong predictors of treatment that are not needed to avoid confounding bias and therefore might share the disadvantages described for the treatment approach.

3.5 | Causal search approach

3.5.1 | Motivation and example

Entner et al. (2013) described a selection algorithm to which we refer as *EHS algorithm* (for the authors Entner, Hoyer, and Spirtes). It can be viewed as a restricted variant of the Fast Causal Inference (FCI) algorithm for causal search (Spirtes, Glymour, & Scheines, 2000). The EHS algorithm is based on two rules. Rule 1 selects sets $\mathbf{X}_{\text{EHS}} \in \mathbf{X}^*$ so that, for a $X' \in \mathbf{X}^* \setminus \mathbf{X}_{\text{EHS}}$, (i) $X' \not\perp\!\!\!\perp Y \mid \mathbf{X}_{\text{EHS}}$ and (ii) $X' \perp\!\!\!\perp Y \mid \mathbf{X}_{\text{EHS}} \cup T$. It can be shown that sets satisfying Rule 1 are sufficient (Entner et al., 2013). Rule 2 identifies null causal effects and is not discussed here. In the example in Figure 1, assuming that the effect of T on Y is not equal to zero, Rule 1 applies for several combinations of X' and $\mathbf{X}_{\text{EHS}}$, including, for example, $(X' = X_1, \mathbf{X}_{\text{EHS}} = \{X_3, X_5\})$, $(X' = X_2, \mathbf{X}_{\text{EHS}} = \{X_3, X_4, X_5, X_6\})$, and $(X' = X_4, \mathbf{X}_{\text{EHS}} = \{X_2, X_3, X_5, X_7, X_8\})$. The advantage of the EHS algorithm is that in contrast to all aforementioned methods, it is sometimes able to infer that a set of covariates is a sufficient adjustment set based only on Assumption 6, which is considerably weaker than Assumptions 4 and 5.

3.5.2 | Caveats

Although the EHS algorithm returns a list of sufficient adjustment sets in theory, a main disadvantage is that when the dependence structure has to be inferred from data, it likely happens that the rules contradict each other and some amount of user discretion is warranted to interpret the results. Further, when an appropriate X' does not exist the EHS algorithm cannot return any result.

3.6 | Estimation approach

3.6.1 | Motivation

As we are primarily interested in estimating a causal effect, a natural approach is to target sufficient adjustment sets that directly optimize desirable properties of the estimator, especially precision.

3.6.2 | Examples

One can regard selection using the *CIE* criterion as aiming for a low-bias estimator with a minimal covariate set: A benchmark effect is first estimated using all covariates, then covariates are gradually removed until any further removal would result in a change in the estimate of more than, for example, 10%, compared to the benchmark estimate. Greenland et al. (2016) suggested instead to evaluate the *change in the mean-squared error* (*CI-MSE*) of the estimate, which they approximate based on the standard error provided by the regression output. A related and more sophisticated procedure is proposed by Vansteelandt et al. (2012). They described *focused confounder selection* (*FCS*) for logistic regression. Their method takes into account that conditional

effects, such as conditional odds ratios, are not collapsible and therefore focuses on the marginal, regression-standardized (log) odds ratio. The MSE of the marginal effect is estimated using either cross-validation or an asymptotic approximation.

3.6.3 | Caveats

The CIE procedure is a heuristic method that can improve as well as corrupt the properties of the estimator (Greenland & Pearce, 2015). For example, the estimator based on the reduced model can be both more biased and more variable than the benchmark effect. All of the example methods rely on Assumption 4, as they all use the full model to obtain a benchmark estimate. However, even under Assumption 4 there are no guarantees that these methods select sufficient adjustment sets.

3.7 | Mechanism of selection

The six types of target sets described in Sections 3.1–3.6 form our first classification criterion (rows in Table 1). The columns correspond to a second criterion, the mechanism of selection. We distinguish between two general mechanisms, preadjustment and wrapper, and subclassify preadjustment methods into knowledge-based, nonparametric, and parametric methods.

Preadjustment methods separate selection from adjustment. For instance, if selection is completely knowledge-based, it is independent of the adjustment process, which might involve, for example, matching or regression adjustment. The analyst can select different adjustment sets for different adjustment methods (e.g., a minimal set for matching and an outcome-oriented set for regression), but crucially, the selection process is not influenced by the resulting estimate. The same is true for many data-driven selection methods, including all variants of univariate screening. We also classify as preadjustment all methods that perform selection for nuisance models such as the PS. This is in line with the notion that PS estimation is part of the design, not the analysis, of a study (Rubin, 2001).

When using wrapper methods, selection and estimation cannot be separated. Instead, the causal effect is repeatedly estimated with different adjustment sets and one set is selected that optimizes some criterion. The selection procedure is wrapped around the estimation procedure, hence the name (see, e.g., Saeys, Inza, & Larrañaga, 2007, for a similar usage of the term "wrapper"). For example, for CIE the selection criterion is the cardinality of the adjustment set. The CIE method as described in Section 3.6 is a backward procedure, meaning one starts with the full set of covariates, then gradually removes covariates without readding them in later stages. Other types of wrapper algorithms are forward, stepwise, and exhaustive search.

Preadjustment methods have the advantage that they offer a certain amount of safeguard against researchers' discretion because the selection process is not influenced by the estimated causal effect. Further, preadjustment selection is generally flexible, that is, it can be combined with different adjustment methods. In contrast, wrapper methods are specific to one adjustment method, usually a regression model. A disadvantage of all parametric methods, either preadjustment or wrapper, is that one has to assume a functional form for each continuous covariate prior to selection. We also note as an important caveat that whenever data-driven methods are used for selection, inference needs to be adjusted for this selection (Leeb & Pötscher, 2005). For example, resampling the entire selection and estimation process is one way to guarantee valid confidence intervals (Heinze, Wallisch, & Dunkler, 2018). Only few methods are robust toward selection without further provisions, see, for example, Belloni et al. (2014) and Dukes, Avagyan, and Vansteelandt (2018).

4 | SIMULATION SETUP

The aim of our simulation study is twofold: First, we compare the target adjustment sets themselves in the context of typical adjustment methods: (a) outcome regression (combined with standardization for noncollapsible measures, such as odds ratios, when estimating a marginal effect parameter), (b) matching on covariates, (c) matching on the estimated PS, and (d) doubly robust estimation, where both treatment and outcome are modeled and estimators are consistent as long as either model is correctly specified. Doubly robust estimation appears especially attractive when the adjustment set is large, such as with the union set approach, as models are then more likely to be misspecified. In practice, method (a) is often combined with outcome-oriented, method (c) with treatment-oriented, and method (d) with union-set selection. However, it is important to note that all adjustment methods can be combined with any type of adjustment set. Second, we evaluate how well common data-driven methods (see Section 4.2) select their target sets and estimate the target causal effect.

We investigated two general setups, the first following Figure 1 and the second following Figure 2. In setup 1, all causal connections were relatively strong, which made it easy for selection algorithms to detect all relevant associations. In setup 2, in contrast, covariates X_1, X_3, X_5, and X_7 influenced treatment only weakly, and covariates X_2, X_4, X_6, and X_8 influenced outcome only weakly, making it more difficult to detect all associations.

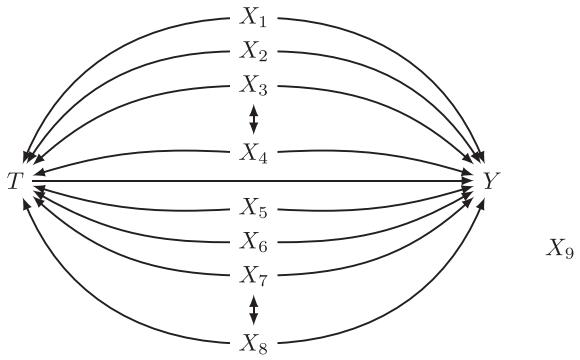


FIGURE 2 Causal diagram used for data generation in setup 2, see Section 4.1. Double-headed arrows indicate correlated error terms, which can be thought of as resulting from an unobserved common cause

For each setup, we investigated two scales of outcome (continuous, binary), three scales of covariates (continuous, mixed scale, binary), two treatment effects (present, absent), and three sample sizes (100, 500, 2,000), resulting in 72 scenarios in total. In the mixed-scale scenarios, the binary covariates were X_3, X_4, X_6, X_7, X_9 in setup 1 and X_1, X_2, X_3, X_4 in setup 2.

In short, the following steps were carried out: (a) Data were generated from the causal diagram in Figure 1 or Figure 2, (b) adjustment sets were determined as the true target sets as well as by applying different selection methods, (c) the causal effect was estimated adjusting, in turn, for the different selected sets. The three steps were repeated 1,000 times. The full code, written for the software package `R` (R Core Team, 2018), is available as Supporting Information on the journal's web page. The online Supporting Information also includes an overview of the `R` packages available for the methods in Table 1.

4.1 | Data generation

The detailed formulas used for data generation are in Appendix A.2. In short, continuous variables were generated according to linear models and binary variables according to logistic models, without interaction effects. Intercepts were chosen such that the prevalence of treatment and outcome, when binary, was about 0.5. In setup 1, for continuous outcome the true ACE is either 0 or 0.5 and an unadjusted analysis has a bias of 0.86/0.58/0.71 for continuous/mixed/binary covariates. For binary outcome, if the treatment is effective the true log(MCOR) is 0.91/0.95/0.88 and an unadjusted analysis has a bias of 0.55/0.37/0.44 for continuous/mixed/binary covariates; if treatment is not effective, the bias is 0.53/0.37/0.47. In setup 2, for continuous outcome the true ACE is either 0 or 0.5 and an unadjusted analysis has a bias of 0.84/1.02/1.15 for continuous/mixed/binary covariates. For binary outcome, the true log(MCOR) is 0.60/0.54/0.51 and an unadjusted analysis has a bias of 0.53/0.61/0.64 for continuous/mixed/binary covariates; if treatment is not effective, the bias is 0.51/0.60/0.63.

4.2 | Implementation of selection methods

The true target adjustment sets for setup 1 are given as follows.

$\mathbf{X}_{\text{min}}$: The minimal target set $\{X_3, X_5\}$ was used.

$\mathbf{X}^*$: The full set $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9\}$ was used.

$\mathbf{X}_Y$: The outcome-oriented target set $\{X_3, X_5, X_6, X_7\}$ was used.

$\mathbf{X}_T$: The treatment-oriented target set $\{X_1, X_2, X_3, X_4\}$ was used.

The following covariate selection strategies were implemented. If not stated otherwise, we used default options for all `R` functions.

AIC: AIC-based selection was implemented as a stepwise (linear or logistic) outcome regression procedure with main effects only, starting with the full set of covariates.

BIC: Analogous to AIC.

Boruta: The `Boruta` package (Kursa & Rudnicki, 2010) was used. All covariates tagged as `Confirmed` by the algorithm were selected into $\mathbf{X}_{\text{Boruta}}$. Boruta is an algorithm for random forest variable selection; covariates are selected based on their so-called variable importance compared to artificially generated noninformative covariates.

CIE: CIE was implemented as a backward (linear or logistic) regression procedure with main effects only. The estimated treatment effect from the full model served as the benchmark estimate. Covariates were then removed one by one such that the CIE, compared to the benchmark estimate, was as small as possible, stopping at a maximum change of 10%. CIE was implemented in two versions: In the coefficient version (CIE_coe), the relevant estimate in each step was the estimated partial regression coefficient of T, corresponding to the effect of T conditional on all other covariates in the model. In the marginal version for binary Y (CIE_mar), regression standardization as described in Section 4.3 was used to estimate the marginal effect of T before assessing the change in the estimate.

CI-MSE: We followed the instructions in Greenland et al. (2016) to implement the CI-MSE method (CI-MSE_coe). In addition, we implemented the variant CI-MSE_mar for binary Y in which the MSE of the marginal treatment effect is estimated in each step, using again standardization, see Section 4.3.

CovSel: For scenarios with only continuous covariates, the package `CovSel` (Häggström, Persson, Waernbaum, & de Luna, 2015) was used. For scenarios with binary covariates, we obtained the source code of the faster `CovSelHigh` (Häggström, 2017; see also Häggström, 2018) from `GitHub`. We modified the code so that continuous covariates were not discretized. We used the modified function with options `method=mmpc, simulate=FALSE, betahat=FALSE`. From the results that were returned by either `cov.sel` or `cov.sel.high`, the sets $\mathbf{X}_{\text{CovSelQ}}$=`Q0 Q1`, $\mathbf{X}_{\text{CovSelX.Y}}$=`X.Y0 X.Y1`, $\mathbf{X}_{\text{CovSelX.TY}}$=`X.T X.Y0 X.Y1`, and $\mathbf{X}_{\text{CovSelZ}}$=`Z0 Z1` were extracted, where $\mathbf{X}_{\text{CovSelQ}}$ is the minimal set selected by Algorithm 1, $\mathbf{X}_{\text{CovSelX.Y}}$ is an outcome-oriented set, $\mathbf{X}_{\text{CovSelX.TY}}$ is the union set of the outcome-oriented and a treatment-oriented set, and $\mathbf{X}_{\text{CovSelZ}}$ is the minimal set selected by Algorithm 2.

doubleAIC: AIC selection as described above was performed separately on the outcome model and the treatment model. All covariates selected by either model were selected into $\mathbf{X}_{\text{doubleAIC}}$.

doubleBIC: Analogous to doubleAIC.

FCS: Focused confounder selection (for binary Y) was implemented as a stepwise procedure for standard logistic regression (without PS), starting from the full model. Code for the main selection algorithm was kindly provided by Prof. Vansteelandt.

uniY: For every covariate $X \in \mathbf{X}^*$, Y was regressed on X and T, using linear regression for continuous Y and logistic regression for binary Y. X was selected into the adjustment set $\mathbf{X}_{\text{uniY}}$ if the p-value of X in this model was ≤ 0.05.

uniTorY: For every covariate $X \in \mathbf{X}^*$, T was regressed on X using logistic regression. X was selected into the adjustment set $\mathbf{X}_{\text{uniTorY}}$ if the p-value of X in this model was ≤ 0.05 or $X \in \mathbf{X}_{\text{uniY}}$ or both.

uniTandY: For every covariate $X \in \mathbf{X}_{\text{uniY}}$, T was regressed on X using logistic regression. X was selected if the p-value of X in this model was ≤ 0.05.

4.3 | Estimation of causal effects

Linear regression: The treatment effect was estimated as the partial regression coefficient corresponding to T and the conventional standard error estimate was calculated. If the linear regression model is correct, this corresponds to the ACE.

Logistic regression: Logistic regression was followed by standardization using the package `stdReg` (Sjölander & Dahlqwist, 2017) to obtain an estimate of the log(MCOR): For $t = 0, 1$, the outcome Y was predicted for each individual with its observed covariate values and T set to t. The mean predicted outcomes μ_0 and μ_1 were obtained and the log(MCOR) was estimated as $\log(\mu_1/(1-\mu_1)/\mu_0 * (1-\mu_0))$.

(*PS*) *matching*: Matching was performed using the package `Matching` (Sekhon, 2011) with `estimand=ATE` (corresponding to the ACE). For PS matching, the PS was estimated with a logistic regression model. For continuous Y, the estimated average treatment effect and the Abadie–Imbens standard error were obtained. For binary Y, the matched sample was employed to estimate μ_0 and μ_1 and the marginal log odds ratio was calculated as described above. Note that none of the matching methods were implemented with any pruning of observations.

Doubly robust estimation: For doubly robust estimation, the `iWeigReg` package (Tan & Shu, 2013) was used. This implements a calibrated likelihood estimator that has been shown to outperform other doubly robust estimation procedures (Tan, 2010). The treatment was modeled by logistic regression, the outcome by linear or logistic regression. Both models are correctly specified if the selected covariates contain the parents of treatment and outcome, but they may be misspecified otherwise.

5 | RESULTS

We discuss the results for two of the 72 scenarios: Scenario A is with setup 1, continuous outcome, continuous covariates, effective treatment, and $N = 500$. Scenario B is similar but with binary outcome. The results for the other scenarios are

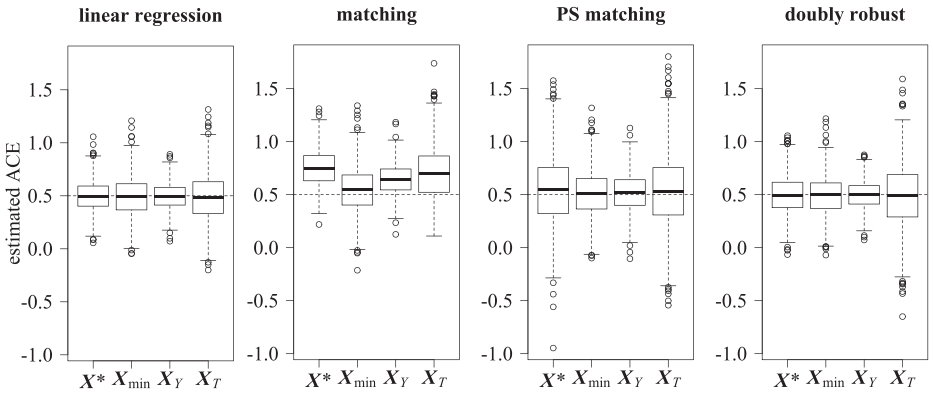


FIGURE 3 Estimated effects in scenario A (continuous outcome, continuous covariates, $N = 500$) adjusted for target sets. Plots show the average causal effect (ACE) estimated by linear regression, matching, propensity score (PS) matching, and doubly robust estimation. The true ACE is 0.5 (dashed line), an unadjusted analysis yields about 1.36

provided as Supporting Information on the journal's webpage. Key findings from the other scenarios are also mentioned in the text.

5.1 | Adjusting for target sets

We start by comparing adjustments with different types of true target adjustment sets. Figure 3 shows box-plots of the estimated ACEs for scenario A with different adjustment methods. Using linear regression, all adjustment sets result in unbiased estimators. The variance is smallest for the outcome-oriented set $\mathbf{X}_Y$ and largest for the treatment-oriented set $\mathbf{X}_T$. In stark contrast, when matching on the covariates, only the minimal set $\mathbf{X}_{\text{min}}$ leads to near unbiased estimators. The bias is largest when adjusting for the full set $\mathbf{X}^*$. The reason is that the more covariates need to be matched on, the worse the matches get with respect to each single covariate. In practice one would prune observations with bad matches; however, as we can see from the improved results, it is not necessary to prune with smaller adjustment sets. Interestingly, the bias for $\mathbf{X}_T$ is notably larger than for $\mathbf{X}_Y$, although these two sets are of the same size. A possible explanation is that covariates strongly associated with treatment tend to be differently distributed in the treatment group versus the control group so that finding good matches is harder. Again, adjusting for $\mathbf{X}_Y$ yields the smallest variance. PS matching results in unbiased or near unbiased estimators for all adjustment sets. The variance is smallest for $\mathbf{X}_Y$, slightly larger for $\mathbf{X}_{\text{min}}$, and considerably larger for $\mathbf{X}^*$ and $\mathbf{X}_T$. The latter confirms previous results (Austin et al., 2007). For doubly robust estimation, similar trends can be observed as for linear regression, with some loss of precision when $\mathbf{X}^*$ or $\mathbf{X}_T$ are used. However, given the greater robustness of this method, it is noteworthy that the loss is only small.

Figure 4 shows the corresponding results of the same analysis for scenario B, that is, binary outcome. For matching and PS matching, the box-plots show very similar trends as in scenario A. When the adjustment method is standardized logistic regression, all estimators are unbiased. The variance is slightly smaller when adjusting for $\mathbf{X}_{\text{min}}$ or $\mathbf{X}_Y$ instead of $\mathbf{X}^*$ or $\mathbf{X}_T$. However, the differences in efficiencies between the types of adjustment sets appear to be quite small.

Varying the sample size or setting the treatment effect to zero does not change the pattern. Further, the trends remain when using all binary covariates. Surprisingly, however, with mixed-scale covariates matching yields close to unbiased estimators for all adjustment sets, see the online Supporting Information.

5.2 | Adjusting for selected sets

Figures 5 and 6 show results obtained by adjusting for sets selected by the different selection methods. The associated tile plots in Figure 7 illustrate how often covariates are selected by which methods. In each plot, the methods are ordered according to their target set. Treatment-oriented selection was not implemented due to inferior results in Section 5. For comparison, we include the box-plot for the full set $\mathbf{X}^*$.

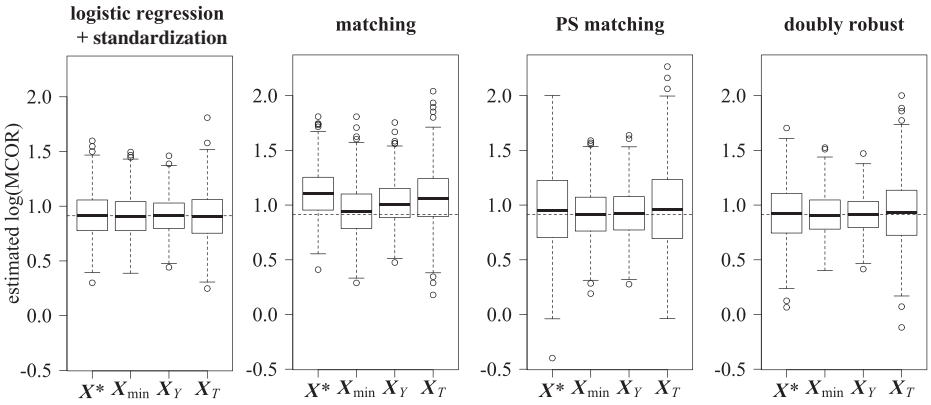


FIGURE 4 Estimated effects in scenario B (binary outcome, continuous covariates, $N = 500$) adjusted for target sets. Plots show the log marginal causal odds ratio (log(MCOR)) estimated by logistic regression with standardization, matching, propensity score (PS) matching, and doubly robust estimation. The true log(MCOR) is about 0.91 (dashed line), an unadjusted analysis yields about 1.46

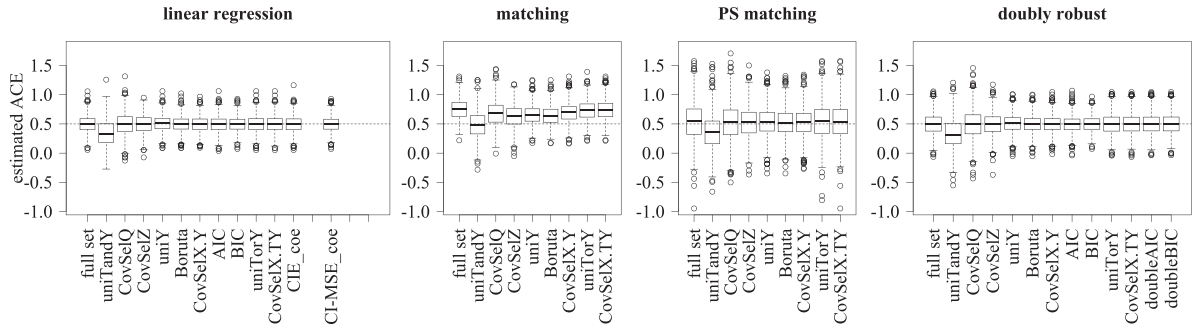


FIGURE 5 Estimated effects in scenario A (continuous outcome, continuous covariates, $N = 500$) adjusted for selected sets. Plots show the average causal effect (ACE) estimated by linear regression, matching, propensity score (PS) matching, and doubly robust estimation. The true ACE is 0.5 (dashed line), an unadjusted analysis yields about 1.36

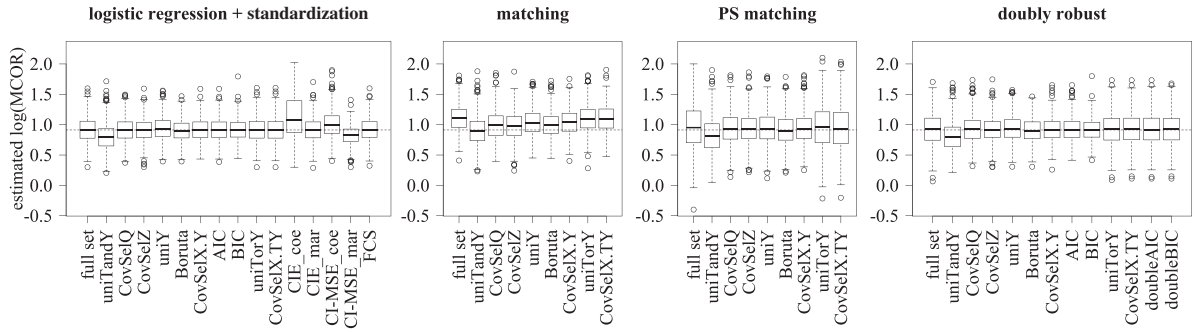


FIGURE 6 Estimated effects in scenario B (binary outcome, continuous covariates, $N = 500$) adjusted for selected sets. Plots show the log marginal causal odds ratio (log(MCOR)) estimated by logistic regression with standardization, matching, propensity score (PS) matching, and doubly robust estimation. The true log(MCOR) is about 0.91 (dashed line), an unadjusted analysis yields about 1.46

5.2.1 | Minimal approach

UniTandY, CovSelQ, and CovSelZ aim at small adjustment sets. For uniTandY, a phenomenon called collider bias can be observed: uniTandY is prone to selecting X_8, a collider on the backdoor path $T \leftarrow X_2 \rightarrow X_8 \leftarrow X_6 \rightarrow Y$. When adjusting for X_8 without also adjusting for either X_2 or X_6, and when all pairwise associations are positive (as in our simulation), a negative association is induced between T and Y, resulting in a negatively biased estimator (Pearl, 2009). This is visible in Figures 5 and

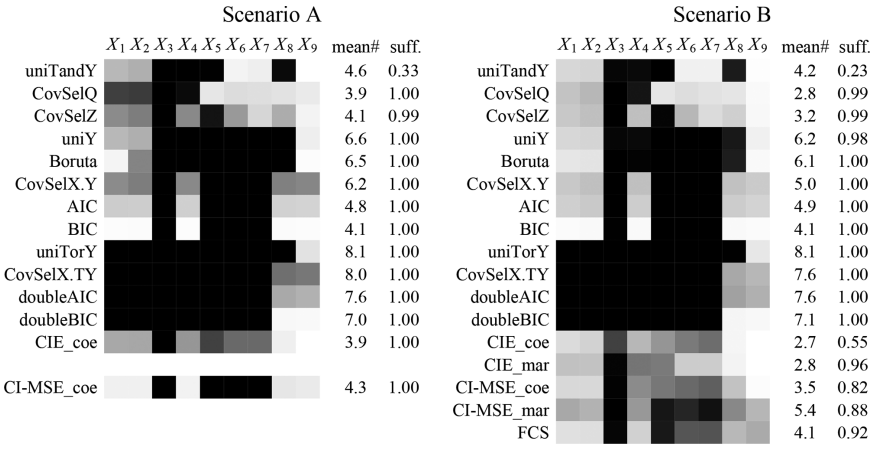


FIGURE 7 Frequencies of covariates selected. The gray level of the tiles is proportional to the number of times the covariate was selected, with white indicating "never selected" and black indicating "always selected." mean#: mean size of selected set; suff.: proportion of times the selected set is sufficient

6 for regression, PS matching, and doubly robust estimation. For matching, the negative bias cancels out with the positive bias from the large number of covariates to be matched on. However, as the size of the set selected by uniTandY strongly depends on the sample size, collider bias cannot be observed for scenario A when the sample size is $N = 100$, as X_8 is then not selected, or when the sample size is $N = 2{,}000$, as X_2 is then selected in addition to X_8 (see online Supporting Information). Ignoring uniTandY, where negative and positive bias cancel out, the smallest bias for matching in scenario A (Figure 5) is achieved by adjusting for the CovSelZ set. The CovSel method performs even better when $N = 2{,}000$ (see online Supporting Information).

5.2.2 | Outcome approach

UniY, Boruta, CovSelX.Y, AIC, and BIC aim at predictors of the outcome. We see in Figure 7 that as expected, uniY tends to select all covariates associated with Y conditional on T. Covariate C_9 is selected in about 5% of cases, reflecting the significance level of the test. The random forest method Boruta selects similar sets but is better able to identify C_9 as an unnecessary covariate. Interestingly, Boruta performs well in combination with matching in many scenarios, including scenario B (Figure 6), although the selected sets are quite large. A possible explanation is that in these scenarios Boruta only rarely selects the strong treatment predictors X_1 and X_2 (Figure 7), for which good matches are especially hard to find. The multivariate methods CovSelX.Y, AIC, and BIC tend to correctly identify $\{X_3, X_5, X_6, X_7\}$ as the direct predictors of Y. The sets selected by AIC and especially BIC are less "noisy" compared to CovSelX.Y, due to the correct parametric assumptions they make. The CovSel method might prove more reliable than AIC and BIC in settings with other than linear influences.

5.2.3 | Union set approach

UniTorY, CovSelX.TY, doubleAIC, and doubleBIC aim at the union set of treatment and outcome predictors. In scenarios A and B, and all other scenarios from setup 1, they tend to select sufficient, but unnecessarily large sets, leading to increased variance and to bias when matching. In setup 2 (see Figure 2), however, where we included many weak associations, the union set methods are the only ones able to find sufficient adjustment sets. As an example, we show in Figure 8 the tile plot for setup 2 with continuous outcome, continuous covariates, effective treatment, and $N = 500$. Only the full set and $\{X_1, \ldots, X_8\}$ are sufficient. The "gaps" in Figure 8 indicate where the selection methods fail to select important covariates due to weak associations.

5.2.4 | Estimation approach

The estimator-oriented selection methods CIE_coe, CIE_mar, CI-MSE_coe, CI-MSE_mar, and FCS are based on parametric regression models (at least in our implementations), hence we show results for their performance only for regression adjustment. In scenario A, selection by CIE_coe (coe for "coefficient version") yields unbiased estimators. The reason is that the full set of covariates is relatively small, so a good estimate is expected without selection (see box-plot "full set" in Figure 5) and the estimate after CIE selection cannot deviate from this by more than 10%, by definition. The same is true for CIE_mar (mar for "marginal version") in scenario B. The coefficient version leads to bias here because the odds ratio is noncollapsible. In general,

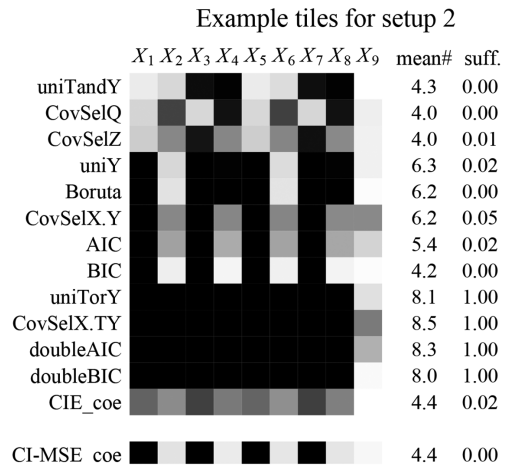


FIGURE 8 Frequencies of covariates selected, for a scenario with setup 2, continuous outcome, continuous covariates, effective treatment, and $N = 500$. The gray level of the tiles is proportional to the number of times the covariate was selected, with white indicating "never selected" and black indicating "always selected." mean#: mean size of selected set; suff.: proportion of times the selected set is sufficient

the CIE method will always perform well when the full model does, especially when the true treatment effect is small. However, this does not mean that the selected covariates are sufficient. For example, in the scenario in Figure 8, CIE selection yields a close to unbiased estimator (see online Supporting Information), yet the selected set is sufficient in only 2% of cases (see Figure 8). Further, it is to be expected that CIE does not perform well when the number of covariates is large so that the estimate from the full model has a large variance. CI-MSE_coe performs well in setup 1 for linear regression, but is unreliable for logistic regression. CI-MSE_mar performs comparable to the union set methods for setup 2 when the sample size is at least 500, but is outperformed in most scenarios by FCS in setup 1 in terms of bias and variance (see online Supporting Information).

5.3 | Coverage

For scenario A, standard ways of estimating the standard error exist; we calculated 95% confidence intervals and assessed how often the true effect of 0.5 was included (Table 2). Note that these standard errors are not corrected for the preceding variable selection. As expected, strong bias leads to severe undercoverage, see, for example, matching on $\mathbf{X}^*$. Moderate undercoverage without bias occurs for doubly robust estimation, especially when the adjustment set is large. For PS matching, the confidence intervals tend to be very wide, resulting in coverage greater than 0.95. For linear regression, mild undercoverage is seen for several adjustment sets, including the target set $\mathbf{X}_Y$. However, we note that 1,000 replications are not sufficient to pin down the mean coverage to the second decimal place, and from the online Supporting Information we find that overall, coverage reaches values close to 0.95 when adjusting for the target sets. When adjusting for the selected sets, the coverage is good in general for $N = 2{,}000$ and decreases with sample size. This is to be expected, as selection is relatively stable when the observations-to-covariates ratio is high (Heinze et al., 2018) so that adjusting for the selected set comes closer to adjusting for a fixed set.

5.4 | Conclusions

In summary, the following conclusions can be drawn from the simulation results. For setup 1 (where only two of nine covariates are necessary for confounding adjustment and the association between causally connected variables is relatively strong), linear regression, logistic regression with standardization, PS matching, and doubly robust estimation worked best when adjustment was for the outcome-oriented target set. Interestingly, however, the results obtained by logistic regression with standardization were quite insensitive to which target set was adjusted for in our simulation. In contrast to all other adjustment methods, matching on the covariates worked best when adjustment was for the minimal set, but was biased when matching was on more than two covariates. In practice, the bias may be reduced by pruning, that is, discarding observations for which no suitable matches can be found. This can be thought of as discarding observations $\mathbf{x}$ for which Assumption 3, positivity, is violated empirically (though not necessarily in the limit). It is important to check for sufficient overlap between the treatment groups regarding the covariates one wants to adjust for, not only for matching but also when using other adjustment methods. In our simulation, regression and PS matching are not affected by insufficient overlap because they are based on correct models that allow for valid extrapolation.

TABLE 2 Coverage analysis for scenario A (continuous outcome, continuous covariates, $N = 500$)

	Linear regression		Matching		PS matching		Doubly robust	
Adjustment set	**Covered**	**Width**	**Covered**	**Width**	**Covered**	**Width**	**Covered**	**Width**
$\mathbf{X}^*$	0.94	0.55	0.68	0.67	0.98	1.48	0.89	0.58
$\mathbf{X}_{\text{min}}$	0.94	0.71	0.94	0.82	0.94	0.87	0.94	0.72
$\mathbf{X}_{Y}$	0.93	0.46	0.82	0.56	0.99	0.88	0.93	0.47
$\mathbf{X}_{T}$	0.95	0.89	0.89	0.99	0.96	1.46	0.90	0.96
$\mathbf{X}_{\text{uniTandY}}$	0.79	0.71	0.92	0.81	0.89	1.11	0.77	0.74
$\mathbf{X}_{\text{CovSelQ}}$	0.96	0.82	0.89	0.92	0.96	1.31	0.92	0.88
$\mathbf{X}_{\text{CovSelZ}}$	0.97	0.70	0.90	0.80	0.97	1.13	0.94	0.73
$\mathbf{X}_{\text{uniY}}$	0.94	0.51	0.83	0.62	0.98	1.17	0.91	0.53
$\mathbf{X}_{\text{Boruta}}$	0.93	0.50	0.84	0.62	0.98	1.13	0.92	0.52
$\mathbf{X}_{\text{CovSelX.Y}}$	0.94	0.50	0.76	0.61	0.99	1.13	0.92	0.52
$\mathbf{X}_{\text{AIC}}$	0.92	0.48	–	–	–	–	0.91	0.49
$\mathbf{X}_{\text{BIC}}$	0.92	0.47	–	–	–	–	0.93	0.47
$\mathbf{X}_{\text{uniTorY}}$	0.94	0.54	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{CovSelX.TY}}$	0.94	0.54	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{doubleAIC}}$	–	–	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{doubleBIC}}$	–	–	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{CIE\_coe}}$	0.96	0.63	–	–	–	–	–	–
$\mathbf{X}_{\text{CI-MSE\_coe}}$	0.92	0.47	–	–	–	–	–	–

Note. Shown are the proportion of times the true effect of 0.5 was included in the 95% confidence interval (Covered) and the mean width of the interval (Width) when adjusting for the different adjustment sets using linear regression, matching, propensity score (PS) matching, or doubly robust estimation.

We confirmed that univariate confounder selection (uniTandY) is prone to select colliders and that the CIE and the CI-MSE procedure must be used in their marginal versions when the effect of interest is a marginal effect. An important result is that for setup 2 (where eight of nine covariates are necessary for sufficient adjustment but each covariate is responsible for only a small amount of confounding), only methods pursuing the union set approach reliably selected sufficient sets in our simulation. This is unfortunate as the union set as a target set generally leads to inefficient estimation and bias when matching is used. Hence, the question of which approach to use can best be answered based on a priori knowledge of the structure and magnitude of the causal relationships between all variables. If such knowledge is lacking, and if avoiding confounding bias, not efficiency, is the main concern then the union set approach appears to be the safest bet, unless matching is used.

6 | DISCUSSION

In this paper, we distinguished six general approaches to covariate selection for causal inference based on the type of target adjustment set. Common theoretically founded and heuristic methods for implementing these approaches were compared with regard to their theoretical as well as empirical properties. It becomes clear that most selection methods aim at covariate *reduction* rather than selection because they assume that the full set of covariates is a sufficient adjustment set (Assumption 4). Moreover, we argued, and illustrated with simulated data, that different adjustment methods need different types of adjustment sets.

For non- or semiparametric methods, especially matching, small or minimal adjustment sets are clearly desirable. If the underlying causal diagram is known, these can be determined with the DAGitty algorithm. Otherwise, under Assumption 4, CovSel can be recommended while neither the common cause criterion nor univariate confounder screening should be used.

Under Assumption 4, selecting all nonredundant outcome predictors, that is, deselecting variables that are conditionally independent of the outcome given the set of included covariates and treatment increased efficiency not only for regression adjustment but also for PS matching and doubly robust estimation in our simulation. Here many of the data-driven approaches, explicitly or implicitly testing for this type of conditional independence, performed well; univariate outcome screening cannot be recommended.

The treatment-oriented approach cannot be recommended as it is outperformed by all other approaches for all methods of adjustment considered. Note an important difference to the outcome approach: including strong treatment predictors not needed

to adjust for confounding is harmful regarding efficiency of all adjustment methods, and can amplify bias when there is residual unobserved confounding. In contrast, and under Assumption 4, including strong outcome predictors, even if not needed to avoid confounding bias, can still increase efficiency. The main reason to use the treatment approach would be to separate covariate selection completely from modeling of the outcome, for example, to avoid postselection bias.

The disjunctive cause criterion has the advantage that it leads to valid adjustment sets under the weaker Assumption 5 without requiring full knowledge of the underlying causal diagram. However, in many situations, one will not even have the expert knowledge to identify the causes of treatment or outcome. Data-driven methods then require again Assumption 4. The union set approach appears most useful in situations where there are many weak confounders, and where avoidance of confounding bias is the primary concern over efficiency. With the resulting typically large size of adjustment set one may be particularly interested in approaches that are robust toward model misspecification.

In the absence of any prior knowledge to justify Assumptions 4 or 5, the EHS algorithm is an interesting alternative but has not demonstrated its practical use in real-life data examples yet.

There are a number of limitations to our investigation. First, the list of example methods we mention is, of course, not exhaustive. Among others, we did not consider methods that combine two or more approaches, such as the combination of the disjunctive cause criterion with model-free backward selection (VanderWeele & Shpitser, 2011), combinations of causal diagrams and CIE (Evans, Chaix, Lobbedez, Verger, & Flahault, 2012; Weng, Hsueh, Messam, & Hertz-Picciotto, 2009), or the adjustment uncertainty algorithm by Crainiceanu, Dominici, and Parmigiani (2008) combining outcome- and treatment-oriented selection with the CIE criterion. Also, we did not consider model averaging approaches as described in Wang, Parmigiani, and Dominici (2012), Zigler and Dominici (2014), and Talbot, Lefebvre, and Atherton (2015).

In our simulation study, we generated data according to linear and logistic models. In practice, other functional forms for relations between variables may be more plausible. Especially when the covariates are continuous, selection of the covariates itself is only one half of the problem, the other half being model specification, for example, selection of higher order terms and interaction terms. One approach is to consider such terms as additional covariates (Belloni et al., 2014). Another approach selects and adjusts nonparametrically, for example, combining CovSel selection with matching on the covariates.

Another important point we touched only briefly is postselection inference. When effects are estimated from the same data as used to select covariates, the standard errors, confidence intervals, p-values, etc. reported by software are inappropriate. Although in our simulation undercoverage was primarily driven by bias, the effects of postselection inference will be more pronounced with more covariates. For this reason, among others, Heinze et al. (2018) advised to refrain from data-driven selection of covariates altogether when the number of covariates is small to moderate. Although this advice is in principle also sensible, under Assumption 4, when the aim is causal inference, the set of potential confounders will often not be small.

Importantly, all selection methods for causal inference rely on untestable assumptions. The assumption that the full set of measured covariates is sufficient for confounding adjustment, as assumed by the majority of methods, is strong and can only be justified by subject-matter knowledge. Although ideally, one would specify a causal diagram to identify the exact desired target adjustment set (Hernán, Hernández-Díaz, Werler, & Mitchell, 2002), this is often not practical. An understanding of the strengths and weaknesses of alternative selection methods is therefore crucial; our classification and comparison contribute to such an understanding.

ACKNOWLEDGMENTS

We thank two referees and the associate editor for their valuable comments and suggestions. We gratefully acknowledge financial support of the German Research Foundation (DFG – Project DI 2372/1-1).

CONFLICTS OF INTEREST

The authors have declared no conflict of interest.

ORCID


Janine Witte http://orcid.org/0000-0003-0346-2633

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*, 734–753.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–972.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*, 608–650.

Bhattacharya, J., & Vogt, W. B. (2007). *Do instrumental variables belong in propensity scores?* (NBER Technical Working Paper No. 343). Cambridge, MA: National Bureau of Economic Research. Revised 2009. Retrieved from http://www.nber.org/papers/t0343

Boulesteix, A.-L., Binder, H., Abrahamowicz, M., & Sauerbrei, W. (2018). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, *60*, 216–218.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149–1156.

Crainiceanu, C. M., Dominici, F., & Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika*, *95*, 635–651.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, *70*, 161–189.

de Luna, X., & Johansson, P. (2014). Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, *2*, 187–199.

de Luna, X., Waernbaum, I., & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, *98*, 861–875.

Dukes, O., Avagyan, V., & Vansteelandt, S. (2018). High-dimensional doubly robust tests for regression parameters. Preprint arXiv:1805.06714v2.

Entner, D., Hoyer, P. O., & Spirtes, P. (2013). Data-driven covariate selection for nonparametric estimation of causal effects (pp. 256–264). In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), Scottsdale, AZ.

Evans, D., Chaix, B., Lobbedez, T., Verger, C., & Flahault, A. (2012). Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology. *BMC Medical Research Methodology*, *12*, 156.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*, 2225–2236.

Glymour, M. M., Weuve, J., & Chen, J. T. (2008). Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: Measurement, selection, and bias. *Neuropsychology Review*, *18*, 194–213.

Greenland, S., Daniel, R., & Pearce, N. (2016). Outcome modelling strategies in epidemiology: Traditional methods and basic alternatives. *International Journal of Epidemiology*, *45*, 565–575.

Greenland, S., & Pearce, N. (2015). Statistical foundations for model-based adjustments. *Annual Review of Public Health*, *36*, 89–108.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*, 37–48.

Guo, H., & Dawid, A. P. (2010). Sufficient covariates and linear propensity analysis (pp. 281–288). Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy.

Häggström, J. (2017). CovSelHigh: Model-free covariate selection in high dimensions [R package version 1.1.1]. Retrieved from https://CRAN.R-project.org/package=CovSel

Häggström, J. (2018). Data-driven confounder selection via Markov and Bayesian networks. *Biometrics*, *74*, 389–398.

Häggström, J., Persson, E., Waernbaum, I., & de Luna, X. (2015). CovSel: An R package for covariate selection when estimating average causal effects. *Journal of Statistical Software*, *68*, 1–20.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection—A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*, 431–449.

Hernán, M. A., Hernández-Díaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, *155*, 176–184.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4–29.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*, 1–13.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, *21*, 21–59.

Li, L., Cook, D. R., & Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 285–299.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., & Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, *174*, 1213–1222.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, *174*, 1223–1227.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, *79*, 321–334.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality* (pp. 69–117). New York: Springer.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169–188.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, *42*, 1–52.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*, 1111–1122.

Sjölander, A., & Dahlqwist, E. (2017). stdReg: Regression standardization. [R package version 2.2.0]. Retrieved from https://CRAN.R-project.org/package=stdReg

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT press.

Talbot, D., Lefebvre, G., & Atherton, J. (2015). The Bayesian causal effect estimation algorithm. *Journal of Causal Inference*, *3*, 207–236.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, *97*, 661–682.

Tan, Z., & Shu, H. (2013). iWeigReg: Improved methods for causal inference and missing data problems [R package version 1.0]. Retrieved from https://CRAN.R-project.org/package=iWeigReg

Textor, J., Hardt, J., & Knüppel, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, *22*, 745.

Textor, J., & Liśkiewicz, M. (2011). Adjustment criteria in causal diagrams: An algorithmic perspective (pp. 681–688). Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011), Barcelona, Spain.

VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, *67*, 1406–1413.

Vansteelandt, S., Bekaert, M., & Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, *21*, 7–30.

Wang, C., Parmigiani, G., & Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, *68*, 661–671.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety*, *13*, 841–853.

Weng, H.-Y., Hsueh, Y.-H., Messam, L. L. McV., & Hertz-Picciotto, I. (2009). Methods of covariate selection: Directed acyclic graphs and the change-in-estimate procedure. *American Journal of Epidemiology*, *169*, 1182–1190.

Wilson, A., & Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, *70*, 852–861.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables? (Working Paper). East Lansing, MI: Michigan State University. Retrieved from http://econ.msu.edu/faculty/wooldridge/docs/treat1r6.pdf

Zhang, Z. (2008). Estimating a marginal causal odds ratio subject to confounding. *Communications in Statistics—Theory and Methods*, *38*, 309–321.

Zigler, C. M., & Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, *109*, 95–107.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Witte J, Didelez V. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*. 2019;61:1270–1289. https://doi.org/10.1002/bimj.201700294

APPENDIX

A.1 Introduction to causal diagrams

Causal diagrams are statistical models associated with *directed acyclic graphs* (DAGs). In a DAG, variables are represented by *nodes* and causal relationships are represented by arrows, also called *directed edges*. A directed edge from node A to node B, $A \rightarrow B$, means that A has a causal effect on B that is not mediated by any other variable in the DAG. A is then called a *parent* or *direct cause* of B. Although an edge may represent a zero effect, the absence of an edge always indicates that there is no effect.

Any sequence of nodes joined by edges in a DAG is called a *path*, regardless of how the edges are oriented. A path is *directed* if all edges have the same direction. If there is a directed path from A to B, then A is called an *ancestor* or *cause* of B and B is called a *descendant* of A. A directed path from a node to itself is called a *cycle* and is not allowed in a DAG. If A is a nondescendant of B, a backdoor path between A and B is defined as a path from A to B that starts with an arrowhead at A. A *collider* on a path is a node at which two arrowheads collide with respect to that path, $\rightarrow C \leftarrow$.

Causal diagrams have to satisfy the condition of causal sufficiency demanding that every common cause of two variables in the DAG is a node in the DAG as well. A causal diagram not only illustrates causal relations, but also represents the conditional independence structure between all variables in the DAG. A conditional independence between variables A and B given variables $\mathbf{C}$ can be read off the graph by checking whether the nodes in $\mathbf{C}$ *block* every path between nodes A and B. A path between A and B is blocked by $\mathbf{C}$ if (a) it contains a noncollider in $\mathbf{C}$ or (b) it contains a collider such that neither the collider itself nor any descendant thereof are in $\mathbf{C}$. The special case of marginal independence between variables A and B can be seen by checking whether the empty set blocks every path between A and B. This is only possible if there is no path or every path contains a collider. Although blocked paths guarantee conditional independencies, an unblocked path means that a (conditional) association is possible.

A.2 Data-generating mechanisms

In all scenarios, continuous covariates with no incoming arrows were drawn from the standard normal distribution. Binary covariates with no incoming arrows were drawn from the Bernoulli distribution with probability $1/(1+\exp(2))$. In setup 2, the correlated covariates (X_3, X_4) were generated from a bivariate normal distribution such that both had a standard normal distribution marginally and their covariance was 0.5. In the scenarios that specified (X_3, X_4) as binary, they were discretized using 0 as a threshold. The same applies for (X_7, X_8). The remaining covariates were generated according to the following formulas.

In setup 1, for binary X_3, $\mathrm{P}(X_3 = 1) = 1/(1+\exp(2-\beta_{3.U}U))$. For continuous X_3, $X_3 = \beta_{3.U}U + \varepsilon_3$ with $\varepsilon_3 \sim \mathcal{N}(0,1)$, and X_3 was standardized afterward to have zero mean and unit variance. For binary X_5, $\mathrm{P}(X_5 = 1) = 1/(1+\exp(2-\beta_{5.4}X_4))$; for continuous X_5, $X_5 = \beta_{5.4}X_4 + \varepsilon_5$ with $\varepsilon_5 \sim \mathcal{N}(0,1)$ followed by standardization. For binary X_8, $\mathrm{P}(X_8 = 1) = 1/(1+\exp(2-\beta_{8.2}X_2 - \beta_{8.6}X_6))$; for continuous X_8, $X_8 = \beta_{8.2}X_2 + \beta_{8.6}X_6 + \varepsilon_8$ with $\varepsilon_8 \sim \mathcal{N}(0,1)$ followed by standardization. The treatment was generated according to $\mathrm{P}(T = 1) = 1/(1+\exp(\beta_0^T - \beta_{T.1}X_1 - \beta_{T.2}X_2 - \beta_{T.3}X_3 - \beta_{T.4}X_4))$ and the outcome according to $Y = \beta_{Y.U}U + \beta_{Y.5}X_5 + \beta_{Y.6}X_6 + \beta_{Y.7}X_7 + \beta_{Y.T}T + \varepsilon_Y$ with $\varepsilon_Y \sim \mathcal{N}(0,1)$ or $\mathrm{P}(Y = 1) = 1/(1+\exp(\beta_0^Y - \beta_{Y.U}U - \beta_{Y.5}X_5 - \beta_{Y.6}X_6 - \beta_{Y.7}X_7 - \beta_{Y.T}T))$, respectively. The treatment effect $\beta_{Y.T}T$ was 0.5 when treatment had an effect and the outcome was continuous, 1.5 when treatment had an effect and the outcome was binary, and 0 otherwise. The other parameter values varied according to the scale of the covariates: For continuous covariates, $\beta_{3.U} = \beta_{5.4} = \beta_{8.2} = \beta_{8.6} = \beta_{T.1} = \beta_{T.2} = \beta_{T.3} = \beta_{T.4} = \beta_{Y.U} = \beta_{Y.5} = \beta_{Y.6} = \beta_{Y.7} = 1$, $\beta_0^T = 0$, and $\beta_0^Y = 0.5$. For mixed-scale covariates, $\beta_{3.U} = \beta_{8.2} = \beta_{T.1} = \beta_{T.2} = \beta_{Y.U} = \beta_{Y.5} = 1$, $\beta_{5.4} = \beta_{8.6} = \beta_{T.3} = \beta_{T.4} = \beta_{Y.6} = \beta_{Y.7} = 3$, $\beta_0^T = 0.7$, and $\beta_0^Y = 1.1$. For binary covariates, $\beta_{3.U} = \beta_{5.4} = \beta_{8.2} = \beta_{8.6} = \beta_{T.1} = \beta_{T.2} = \beta_{T.3} = \beta_{T.4} = \beta_{Y.U} = \beta_{Y.5} = \beta_{Y.6} = \beta_{Y.7} = 3$, $\beta_0^T = 1.4$, and $\beta_0^Y = 1.9$.

In setup 2, $\mathrm{P}(T = 1) = 1/(1+\exp(\beta_0^T - \beta_{T.1}X_1 - \beta_{T.2}X_2 - \beta_{T.3}X_3 - \beta_{T.4}X_4 - \beta_{T.5}X_5 - \beta_{T.6}X_6 - \beta_{T.7}X_7 - \beta_{T.8}X_8))$ and $Y = \beta_{Y.1}X_1 + \beta_{Y.2}X_2 + \beta_{Y.3}X_3 + \beta_{Y.4}X_4 + \beta_{Y.5}X_5 + \beta_{Y.6}X_6 + \beta_{Y.7}X_7 + \beta_{Y.8}X_8 + \beta_{Y.T}T + \varepsilon_Y$ with $\varepsilon_Y \sim \mathcal{N}(0,1)$ or $\mathrm{P}(Y = 1) = 1/(1+\exp(\beta_0^Y - \beta_{Y.1}X_1 - \beta_{Y.2}X_2 - \beta_{Y.3}X_3 - \beta_{Y.4}X_4 - \beta_{Y.5}X_5 - \beta_{Y.6}X_6 - \beta_{Y.7}X_7 - \beta_{Y.8}X_8 - \beta_{Y.T}T))$, respectively. The treatment effect $\beta_{Y.T}$ was 0.5 when treatment had an effect and the outcome was continuous, 1 when treatment had an effect and the outcome was binary, and 0 otherwise. The other parameter values varied according to the scale of the covariates: For continuous covariates, $\beta_{T.2} = \beta_{T.4} = \beta_{T.6} = \beta_{T.8} = \beta_{Y.1} = \beta_{Y.3} = \beta_{Y.5} = \beta_{Y.7} = 1$, $\beta_{T.1} = \beta_{T.3} = \beta_{T.5} = \beta_{T.7} = \beta_{Y.2} = \beta_{Y.4} = \beta_{Y.6} = \beta_{Y.8} = 0.05$, $\beta_0^T = 0$, and $\beta_0^Y = 0.5$. For mixed-scale covariates, $\beta_{T.2} = \beta_{T.4} = \beta_{Y.1} = \beta_{Y.3} = 3$, $\beta_{T.6} = \beta_{T.8} = \beta_{Y.5} = \beta_{Y.7} = 1$, $\beta_{T.1} = \beta_{T.3} = \beta_{Y.2} = \beta_{Y.4} = 0.15$, $\beta_{T.5} = \beta_{T.7} = \beta_{Y.6} = \beta_{Y.8} = 0.05$, $\beta_0^T = -1.2$, and $\beta_0^Y = -0.7$. For binary covariates, $\beta_{T.2} = \beta_{T.4} = \beta_{T.6} = \beta_{T.8} = \beta_{Y.1} = \beta_{Y.3} = \beta_{Y.5} = \beta_{Y.7} = 3$, $\beta_{T.1} = \beta_{T.3} = \beta_{T.5} = \beta_{T.7} = \beta_{Y.2} = \beta_{Y.4} = \beta_{Y.6} = \beta_{Y.8} = 0.15$, $\beta_0^T = -2.5$, and $\beta_0^Y = -2$.

4 The optimal adjustment set

In this chapter, I follow up on the connection between graph-aided confounder selection, backward outcome regression selection and efficient adjustment. We conjectured in Witte and Didelez (2019) that outcome-oriented adjustment yields small standard errors for a number of estimators, which we supported by theoretical and empirical evidence from the literature as well as our own simulation results. However, our graphical characterisation of the target set of outcome-oriented selection remained vague: 'In terms of a causal diagram, the desired set is sufficient for adjustment and additionally includes all direct causes of the outcome.' (Section 3.2.1 of Paper 1, Witte and Didelez, 2019).

It is intuitively clear that adjusting for direct causes, i.e. parents, of the outcome is beneficial in terms of the efficiency of e.g. the ordinary least squares (OLS) estimator. However, the set of parents of the outcome may violate the adjustment criterion for DAGs (Definition 13), as illustrated by the causal DAG in Figure 9. Suppose that X is the treatment and Y is the outcome of interest. The only valid adjustment set is $\{A\}$, but A is not a parent of Y. Node B, on the other hand, is a parent of Y, but is a forbidden node that must not be adjusted for. Thus, in this example the parent set of the outcome satisfies neither of the two conditions of the adjustment criterion.

In this chapter, I define the optimal adjustment set (**O**-set), which is a valid adjustment set and, loosely speaking, as 'close' to the set of parents of Y as possible. A central result will be that in a Gaussian setting, adjusting for the **O**-set is more efficient in terms of the asymptotic variance of the OLS estimator than adjusting for

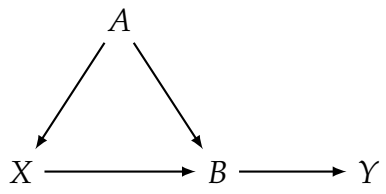


Figure 9: Example causal DAG in which the set of parents of the outcome Y is not a valid adjustment set relative to (X, Y).

any other valid adjustment set. Importantly, this does not rely on the parameters of the Gaussian distribution, but only on the graphical structure. Further, I show that the **O**-set is the target set of Procedure 5A (backward outcome regression selection) above, thereby answering another question left open in Chapter 3.

While Chapter 3 was focussed on causal DAGs, I here consider the larger class of causal MPDAGs. It is useful to distinguish between so-called *amenable* and *non-amenable* MPDAGs, where amenability is defined with respect to a given treatment-outcome pair $(\mathbf{X}, \mathbf{Y})$ (Perković, 2020). In an amenable causal MPDAG, the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ is the same in all represented possibly causal DAGs. Amenable MPDAGs behave similarly to DAGs in several aspects relevant for adjustment, which makes generalisations of results from DAGs to amenable MPDAGs relatively straight-forward. For non-amenable MPDAGs, where the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ differs among the represented DAGs, Maathuis et al. (2009) developed the *IDA algorithm* (Intervention Calculus When the DAG is Absent) for identifying and estimating all possible causal effects compatible with the MPDAG in a computationally efficient manner. Adjustment in amenable MPDAGs and the IDA algorithm are the topic of Section 4.1.

In Section 4.2, I introduce the *forbidden projection* for singleton X and Y in an amenable MPDAG. This special latent projection simplifies a given amenable MPDAG, while retaining all information relevant for identification by adjustment. I then use the forbidden projection to define the **O**-set and to prove that it is a valid adjustment set. Section 4.3 contains the central optimality result. All proofs presented in Sections 4.2 and 4.3 are my own work.

While working on this, I became aware of the (at that time unpublished) work of Henckel et al. (2019), who gave an alternative proof for the optimality of the **O**-set in a setting more general than the one in Section 4.3. In particular, they showed that the result holds also for sets $\mathbf{X}$ and $\mathbf{Y}$, and under weaker parametric assumptions requiring a linear system but not Gaussianity. Their results are summarised in Section 4.4, together with more recent results concerning non-parametric estimation.

The publication for this chapter, *Witte, Henckel, Maathuis and Didelez (2020)*, is included in Section 4.5. It contains three aspects of efficient adjustment using the **O**-set: First the forbidden projection in a slightly more general version than in Section 4.2, second a new version of the IDA algorithm that uses the **O**-set, and third a proof that the **O**-set is the target set of backward outcome regression selection.

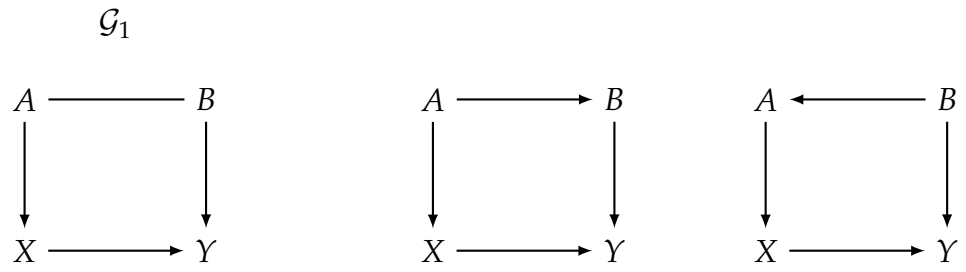


Figure 10: Amenable causal MPDAG $\mathcal{G}_1$ and represented DAGs. $\{A\}$, $\{B\}$ and $\{A, B\}$ are valid adjustment sets relative to (X, Y) in $\mathcal{G}_1$.

4.1 Adjustment criteria for MPDAGs

A valid adjustment set in a causal MPDAG is defined analogously to a valid adjustment set in a causal DAG (see Definition 10):

Definition 30 (Valid adjustment set in a causal MPDAG; Perković et al., 2017)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal MPDAG* $\mathcal{G}$*, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *is a* valid adjustment set *relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$ *if* $\mathbf{Z}$ *is a* valid adjustment set *(according to Definition 9) relative to* $(\mathbf{X}, \mathbf{Y})$ *in every density compatible with* $\mathcal{G}$*.*

It follows that a set $\mathbf{Z}$ is a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in an MPDAG $\mathcal{G}$ if and only if it is a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in every possibly causal DAG in $[\mathcal{G}]$. For illustration, consider first the MPDAG $\mathcal{G}_1$ in Figure 10. It represents two DAGs, also shown in the figure, one with $A \to B$ and one with $A \leftarrow B$. If X and Y are the treatment and outcome, respectively, then there exists exactly one non-causal path in either DAG, with non-colliders A and B. Hence, there are three valid adjustment sets each relative to (X, Y) in $\mathcal{G}_1$ and in the two DAGs: $\{A\}$, $\{B\}$ and $\{A, B\}$. In contrast, consider the MPDAG $\mathcal{G}_2$ and the two DAGs it represents in Figure 11. One of the DAGs contains the causal path $X \to A \to Y$, whereas the other one contains the non-causal path $X \leftarrow A \to Y$. Hence, in one DAG the empty set is a valid adjustment set relative to (X, Y) and $\{A\}$ is not valid, while in the other one $\{A\}$ is valid and the empty set is not. In consequence, there exists no valid adjustment set relative to (X, Y) in $\mathcal{G}_2$.

4.1.1 Amenable MPDAGs

Perković et al. (2017) formulated a sound and complete adjustment criterion for causal MPDAGs (Definition 31, Proposition 32). Using the criterion, valid adjustment sets can be read off a causal MPDAG without first considering the repres-

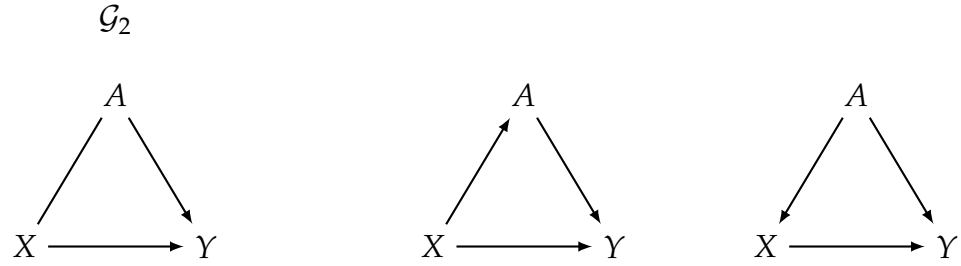


Figure 11: Non-amenable causal MPDAG $\mathcal{G}_2$ and represented DAGs. No valid adjustment set relative to (X, Y) in $\mathcal{G}_2$ exists .

ented DAGs. In a causal MPDAG $\mathcal{G}$ containing disjoint node sets $\mathbf{X}$ and $\mathbf{Y}$, the *possibly causal nodes* $\text{posscn}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$ are all nodes on possibly causal paths from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{G}$, excluding the nodes in $\mathbf{X}$. The *forbidden set* with respect to $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}$ is defined as $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \text{possde}(\text{posscn}(\mathbf{X}, \mathbf{Y}, \mathcal{G}), \mathcal{G})$.

Definition 31 (Adjustment criterion for MPDAGs; Perković et al., 2017)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal MPDAG* $\mathcal{G}$*, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *satisfies the* adjustment criterion for MPDAGs *relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$ *if*

(i) all proper possibly causal paths from $\mathbf{X}$ *to* $\mathbf{Y}$ *contain an edge out of* $\mathbf{X}$*,*

(ii) $\mathbf{Z} \cap \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \emptyset$*,*

(iii) all proper non-causal definite-status paths from $\mathbf{X}$ *to* $\mathbf{Y}$ *in* $\mathcal{G}$ *are blocked given* $\mathbf{Z}$*.*

Proposition 32 (Theorem 4.6 in Perković et al., 2017)
Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal MPDAG* $\mathcal{G}$*, where* $\mathbf{Z}$ *is possibly empty. Then* $\mathbf{Z}$ *satisfies the adjustment criterion for MPDAGs relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$ *if and only if* $\mathbf{Z}$ *is a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$*.*

Condition (i) in Definition 31 is also called *amenability relative to* $(\mathbf{X}, \mathbf{Y})$. It has been shown that an MPDAG $\mathcal{G}$ is amenable relative to $(\mathbf{X}, \mathbf{Y})$ if and only if the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ is non-parametrically identified in $\mathcal{G}$, though not necessarily identified by adjustment (Perković, 2020). For singleton X and Y such that $Y \notin \text{pa}(X, \mathcal{G})$, it holds that $\mathcal{G}$ is amenable relative to (X, Y) if and only if there exists a valid adjustment set relative to (X, Y) in $\mathcal{G}$. Further, the set $\text{pa}(X, \mathcal{G})$ is a valid adjustment set in this case (Perković, 2020). This is analogous to DAGs.

Consider again the causal MPDAGs in Figures 10 and 11. The MPDAG $\mathcal{G}_1$ is

amenable relative to (X, Y), as the only possibly causal path from X to Y is the directed path $X \to Y$. The MPDAG $\mathcal{G}_2$ is non-amenable relative to (X, Y), as the possibly causal path $X - A \to Y$ contains an undirected edge between X and A.

4.1.2 Non-amenable MPDAGs — IDA

In a non-amenable MPDAG, where no valid adjustment set exists, it may be of interest to identify and estimate the set of all possible causal effects consistent with the MPDAG. For this aim, Maathuis et al. (2009) devised the IDA algorithm. Conceptually, the idea is to enumerate all DAGs represented by the MPDAG, and determine the set of parents of the treatment in each of them; the set of all possible causal effects is then estimated by adjusting for all possible parent sets in turn. Naively implemented, such an algorithm does not scale well, as the enumeration of DAGs is computationally intensive when there are many undirected edges in the MPDAG. However, Maathuis et al. (2009) showed that the set of possible parent sets can be determined locally without enumerating all DAGs, thus making IDA feasible also for large MPDAGs with many undirected edges.

Maathuis et al. (2009) described IDA for a singleton treatment in a causal CPDAG. Generalisations to a set of treatments and causal MPDAGs were proposed by Nandy et al. (2017) and Perković et al. (2017), respectively. Malinsky and Spirtes (2017) developed a version of IDA that can be used in the presence of latent variables. All of them assumed that the variables represented in the graph follow a multivariate normal distribution, and proposed treatment effect estimators based on linear regression (see also Section 4.3.1 below). However, the part of IDA that determines the adjustment sets does not require parametric assumptions, and can in principle be combined with any estimation method as long as marginal effects are estimated (Witte and Didelez, 2018).

4.2 The forbidden projection and the O-set

In this section, I introduce the forbidden projection, which is a latent projection over the forbidden nodes, save the treatment and outcome nodes. The intuition behind the forbidden projection is that the causal structure among the forbidden nodes is not relevant for choosing an adjustment set, as valid adjustment sets do not contain forbidden nodes, see the adjustment criterion for MPDAGs in Definition 31. The forbidden projection removes the forbidden nodes from the graph, while retaining the causal structure among the remaining nodes.

The focus of the next few sections is on amenable MPDAGs containing a singleton treatment X and a singleton outcome Y of interest. In Paper 2, Witte et al. (2020), in Section 4.5, we define and investigate the forbidden projection for the more general case that $\mathbf{X}$ is a set. Proofs that are given in the publication for the more general case are not repeated here.

Definition 33 (Forbidden projection; Definition 17 in Witte et al., 2020)
Let $\mathcal{G}$ be an MPDAG with node set $\mathbf{V}$, and let X and Y be two nodes in $\mathbf{V}$ such that $\mathcal{G}$ is amenable relative to (X, Y). Define $\mathbf{F} = \text{forb}(X, Y, \mathcal{G}) \setminus \{X, Y\}$. The forbidden projection *$\mathcal{G}^{XY}$ of $\mathcal{G}$ is a graph with node set $\mathbf{V} \setminus \mathbf{F}$ and edges as follows: For distinct nodes $W_i, W_j \in \mathbf{V} \setminus \mathbf{F}$,*

(i) $\mathcal{G}^{XY}$ contains a directed edge $W_i \to W_j$ if and only if $\mathcal{G}$ contains a directed path $W_i \to \cdots \to W_j$ on which all non-endpoint nodes are in $\mathbf{F}$,

(ii) $\mathcal{G}^{XY}$ contains a bi-directed edge $W_i \leftrightarrow W_j$ if and only if $\mathcal{G}$ contains a path, with at least one non-endpoint node, of the form $W_i \leftarrow \cdots \to W_j$ on which all non-endpoints are non-colliders and in $\mathbf{F}$,

(iii) $\mathcal{G}^{XY}$ contains an undirected edge $W_i - W_j$ if and only if $\mathcal{G}$ contains $W_i - W_j$.

See Figure 12 for illustration. The following properties make the forbidden projection a useful simplification of a given causal MPDAG:

Proposition 34 (Proposition 22 in Witte et al., 2020)
Let $\mathcal{G}$ be an MPDAG with node set $\mathbf{V}$ and let X and Y be two nodes in $\mathbf{V}$ such that $\mathcal{G}$ is amenable relative to (X, Y). Denote the set of DAGs represented by $\mathcal{G}$ as $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$. Then the forbidden projection $\mathcal{G}^{XY}$ is the causal MPDAG representing the DAGs in $\{\mathcal{D}_1^{XY}, \mathcal{D}_2^{XY}, \ldots, \mathcal{D}_M^{XY}\}$.

Proposition 34 implies that for singletons X and Y, the forbidden projection does not contain bi-directed edges and has the same causal interpretation as the original graph. In the special case that $\mathcal{G}$ is a DAG, Proposition 34 implies that $\mathcal{G}^{XY}$ is also a DAG. In the forbidden projection, all causal paths from X to Y are collapsed into a single arrow, as shown next.

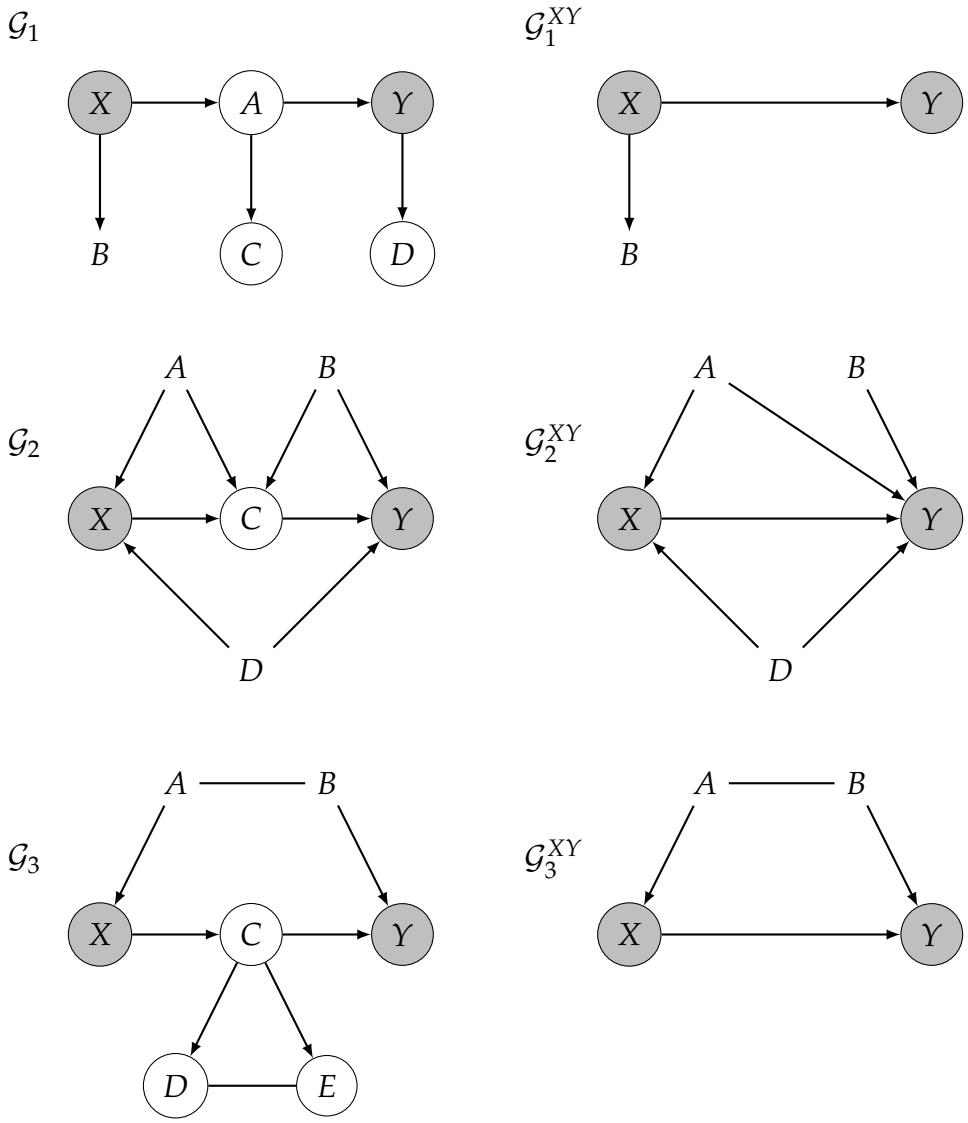


Figure 12: Example causal MPDAGs and their forbidden projections with respect to (X, Y). The forbidden nodes with respect to (X, Y) are shown as circles, treatment and outcome are additionally highlighted in grey.

Proposition 35

Let $\mathcal{G}$ be a causal MPDAG with node set $\mathbf{V}$ and let X and Y be two nodes in $\mathbf{V}$ such that $\mathcal{G}$ is amenable relative to (X, Y), and $Y \in \text{possde}(X, \mathcal{G})$. Then (i) $X \to Y$ is in $\mathcal{G}^{XY}$ and (ii) $X \to Y$ is the only possibly causal path from X to Y in $\mathcal{G}^{XY}$.

Proof. Statement (i) follows immediately from Lemma 18 in Witte et al., 2020, together with Definition 33 of the forbidden projection.

To see that statement (ii) holds, assume for contradiction that there was an additional possibly causal path p from X to Y in $\mathcal{G}^{XY}$. The path p cannot be equal to $X - Y$, as by Definition 33 of the forbidden projection, every undirected edge in $\mathcal{G}^{XY}$ also occurs in $\mathcal{G}$, but $\mathcal{G}$ is amenable relative to (X, Y). Hence, p must

contain at least one non-endpoint node. Denote one such node as A. Then $A \in \text{possde}(X, \mathcal{G}^{XY})$ and $A \in \text{possan}(Y, \mathcal{G}^{XY})$. By Definition 33 of the forbidden projection, $A \in \text{possde}(X, \mathcal{G})$ and $A \in \text{possan}(Y, \mathcal{G})$. However, this implies that $A \in \text{forb}(X, Y, \mathcal{G})$ and thus A is not in $\mathcal{G}^{XY}$, which is a contradiction. □

Proposition 36 (Proposition 25 in Witte et al., 2020)
Let $\mathcal{G}$ be an MPDAG with node set $\mathbf{V}$ and let X and Y be two nodes in $\mathbf{V}$ such that $\mathcal{G}$ is amenable relative to (X, Y). Then a set $\mathbf{Z}$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$ if and only if it is a valid adjustment set relative to (X, Y) in $\mathcal{G}^{XY}$.

Proposition 36 formalises how $\mathcal{G}^{XY}$ retains all information relevant for choosing an adjustment set. The forbidden projection is not useful if other identification strategies (e.g. front-door adjustment, see Pearl, 2009) are of interest.

I now use the forbidden projection in order to define the **O**-set and to prove that it is a valid adjustment set.

Definition 37 (**O**-set)
Let X and Y be two nodes in a causal MPDAG $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to (X, Y). The **O**-set *with respect to X, Y in $\mathcal{G}$ is defined as $\mathbf{O}(X, Y, \mathcal{G}) = \text{pa}(Y, \mathcal{G}^{XY}) \setminus \{X\}$.*

Proposition 38
Let X and Y be two nodes in a causal MPDAG $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to (X, Y). If $Y \in \text{possde}(X, \mathcal{G})$, then $\mathbf{O}(X, Y, \mathcal{G})$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$.

Proof. I show that $\mathbf{O}(X, Y, \mathcal{G})$ satisfies the adjustment criterion for MPDAGs (Definition 31) relative to (X, Y) in $\mathcal{G}^{XY}$. By Proposition 32, this implies that $\mathbf{O}(X, Y, \mathcal{G})$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}^{XY}$, and by Proposition 36, it is also a valid adjustment set relative to (X, Y) in $\mathcal{G}$.

Condition (i) of Definition 31 requires that all possibly causal paths from X to Y in $\mathcal{G}^{XY}$ contain a directed edge out of X, and condition (ii) requires that $\mathbf{O}(X, Y, \mathcal{G}) \cap \text{forb}(X, Y, \mathcal{G}^{XY}) = \emptyset$. By Proposition 35, the only possibly causal path from X to Y in $\mathcal{G}^{XY}$ is $X \to Y$, and $\text{forb}(X, Y, \mathcal{G}) = \{X, Y\}$. By Definition 37 of $\mathbf{O}(X, Y, \mathcal{G})$, $\mathbf{O}(X, Y, \mathcal{G}) \cap \{X, Y\} = \emptyset$. Hence conditions (i) and (ii) are satisfied. Condition (iii) requires that all definite-status non-causal paths from X to Y are blocked

given $\mathbf{O}(X, Y, \mathcal{G})$ in $\mathcal{G}^{XY}$. Since it is assumed that $Y \in \text{possde}(X, \mathcal{G})$, all nodes in $\text{possde}(Y, \mathcal{G})$ are forbidden nodes, hence Y does not have children or siblings in $\mathcal{G}^{XY}$. It follows that all non-causal paths from X to Y in $\mathcal{G}^{XY}$ contain a non-collider in $\text{pa}(Y, \mathcal{G}^{XY})$ and are thus blocked given $\mathbf{O}(X, Y, \mathcal{G})$. Hence, condition (iii) is satisfied as well. □

Note that if $Y \notin \text{possde}(X, \mathcal{G})$, then it can be read off the graph that X has no causal effect on Y. The assumption that $Y \in \text{possde}(X, \mathcal{G})$ does therefore not limit the usefulness of the **O**-set for identifying causal effects.

4.3 Optimal adjustment under Gaussianity

The **O**-set contains all 'direct predictors' of the outcome that are not in the forbidden set. Intuitively, adjusting for the **O**-set e.g. in OLS regression reduces the residual variance of the outcome and therefore yields an efficient estimator of the treatment effect. In this section, I formally prove the optimality of the **O**-set for causal inference using OLS regression.

The graphical setting considered here is the same as in the previous section: I assume that the interest lies in the causal effect of a singleton treatment X on a singleton outcome Y in an amenable MPDAG. I further assume that the variables in the graph jointly follow a multivariate normal distribution and that the target of inference is the average causal effect of X on Y, $\tau = \text{E}(Y \mid do(x+1)) - \text{E}(Y \mid do(x))$, which is constant in x under the multivariate normal assumption (see Section 3.1). The estimated coefficient $\hat{\beta}_{yx.\mathbf{z}}$ of X in a linear regression of Y on X and a valid adjustment set $\mathbf{Z}$ is a consistent, asymptotically normal estimator of τ. In other words, the sequence of estimators $(\hat{\beta}^{(n)}_{yx.\mathbf{z}})_{n \in \mathbb{N}}$, where n denotes the sample size used for the estimation, converges in distribution to a normal distribution with mean $\beta_{yx.\mathbf{z}}$ and *asymptotic variance* $a.var(\hat{\beta}_{yx.\mathbf{z}})$.

Different choices of valid adjustment sets lead to different asymptotic variances, and I will show that the asymptotic variance obtained by adjusting for the **O**-set is at least as small of the asymptotic variance obtained by adjusting for an alternative valid adjustment set. I start by stating relevant properties of the normal distribution.

4.3.1 Properties of the multivariate normal distribution

A vector $\mathbf{V} = (V_1, \ldots, V_K)^T$ of random variables with mean vector $\boldsymbol{\mu} = (\mu_1 \ldots, \mu_K)$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$ follows a multivariate normal or Gaussian distribution if its joint density is

$$f(v_1, \ldots, v_K) = \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{\Sigma})}} \exp\left[-\frac{1}{2}(\mathbf{v} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{v} - \boldsymbol{\mu})\right]$$

(Anderson, 1984, p. 17).

Consider a partition of the vector $\mathbf{V} = (V_1, \ldots, V_K)^T$ into subvectors $\mathbf{S} = (V_1, \ldots, V_l)^T$ and $\mathbf{T} = (V_{l+1}, \ldots, V_K)^T$, for $1 \leq l < K$. I denote the partitioned mean vector by $\boldsymbol{\mu} = (\boldsymbol{\mu}_\mathbf{S}^T, \boldsymbol{\mu}_\mathbf{T}^T)^T$ and the partitioned covariance matrix by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{SS}} & \boldsymbol{\Sigma}_{\mathbf{ST}} \\ \boldsymbol{\Sigma}_{\mathbf{TS}} & \boldsymbol{\Sigma}_{\mathbf{TT}} \end{pmatrix}.$$

If $l = 1$, then $\mathbf{S}$ has only one element, i.e. $\mathbf{S} = V_1$. In this case, $\boldsymbol{\Sigma}_{\mathbf{SS}} = \mathrm{Var}(V_1)$, and $\boldsymbol{\Sigma}_{\mathbf{ST}}$ and $\boldsymbol{\Sigma}_{\mathbf{TS}}$ are vectors.

The family of multivariate normal distributions is closed under marginalisation and conditioning, as formalised in the following two lemmas. Proofs can be found in textbooks on multivariate analysis, e.g. Mardia et al. (1979).

Lemma 39
Let $\mathbf{S}$ *and* $\mathbf{T}$ *be random vectors such that* $(\mathbf{S}^T, \mathbf{T}^T)^T$ *follows a multivariate normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$*. Then* $\mathbf{S}$ *is multivariate normal with mean vector* $\boldsymbol{\mu}_\mathbf{S}$ *and covariance matrix* $\boldsymbol{\Sigma}_{\mathbf{SS}}$*.*

Lemma 40
Let $\mathbf{S}$ *and* $\mathbf{T}$ *be random vectors such that* $(\mathbf{S}^T, \mathbf{T}^T)^T$ *follows a multivariate normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$*. Then the conditional distribution of* $\mathbf{S}$ *given* $\mathbf{T} = t$ *is multivariate normal with mean vector* $\boldsymbol{\mu}_\mathbf{S} + \boldsymbol{\Sigma}_{\mathbf{ST}} \boldsymbol{\Sigma}_{\mathbf{TT}}^{-1} (\mathbf{t} - \boldsymbol{\mu}_\mathbf{T})$ *and covariance matrix* $\boldsymbol{\Sigma}_{\mathbf{S}|\mathbf{T}} = \boldsymbol{\Sigma}_{\mathbf{SS}} - \boldsymbol{\Sigma}_{\mathbf{ST}} \boldsymbol{\Sigma}_{\mathbf{TT}}^{-1} \boldsymbol{\Sigma}_{\mathbf{TS}}$*.*

The matrix $\boldsymbol{\Sigma}_{\mathbf{ST}} \boldsymbol{\Sigma}_{\mathbf{TT}}^{-1}$ is the matrix of regression coefficients of $\mathbf{T}$ on $\mathbf{S}$. I denote the (i, j)-th element of $\boldsymbol{\Sigma}_{\mathbf{ST}} \boldsymbol{\Sigma}_{\mathbf{TT}}^{-1}$ as $\beta_{t_j s_i . \mathbf{s}_{-i}}$, where $\mathbf{S}_{-i} = \mathbf{S} \setminus \{S_i\}$. Thus, $\beta_{t_j s_i . \mathbf{s}_{-i}}$ is the partial regression coefficient of S_i in a linear regression of T_j on $\mathbf{S}$. Note

that the conditional covariance matrix $\mathbf{\Sigma_{S|T}}$ does not depend on the value $\mathbf{t}$ but is a constant. This is a specific property of normal distributions and allows for meaningful comparisons of different conditional variances given random vectors using '$<$' and '$>$' in the proofs below.

Consider now a causal DAG $\mathcal{D} = (\mathbf{V}, \mathbf{E})$ with node set $\mathbf{V}$ such that the set of variables $\mathbf{V}$ follows a multivariate normal distribution compatible with $\mathcal{D}$. If $\mathcal{D}$ is subjected to the forbidden projection, then by Lemma 39, the distribution of the variables represented in the projection is also multivariate normal. In other words, the joint distribution collapses together with the graph.

4.3.2 Optimality in DAGs

I now establish the optimality of the **O**-set in a causal DAG (Proposition 43). The result is based on the following well-known formula for the asymptotic variance of the OLS estimator $\hat{\beta}_{XY.\mathbf{Z}}$:

$$a.var(\hat{\beta}_{XY.\mathbf{Z}}) = \frac{\mathrm{Var}(Y \mid X, \mathbf{Z})}{\mathrm{Var}(X \mid \mathbf{Z})}$$

(see e.g. Goldberger, 1991, p. 272). Further, I will be using that the inverse of a positive definite matrix is itself positive definite (see e.g. Anderson, 1984, p. 584).

Lemma 41
Let $(V, \mathbf{U}, \mathbf{W})$ *be a random vector following a multivariate normal distribution. Then* $\mathrm{Var}(V \mid \mathbf{W}, \mathbf{U}) \leq \mathrm{Var}(V \mid \mathbf{W})$.

Proof. By Lemma 40, the conditional distribution of $(V, \mathbf{U})$ given $\mathbf{W}$ is multivariate normal. Consider the following partition of the covariance matrix $\mathbf{\Sigma}_{V\mathbf{U}|\mathbf{W}}$ of $(V, \mathbf{U})$ given $\mathbf{W}$:

$$\mathbf{\Sigma}_{V\mathbf{U}|\mathbf{W}} = \begin{pmatrix} \mathrm{Var}(V \mid \mathbf{W}) & \mathbf{\Sigma}_{V\mathbf{U}|\mathbf{W}} \\ \mathbf{\Sigma}_{\mathbf{U}V|\mathbf{W}} & \mathbf{\Sigma}_{\mathbf{UU}|\mathbf{W}} \end{pmatrix}$$

By Lemma 39, the conditional distribution of $\mathbf{U}$ given $\mathbf{W}$ is multivariate normal with positive definite covariance matrix $\mathbf{\Sigma}_{\mathbf{UU}|\mathbf{W}}$. Hence, $\mathbf{\Sigma}^{-1}_{\mathbf{UU}|\mathbf{W}}$ is positive definite as well.

Consider now additionally conditioning on $\mathbf{U}$. By Lemma 40, the resulting conditional distribution of V given $(\mathbf{U}, \mathbf{W})$ is normal with variance $\mathrm{Var}(V \mid \mathbf{U}, \mathbf{W}) =$

$\text{Var}(V \mid \mathbf{W}) - \boldsymbol{\Sigma}_{V\mathbf{U}|\mathbf{W}}\boldsymbol{\Sigma}^{-1}_{\mathbf{UU}|\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{U}V|\mathbf{W}}$. As $\boldsymbol{\Sigma}^{-1}_{\mathbf{UU}|\mathbf{W}}$ is positive definite and $\boldsymbol{\Sigma}_{V\mathbf{U}|\mathbf{W}}$ and $\boldsymbol{\Sigma}_{\mathbf{U}V|\mathbf{W}}$ are non-zero vectors such that $\boldsymbol{\Sigma}_{V\mathbf{U}|\mathbf{W}} = \boldsymbol{\Sigma}^{T}_{\mathbf{U}V|\mathbf{W}}$, it follows that $\text{Var}(V \mid \mathbf{W}, \mathbf{U}) \leq \text{Var}(V \mid \mathbf{W})$. □

Lemma 42
Let $(V, \mathbf{U}, \mathbf{W})$ *be a random vector following a multivariate normal distribution such that* $V \perp\!\!\!\perp \mathbf{U} \mid \mathbf{W}$. *Then* $\text{Var}(V \mid \mathbf{W}, \mathbf{U}) = \text{Var}(V \mid \mathbf{W})$.

Proof. This follows immediately from the definition of conditional independence. □

Proposition 43
Let $\mathcal{D} = (\mathbf{V}, \mathbf{E})$ *be a causal DAG such that* $\mathbf{V}$ *follows a multivariate normal distribution compatible with* $\mathcal{D}$. *Let* X *and* Y *be two nodes in* $\mathcal{D}$, *let* $\mathbf{Z}$ *be a valid adjustment set relative to* (X, Y) *in* $\mathcal{D}$, *and let* $\mathbf{O} = \mathbf{O}(X, Y, \mathcal{D})$. *Then* $a.var(\hat{\beta}_{XY.\mathbf{O}}) \leq a.var(\hat{\beta}_{XY.\mathbf{Z}})$.

Proof. The asymptotic variances can be written as

$$a.var(\hat{\beta}_{XY.\mathbf{O}}) = \frac{\text{Var}(Y \mid X, \mathbf{O})}{\text{Var}(X \mid \mathbf{O})}$$

and

$$a.var(\hat{\beta}_{XY.\mathbf{Z}}) = \frac{\text{Var}(Y \mid X, \mathbf{Z})}{\text{Var}(X \mid \mathbf{Z})}.$$

Consider first the numerators. I will show that $\text{Var}(Y \mid X, \mathbf{O}) \overset{(1)}{=} \text{Var}(Y \mid X, \mathbf{O}, \mathbf{Z}) \overset{(2)}{\leq} \text{Var}(Y \mid X, \mathbf{Z})$.

(1) Define $\mathbf{Z}' = \mathbf{Z} \setminus \mathbf{O}$. As $\mathbf{Z}$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$, $X \notin \mathbf{Z}$; hence $X \notin \mathbf{Z}'$. By Definition 37 of the $\mathbf{O}$-set, all nodes in $\mathbf{O}$ are in $\mathcal{D}^{XY}$, and $\mathbf{O} \cup \{X\} = \text{pa}(Y, \mathcal{D}^{XY})$. By Definition 33 of the forbidden projection, $\text{ch}(Y, \mathcal{D}^{XY}) = \emptyset$. Thus, $Y \perp_{\mathcal{D}^{XY}} \mathbf{V} \setminus (\{X\} \cup \mathbf{O}) \mid \{X\} \cup \mathbf{O}$. By the 'decomposition' property of DAG-induced independence models (see Definition 1), $Y \perp_{\mathcal{D}^{XY}} \mathbf{Z}' \mid \{X\} \cup \mathbf{O}$, which by compatibility implies $Y \perp\!\!\!\perp \mathbf{Z}' \mid (X, \mathbf{O})$. By Lemmas 39 and 42, $\text{Var}(Y \mid X, \mathbf{O}, \mathbf{Z}) = \text{Var}(Y \mid X, \mathbf{O})$.

(2) By Lemmas 39 and 41, $\text{Var}(Y \mid X, \mathbf{O}, \mathbf{Z}) \leq \text{Var}(Y \mid X, \mathbf{Z})$.

Consider now the denominators. I will show that $\text{Var}(X \mid \mathbf{Z}) \overset{(3)}{=} \text{Var}(X \mid \mathbf{O}, \mathbf{Z}) \overset{(4)}{\leq} \text{Var}(X \mid \mathbf{O})$.

(3) Define $\mathbf{O}' = \mathbf{O} \setminus \mathbf{Z}$. By Definition 37 of the $\mathbf{O}$-set, all nodes in $\mathbf{O}'$ are in $\mathcal{D}^{XY}$, and $\mathbf{O}' \subseteq \text{pa}(Y, \mathcal{D}^{XY})$. I show that any paths between X and $\mathbf{O}'$ in $\mathcal{D}^{XY}$ are blocked given $\mathbf{Z}$. If there are no paths between X and $\mathbf{O}'$ in $\mathcal{D}^{XY}$, this follows trivially. Suppose, therefore, that at least one such path exists, and denote it as p. There are now two cases. (i) p contains Y. Then Y must be a collider on p, as $\text{ch}(Y, \mathcal{D}^{XY}) = \emptyset$ by Definition 33 of the forbidden projection. Since $\mathbf{Z}$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$, $Y \notin \mathbf{Z}$. Hence, p is blocked given $\mathbf{Z}$ in $\mathcal{D}^{XY}$. (ii) p does not contain Y. Then p does not contain an edge out of X, as by Definition 33 of the forbidden projection, $\text{ch}(X, \mathcal{D}^{XY}) \subseteq \{Y\}$. Hence, p is of the form $X \leftarrow \cdots O$, with $O \in \mathbf{O}'$. As $O \in \text{pa}(Y, \mathcal{D}^{XY})$, there exists a path p' in $\mathcal{D}^{XY}$ constructed by adding the edge $O \to Y$ to p. This path p' is a non-causal path from X to Y in $\mathcal{D}^{XY}$, and is thus blocked given $\mathbf{Z}$ in $\mathcal{D}^{XY}$, see Proposition 36. It follow that p is blocked given $\mathbf{Z}$ in $\mathcal{D}^{XY}$. Thus, all paths between X and $\mathbf{O}'$ in $\mathcal{D}^{XY}$ are blocked given $\mathbf{Z}$ in $\mathcal{D}^{XY}$, which by compatibility implies $X \perp\!\!\!\perp \mathbf{O}' \mid \mathbf{Z}$. By Lemmas 39 and 42, $\text{Var}(X \mid \mathbf{O}, \mathbf{Z}) = \text{Var}(X \mid \mathbf{Z})$.

(4) By Lemmas 39 and 41, $\text{Var}(X \mid \mathbf{O}, \mathbf{Z}) \leq \text{Var}(X \mid \mathbf{O})$.

Together, we have that $\text{Var}(Y \mid X, \mathbf{O}) \leq \text{Var}(Y \mid X, \mathbf{Z})$ and $\text{Var}(X \mid \mathbf{Z}) \leq \text{Var}(X \mid \mathbf{O})$, thus $a.var(\hat{\beta}_{XY.\mathbf{O}}) \leq a.var(\hat{\beta}_{XY.\mathbf{Z}})$. □

A small note is in order here. Many textbooks on regression show that the asymptotic variance of OLS estimators always *increases* when more predictors are added, and *decreases* when predictors are removed, regardless of the correlation structure between the variables in the model. A concise proof is given in Rao (1971). At first glance, these results appear to contradict the results in this thesis. However, Rao and others assume that all predictor variables in the model are *fixed* by design, while I assume them to be random variables. This difference between fixed and random predictors is often neglected in textbooks, as the OLS estimator is consistent for the true coefficients in both cases. However, as seen here, the implications for its asymptotic variance are very different.

4.3.3 Generalisation to amenable MPDAGs

In order to show that the optimality result in Proposition 43 can be generalised to amenable MPDAGs, I use that the $\mathbf{O}$-set relative to (X, Y) in an amenable MPDAG is equal to the $\mathbf{O}$-set relative to (X, Y) in each represented DAG (Lemma 44). I show that the set of alternative valid adjustment sets also coincides in the MPDAG and each represented DAG (Lemma 47). From these two results, the optimality

result for amenable MPDAGs immediately follows (Proposition 48).

Lemma 44 (Lemma E.7 in Henckel et al., 2019)
Let X and Y be two nodes in a causal MPDAG $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to (X, Y), and let $\mathcal{D}$ be a DAG in $[\mathcal{G}]$. Then $\mathbf{O}(X, Y, \mathcal{G}) = \mathbf{O}(X, Y, \mathcal{D})$.

Lemma 45 (Lemma E.8 in Henckel et al., 2019)
Let X and Y be two nodes in a causal MPDAG $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to (X, Y), and let $\mathcal{D}$ be a DAG in $[\mathcal{G}]$. Then $\text{forb}(X, Y, \mathcal{G}) = \text{forb}(X, Y, \mathcal{D})$.

Lemma 46 (Lemma 10 in Perković et al., 2018)
Let $\{X\}$, $\{Y\}$ and $\mathbf{Z}$ be disjoint node sets in a causal MPDAG $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to (X, Y). If $\mathbf{Z} \cap \text{forb}(X, Y, \mathcal{G}) = \emptyset$, then the following statements are equivalent:

(i) all non-causal definite-status paths from X to Y are blocked given $\mathbf{Z}$ in $\mathcal{G}$,

(ii) all non-causal paths from X to Y are blocked given $\mathbf{Z}$ in a DAG $\mathcal{D} \in [\mathcal{G}]$.

Proof. Perković et al. (2018) gave a proof for the case that $\mathcal{G}$ is a CPDAG. As the proof does not use any properties specific to MPDAGs but not CPDAGs, the lemma also holds for MPDAGs. □

Lemma 47
Let X and Y be two nodes in a causal MPDAG $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to (X, Y), and let $\mathcal{D}$ be a DAG in $[\mathcal{G}]$. Then the set of valid adjustment sets relative to (X, Y) in $\mathcal{G}$ is equal to the set of valid adjustment sets relative to (X, Y) in $\mathcal{D}$.

Proof. By definition, any valid adjustment set relative to (X, Y) in $\mathcal{G}$ is a valid adjustment set relative to (X, Y) in all DAGs in $\mathcal{G}$, including $\mathcal{D}$. What needs to be shown is that any valid adjustment set relative to (X, Y) in $\mathcal{D}$ is also a valid adjustment set relative to (X, Y) in $\mathcal{G}$. Consider thus a set $\mathbf{Z}$ that is a valid adjustment set relative to (X, Y) in $\mathcal{D}$. By Proposition 14, $\mathbf{Z}$ satisfies the adjustment criterion for DAGs (Definition 13), i.e. the following holds: (1) $\mathbf{Z} \cap \text{forb}(X, Y, \mathcal{D}) = \emptyset$, (2) all non-causal paths from X to Y in $\mathcal{G}$ are blocked given $\mathbf{Z}$. By (1) and Lemma 45, $\text{forb}(X, Y, \mathcal{D}) = \text{forb}(X, Y, \mathcal{G})$, hence $\mathbf{Z}$ satisfies condition (ii) of the adjustment criterion for MPDAGs (Definition 31) with respect to (X, Y) and $\mathcal{G}$. By (2) and Lemma 46, $\mathbf{Z}$ also satisfies condition (iii) of Definition 31 with respect to (X, Y)

and $\mathcal{G}$. Condition (i) of Definition 31, amenability of $\mathcal{G}$ relative to (X, Y), is satisfied by assumption. Hence, by Proposition 32, $\mathbf{Z}$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$. □

Intuitively, Lemma 47 shows that the undirected edges in amenable MPDAGs can be ignored for the purpose of selecting a valid adjustment set.

Proposition 48
Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a causal MPDAG such that $\mathbf{V}$ follows a multivariate normal distribution compatible with $\mathcal{D}$. Let X and Y be two nodes in $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to (X, Y), let $\mathbf{Z}$ be a valid adjustment set relative to (X, Y) in $\mathcal{G}$, and let $\mathbf{O} = \mathbf{O}(X, Y, \mathcal{G})$. Then $a.var(\hat{\beta}_{XY.\mathbf{O}}) \leq a.var(\hat{\beta}_{XY.\mathbf{Z}})$.

Proof. Pick a DAG in $[\mathcal{G}]$ and denote it as $\mathcal{D}$. By Lemma 47, the set of valid adjustment sets relative to (X, Y) in $\mathcal{G}$ is equal to the set of valid adjustment sets relative to (X, Y) in $\mathcal{D}$. By Lemma 44, $\mathbf{O}(X, Y, \mathcal{G}) = \mathbf{O}(X, Y, \mathcal{D})$. It now follows immediately from the result for DAGs in Proposition 43 that $a.var(\hat{\beta}_{XY.\mathbf{O}}) \leq a.var(\hat{\beta}_{XY.\mathbf{Z}})$. □

4.4 Generalisations and further results

The **O**-set was also described and investigated in Henckel et al. (2019). They defined the **O**-set for sets $\mathbf{X}$ and $\mathbf{Y}$ in terms of the original graph (instead of the forbidden projection): $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \text{pa}(\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{G}), \mathcal{G}) \setminus \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$. As shown in Paper 2 (Section 4.5), this definition is equivalent, for singleton X and Y, to my definition of the **O**-set in Definition 37. Henckel et al. (2019) presented an optimality result similar to my Proposition 48, but for sets $\mathbf{X}$ and $\mathbf{Y}$ and under more general distributional assumptions. In particular, Henckel et al. (2019) assumed that for each variable V_i represented in graph, $V_i = \sum_{V_j \in \text{pa}(V_i)} \alpha_{ij} V_j + \varepsilon_{V_i}$, where α_{ij} is a constant and ε_{V_i} is an error term with mean zero and finite variance. Importantly, the error term is not required to be normally distributed. This model is called the *causal linear model* in Henckel et al. (2019).

Under the same assumptions, Henckel et al. (2019) also devised a *pruning procedure* that reduces a given valid adjustment set $\mathbf{Z}$ to a smaller valid adjustment set $\mathbf{Z}'$ such that $a.var(\hat{\beta}_{XY.\mathbf{Z}'}) \leq a.var(\hat{\beta}_{XY.\mathbf{Z}})$. If $\mathbf{O}(X, Y, \mathcal{G}) \subseteq \mathbf{Z}$, then under faithfulness $\mathbf{Z}' = \mathbf{O}(X, Y, \mathcal{G})$, but $a.var(\hat{\beta}_{XY.\mathbf{Z}'}) \leq a.var(\hat{\beta}_{XY.\mathbf{Z}})$ holds also if $\mathbf{O}(X, Y, \mathcal{G}) \not\subseteq \mathbf{Z}$.

The publication of Henckel et al. (2019) and our Paper 2 (Section 4.5) fostered a number of follow-up papers. Rotnitzky and Smucler (2020) showed that surprisingly, the optimality of the **O**-set also holds in a non-parametric setting when the average causal effect is estimated using a regular asymptotically linear estimator. Estimators in this class include non-parametric outcome regression (Hahn, 1998), non-parametric propensity score weighting (Hirano et al., 2003) and double machine learning (Chernozhukov et al., 2018; Smucler et al., 2019). In addition to adjustment, Rotnitzky and Smucler (2020) considered alternative non-parametric identification strategies, e.g. the front-door strategy (Pearl, 2009). Guo and Perković (2020) investigated estimators based on recursive least squares assuming a causal linear model.

Smucler et al. (2021) and Runge (2021) considered efficient adjustment in the case that the underlying causal DAG includes latent variables. Smucler et al. (2021) gave sufficient conditions under which an optimal adjustment set exists for regular asymptotically linear estimators. Essentially, they assumed that some valid adjustment sets exists, and that the set of observed variables is a subset of the ancestors of the treatment and the outcome, excluding M-structures. The optimal adjustment set can then be determined from a projection graph that can be viewed as an undirected version of the forbidden projection. Runge (2021) presented a necessary and sufficient criterion for the existence of an optimal adjustment set for a class of estimators satisfying certain information-theoretic properties. This includes the OLS estimator and is conjectured to also include the regular asymptotically linear estimators.

An interesting question for future research is whether the optimality of the **O**-set holds for parametric estimators other than the OLS estimator. The simulation results in Witte and Didelez (2018) (Section 3.5) and in other publications, e.g. Brookhart et al. (2006), Austin et al. (2007) and Chatton et al. (2020), suggest that adjustment for the **O**-set might be optimal also for standardised logistic regression and propensity score methods. Hurink (2020) provided further empirical evidence and a partial proof for the optimality of the **O**-set in a standardised logistic regression setting. Diop et al. (2021) consider mediation analysis; from their theoretical results it follows that an optimal adjustment set does not exist for the estimation of the natural direct and indirect effect based on linear regressions.

I conjecture that a generally useful strategy for proving the optimality of the **O**-set in different settings could be as follows, for treatment X, outcome Y, **O**-set $\mathbf{O}$ and valid adjustment set $\mathbf{Z}$: First, compare $\mathbf{Z}$ and $\mathbf{Z} \cup \mathbf{O}$ — both are valid adjustment sets relative to (X, Y), and $\mathbf{O}' = \mathbf{O} \setminus \mathbf{Z}$ satisfies what is sometimes called the *ex-*

clusion restriction (Hahn, 2004) on the treatment model, i.e. $X \perp\!\!\!\perp \mathbf{O}' \mid \mathbf{Z}$ (see the arguments in Proposition 43). It needs to be shown that adding $\mathbf{O}'$ to $\mathbf{Z}$ does not increase the asymptotic variance of the considered estimator. Then, compare $\mathbf{Z} \cup \mathbf{O}$ and $\mathbf{O}$ — again, both are valid adjustment sets relative to (X, Y), and $\mathbf{Z}' = \mathbf{Z} \setminus \mathbf{O}$ satisfies the exclusion restriction on the outcome model, i.e. $Y \perp\!\!\!\perp \mathbf{Z}' \mid \mathbf{O}$ (see the arguments in Proposition 43). Here is needs to be shown that removing $\mathbf{Z}'$ from $\mathbf{Z} \cup \mathbf{O}$ does not increase the asymptotic variance of the estimator. The exclusion restriction has been proven useful for investigations into estimation efficiency in Hahn (2004) and Lunceford and Davidian (2004).

4.5 Paper 2: *Witte, Henckel, Maathuis and Didelez (2020)*

Witte et al. (2020) builds on results as stated in Henckel et al. (2019) and Rotnitzky and Smucler (2020). It makes three contributions to the literature: First, we present the forbidden projection and the associated intuitive definition of the **O**-set given above. Second, we propose to adjust for the **O**-set instead of the parents of the treatment in the IDA algorithm, and we show that this can be done semi-locally. This extends the applicability of the **O**-set to non-amenable MPDAGs. Third, we point out that backward regression selection (see Procedure 5A) can be viewed as an implementation of the pruning procedure proposed in Henckel et al. (2019), thereby drawing the link between the graphically defined **O**-set and data-driven confounder selection.

The notation in Witte et al. (2020) deviates from the notation used in this frame text as follows: The forbidden-projection definition of the **O**-set is denoted with $\mathbf{O}^*$ in the paper, and the definition by Henckel et al. (2019) with **O**. We call the interventional distribution *post-intervention distribution* and use *maxPDAG* instead of MPDAG as an abbreviation for 'maximally oriented partially directed acyclic graph'.

Own contributions

The forbidden projection, optimal IDA and the link between **O**-set and regression selection were my own ideas. I co-designed the simulation study and took the lead in programming. I wrote the first draft of the manuscript, including all proofs, and led the revision process.

Journal of Machine Learning Research 21 (2020) 1-45 Submitted 2/20; Revised 11/20; Published 12/20

On Efficient Adjustment in Causal Graphs

Janine Witte WITTE@LEIBNIZ-BIPS.DE
Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany
and Faculty of Mathematics and Computer Science, University of Bremen, Germany

Leonard Henckel HENCKEL@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zurich, Switzerland

Marloes H. Maathuis MAATHUIS@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zurich, Switzerland

Vanessa Didelez DIDELEZ@LEIBNIZ-BIPS.DE
Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany
and Faculty of Mathematics and Computer Science, University of Bremen, Germany

Editor: Peter Spirtes

Abstract

We consider estimation of a total causal effect from observational data via covariate adjustment. Ideally, adjustment sets are selected based on a given causal graph, reflecting knowledge of the underlying causal structure. Valid adjustment sets are, however, not unique. Recent research has introduced a graphical criterion for an 'optimal' valid adjustment set (**O**-set). For a given graph, adjustment by the **O**-set yields the smallest asymptotic variance compared to other adjustment sets in certain parametric and non-parametric models. In this paper, we provide three new results on the **O**-set. First, we give a novel, more intuitive graphical characterisation: We show that the **O**-set is the parent set of the outcome node(s) in a suitable latent projection graph, which we call the forbidden projection. An important property is that the forbidden projection preserves all information relevant to total causal effect estimation via covariate adjustment, making it a useful methodological tool in its own right. Second, we extend the existing IDA algorithm to use the **O**-set, and argue that the algorithm remains semi-local. This is implemented in the `R`-package `pcalg`. Third, we present assumptions under which the **O**-set can be viewed as the target set of popular non-graphical variable selection algorithms such as stepwise backward selection.

Keywords: causal discovery, causal inference, confounder selection, confounding, efficiency, graphical models, IDA algorithm, model selection, sufficient adjustment set

1. Introduction

In typical analyses of observational data, we wish to estimate the total causal effect of a (possibly multivariate) treatment or exposure $\mathbf{X}$ on a (possibly multivariate) outcome $\mathbf{Y}$. Ideally, we can fully specify the underlying causal directed acyclic graph (DAG). We can then use a graphical adjustment criterion, e.g. Pearl's back-door criterion (Pearl, 2009) or the generalised adjustment criterion (Perković et al., 2015, 2018; Shpitser et al., 2010), to check whether a set of covariates is valid for adjustment. However, there may be more than one valid adjustment set. Although all resulting estimators are then consistent, their variances may differ considerably.

There are several approaches to choose an adjustment set among all valid adjustment sets. For example, one can pick a minimal adjustment set (de Luna et al., 2011; Textor and Liśkiewicz, 2011). An alternative strategy is to aim at decreasing the causal effect estimator's variance by including variables associated with the outcome (e.g. Brookhart et al., 2006; Lunceford and Davidian, 2004; Shortreed and Ertefaie, 2017). Witte and Didelez (2019) referred to this strategy as the 'outcome-oriented' approach. It is especially popular when little graphical knowledge is available. A major advancement for the outcome-oriented approach was the graphical characterisation of the 'optimal' adjustment set (**O**-set) by Henckel et al. (2019) (HPM19). They showed that under a linear model, adjusting for the **O**-set yields the smallest asymptotic variance for the causal effect estimator compared to all other valid adjustment sets, under assumptions detailed below. Strengthening this result, Rotnitzky and Smucler (2020) (RS20) recently showed that the minimal variance property of the **O**-set is retained for a class of non-parametric estimators. All these results apply to DAGs, as well as so-called amenable completed partially directed acyclic graphs (CPDAGs; see e.g. Andersson et al., 1997) and amenable maximally oriented partially directed acyclic graphs (maxPDAGs; see Perković et al., 2017). These are larger classes of graphs allowing for undirected edges where the direction cannot be decided. Amenability implies that despite the undirected edges, an adjustment set can be identified from the CPDAG (or maxPDAG) so that this set is valid for adjustment in all DAGs in the equivalence class. If a CPDAG (or maxPDAG) is not amenable, no common adjustment set for all DAGs in the equivalence class exists (Perković et al., 2018), and hence different DAGs may imply different true causal effects of $\mathbf{X}$ on $\mathbf{Y}$. However, it is then still possible to estimate a multiset of possible causal effects (meaning that all effects in the multiset are compatible with the non-amenable graph) using the IDA algorithm by Maathuis et al. (2009, 2010).

In this paper, we provide three new results on efficient causal effect estimation. First, after briefly reviewing the results of HPM19 and RS20 (Section 2), we provide an alternative, intuitive characterisation of the **O**-set. This is based on the new concept of a forbidden projection, which has many interesting properties regarding adjustment for confounding (Section 3). Second, we extend the application of the **O**-set to non-amenable CPDAGs and maxPDAGs, by incorporating optimal adjustment into the IDA algorithm (Section 4). Third, we discuss how and under what assumptions the **O**-set can be viewed as the target set of data-driven variable selection methods such as backward model selection (Section 5).

2. Optimal Adjustment for Known Causal Structure

We begin by clarifying our setting and defining the **O**-set, before proposing an alternative definition in Section 3. We defer most of the terminology and formal definitions to Appendix A; here we only state some key concepts.

(Possibly) causal nodes and forbidden nodes. Let $\mathcal{G}$ be a causal DAG, CPDAG or maxPDAG. A path $(V_1, \dots, V_m)$ in $\mathcal{G}$ is called *causal* from V_1 to V_m if $V_i \to V_{i+1}$ for all $i \in \{1, \dots, m-1\}$. It is called *possibly causal* if there are no $i, j \in \{1, \dots, m\}$, $i < j$, such that $V_i \leftarrow V_j$. Otherwise it is called *non-causal* from V_1 to V_m. A path from $\mathbf{X}$ to $\mathbf{Y}$ is *proper* if only its first node is in $\mathbf{X}$. If there is a causal path from V_1 to V_m in $\mathcal{G}$, then V_m is called a *descendant* of V_1 in $\mathcal{G}$. Analogously, if there is a possibly causal path from V_1 to V_m in $\mathcal{G}$, then V_m is called a *possible descendant* of V_1 in $\mathcal{G}$. The set of all

descendants of V_1 in $\mathcal{G}$ is denoted by $\mathrm{de}(V_1, \mathcal{G})$, and the set of all possible descendants by $\mathrm{possde}(V_1, \mathcal{G})$. The *causal nodes* with respect to ($\mathbf{X}$, $\mathbf{Y}$) in $\mathcal{G}$, denoted by $\mathrm{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$, are the nodes on proper causal paths from $\mathbf{X}$ to $\mathbf{Y}$, excluding $\mathbf{X}$ itself. The *possibly causal nodes* $\mathrm{posscn}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$ are defined analogously. The *forbidden set* with respect to ($\mathbf{X}$, $\mathbf{Y}$) and $\mathcal{G}$ is defined as $\mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \mathrm{possde}(\mathrm{posscn}(\mathbf{X}, \mathbf{Y}, \mathcal{G}), \mathcal{G}) \cup \mathbf{X}$. In a DAG, this simplifies to $\mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \mathrm{de}(\mathrm{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{G}), \mathcal{G}) \cup \mathbf{X}$. The nodes in the forbidden set are called *forbidden nodes*. It can be shown that valid adjustment sets never contain forbidden nodes (Perković et al., 2018).

Valid adjustment sets. We consider a set of treatments $\mathbf{X}$ and a set of outcomes $\mathbf{Y}$. A (possibly empty) set $\mathbf{Z}$ is a *valid adjustment set* relative to $(\mathbf{X}, \mathbf{Y})$ if the interventional distribution $f(\mathbf{y} \mid do(\mathbf{x}))$ of $\mathbf{Y}$, given we set $\mathbf{X}$ to $\mathbf{x}$ by intervention, factorises as follows:

$$f(\mathbf{y} \mid do(\mathbf{x})) = \begin{cases} f(\mathbf{y} \mid \mathbf{x}) & \text{if } \mathbf{Z} = \emptyset, \\ \int_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) f(\mathbf{z}) d\mathbf{z} & \text{otherwise.} \end{cases}$$

Valid adjustment sets can be read off from a given causal DAG, CPDAG or maxPDAG $\mathcal{G}$ using the *generalised adjustment criterion* (Perković et al., 2017, 2018; Shpitser et al., 2010), which generalises Pearl's back-door criterion (Pearl, 2009): $\mathbf{Z}$ is a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if and only if the following three conditions hold: (a) every proper possibly causal path from ($\mathbf{X}$ to $\mathbf{Y}$) starts with a directed edge out of $\mathbf{X}$, (b) $\mathbf{Z} \cap \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \emptyset$, (c) all proper non-causal definite-status paths from $\mathbf{X}$ to $\mathbf{Y}$ are blocked by $\mathbf{Z}$. Property (a) is called *amenability*. See Appendix A for the definition of a definite-status path. In a DAG, all paths are of definite status.

We consider two model classes and corresponding strategies for estimating causal effects when a valid adjustment set is available: (i) the causal linear model with possibly non-Gaussian error terms, where causal effects are estimated using linear regression (HPM19), and (ii) the more general non-parametric causal model, where estimation proceeds non-parametrically (RS20). In both settings, we assume an underlying causal DAG, and that we observe all variables displayed as nodes in the DAG, i.e. there are no latent variables.

Causal linear models (HPM19). A causal linear model is a causal DAG where every edge represents a linear causal effect. In a causal linear model, the (joint) causal effect of $\mathbf{X} = \{X_1, \ldots, X_{k_x}\}$ on $\mathbf{Y} = \{Y_1, \ldots, Y_{k_y}\}$ is defined as the matrix $\boldsymbol{\tau}_{\mathbf{yx}}$ with elements

$$\begin{aligned} (\boldsymbol{\tau}_{\mathbf{yx}})_{j,i} &= \frac{\partial}{\partial x_i} \mathrm{E}(Y_j \mid do(x_1, \ldots, x_{k_x})) \\ &= \mathrm{E}(Y_j \mid do(x_1, \ldots, x_i + 1, \ldots, x_{k_x})) - \mathrm{E}(Y_j \mid do(x_1, \ldots, x_{k_x})), \end{aligned}$$

where element $(\boldsymbol{\tau}_{\mathbf{yx}})_{j,i}$ corresponds to the controlled direct effect (Robins and Greenland, 1992; Pearl, 2001) of X_i on Y_j relative to $\mathbf{X}$. In other words, $(\boldsymbol{\tau}_{\mathbf{yx}})_{j,i}$ is the difference in $\mathrm{E}(Y_j)$ when $\mathbf{X}$ is set to $(x_1, \cdots, x_i + 1, \ldots, x_{k_x})$ by intervention, compared to when $\mathbf{X}$ is set to $(x_1, \ldots, x_{k_x})$ by intervention. We can compute the effect of more general interventions as functions of the elements of $\boldsymbol{\tau}_{\mathbf{yx}}$; for example, the sum of the first row corresponds to the effect on Y_1 of increasing all elements of $(x_1, \ldots, x_{k_x})$ by one. Given a valid adjustment set $\mathbf{Z}$ for the effect of $\mathbf{X}$ on $\mathbf{Y}$, $\boldsymbol{\tau}_{\mathbf{yx}}$ can be rewritten as a matrix of regression coefficients as follows: Denote by $\boldsymbol{\beta}_{\mathbf{yx.z}}$ the $(k_y \times k_x)$-matrix whose (j, i)-th element is the regression coefficient $\beta_{y_j x_i . \mathbf{x}_{-\mathbf{i}} \mathbf{z}}$ of X_i in a linear regression of Y_j on X_i and $\mathbf{Z} \cup \mathbf{X}_{-i}$, where $\mathbf{X}_{-i} = \mathbf{X} \setminus \{X_i\}$. Then

$\boldsymbol{\tau}_{\mathbf{yx}} = \boldsymbol{\beta}_{\mathbf{yx.z}}$. The ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{\mathbf{yx.z}}$ is a consistent estimator of $\boldsymbol{\beta}_{\mathbf{yx.z}}$. We denote the asymptotic variance of $\hat{\beta}_{y_j x_i . \mathbf{x_{-i} z}}$ by $a.var(\hat{\beta}_{y_j x_i . \mathbf{x_{-i} z}})$.

Non-parametric estimation of causal effects (RS20). In the more general setting of a causal DAG without linearity or other assumptions on the functional form, we define the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ as follows. Let $\mathcal{X}$ be the set of values that $\mathbf{X}$ can take. For a pair of vectors $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, the causal effect of intervening to set $\mathbf{X}$ to $\mathbf{x}$ vs. $\mathbf{x}'$ is the vector $\boldsymbol{\Delta}_{\mathbf{yxx'}}$ with elements

$$(\boldsymbol{\Delta}_{\mathbf{yxx'}})_j = \mathrm{E}(Y_j \mid do(\mathbf{x})) - \mathrm{E}(Y_j \mid do(\mathbf{x}')).$$

Note that in the non-parametric case, it is not possible to compactly represent the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ in a $(k_y \times k_x)$-matrix. RS20 considered the class of regular asymptotically linear estimators for the non-parametric estimation of $\boldsymbol{\Delta}_{\mathbf{yxx'}}$. This class includes inverse probability weighting by a non-parametrically estimated propensity score (Hirano et al., 2003), non-parametric outcome regression (Hahn, 1998), and double machine learning (Chernozhukov et al., 2018; Smucler et al., 2019). We use $\hat{\boldsymbol{\Delta}}_{\mathbf{yxx'.z}}$ to denote an estimator from this class that estimates $\boldsymbol{\Delta}_{\mathbf{yxx'}}$ adjusting for a valid adjustment set $\mathbf{Z}$. Under a causal DAG model and certain smoothness and complexity restrictions, $\hat{\boldsymbol{\Delta}}_{\mathbf{yxx'.z}}$ is a consistent estimator of $\boldsymbol{\Delta}_{\mathbf{yxx'}}$. For given $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{x}'$, the asymptotic distribution of estimators from this class depends only on $\mathbf{Z}$, therefore we do not further distinguish between the estimators. We denote the asymptotic variance of $(\hat{\boldsymbol{\Delta}}_{\mathbf{yxx'.z}})_j = \hat{\Delta}_{y_j \mathbf{x x'.z}}$ by $a.var(\hat{\Delta}_{y_j \mathbf{x x'.z}})$. See RS20 and the references therein for more details on regular asymptotically linear estimators.

Definition 1 (O-set; HPM19 Definition 3.8) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint node sets in a DAG, CPDAG or maxPDAG* $\mathcal{G}$*. Then* $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$ *is defined as:*

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

An example is given in Figure 1. It shows the causal relations between 12 symptoms of prodromal schizophrenia as measured by the *Schizotypic Syndrome Questionnaire* (van Kampen, 2006). The DAG was constructed using a combination of expert knowledge and data-driven structure learning (van Kampen, 2014). For illustration, we here take this given DAG as ground truth. Suppose we are interested in the causal effect of Alienation (ALN) on Delusional Thinking (DET). The bold edges indicate the causal paths with causal nodes {PER, SUS, FTW, DET} (circles). The parents of the causal nodes are {ALN, PER, SUS, FTW, AIS, CDR}, the forbidden set is {ALN, PER, SUS, FTW, DET, HOS, EGC} and the **O**-set is {ALN, PER, SUS, FTW, AIS, CDR} \ {ALN, PER, SUS, FTW, DET, HOS, EGC}={AIS, CDR} (shown in boxes). Other valid adjustment sets are, for example, {AFF, SAN}, {AIS, CDR, AFF} and {AFF, APA, AIS, CDR, SAN}. This can be checked using the generalised adjustment criterion stated above.

Note that in many applications it might be possible to augment a causal graph e.g. with further parents of Y that are marginally independent of all other non-descendants of Y. This induces a different **O**-set illustrating that this set depends on what variables are included in the graph. Note also that the **O**-set is defined even if no valid adjustment set exists, but this case will rarely be of interest.

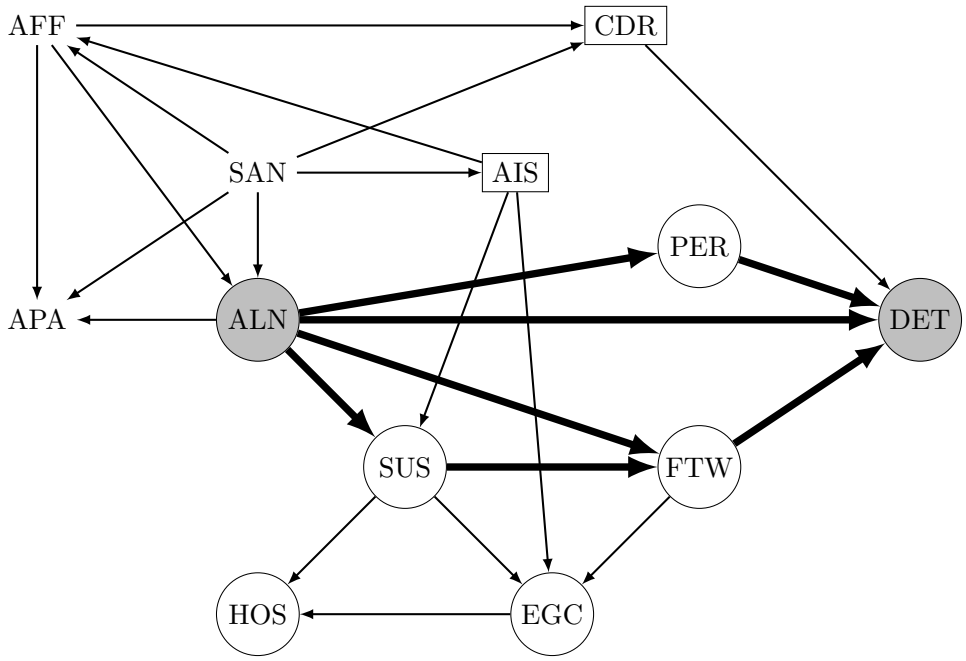


Figure 1: DAG from van Kampen (2014) illustrating the assumed causal relations between 12 prodromal symptoms of schizophrenia: AFF=Affective Flattening, AIS=Active Isolation, ALN=Alienation, APA=Apathy, CDR=Cognitive Derailment, DET=Delusional Thinking, EGC=Egocentrism, FTW=Living in a Fantasy World, HOS=Hostility, PER=Perceptual Aberrations, SAN=Social Anxiety, SUS=Suspiciousness. We are interested in the causal effect of ALN on DET, both shown in grey circles. Bold arrows show the causal paths from ALN to DET. The forbidden nodes are shown as circles, nodes in the **O**-set are shown as boxes.

Proposition 2 (HPM19 Theorem 3.10 (1)) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint subsets of the node set* $\mathbf{V}$ *of a causal DAG, CPDAG or maxPDAG* $\mathcal{G}$*. The set* $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$ *is a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$ *if (i)* $\mathbf{Y} \subseteq \text{possde}(\mathbf{X}, \mathcal{G})$ *and (ii) a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$ *exists.*

Condition (i) can be checked using a simple query on $\mathcal{G}$. If $\mathbf{Y} \not\subseteq \text{possde}(\mathbf{X}, \mathcal{G})$, we know that the causal effect of $\mathbf{X}$ on $\mathbf{Y} \setminus \text{possde}(\mathbf{X}, \mathcal{G})$ is zero. Hence, without loss of generality, we can consider the set of outcome variables $\mathbf{Y} \cap \text{possde}(\mathbf{X}, \mathcal{G})$ instead of $\mathbf{Y}$. Condition (ii) is satisfied if $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$ or any other subset of $\mathbf{V} \setminus \{\mathbf{X}, \mathbf{Y}\}$ fulfils the generalised adjustment criterion stated above. For the DAG in Figure 1, it can easily be seen that DET $\in$ de(ALN), hence condition (i) is satisfied. Under condition (i), condition (ii) is always satisfied for univariate treatment and outcome in a DAG, because the parents of treatment then form a valid adjustment set (see Pearl, 2009, p. 72f.).

The following proposition, which builds on earlier work by Kuroki and Miyakawa (2003) and Kuroki and Cai (2004), establishes the optimality of the **O**-set in terms of the asymptotic variance in the linear and in the non-parametric setting.

Proposition 3 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint subsets of the node set* $\mathbf{V}$ *of a causal DAG, CPDAG or maxPDAG* $\mathcal{G}$*, such that* $\mathbf{Y} \subseteq \text{possde}(\mathbf{X}, \mathcal{G})$*. Let* $\mathbf{Z}$ *be a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$ *and let* $\mathbf{O} = \mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$*.*

(a) (HPM19 Theorem 3.10 (2)) If the variables $\mathbf{V}$ *follow a linear causal model compatible with* $\mathcal{G}$*, then, for every* $X_i \in \mathbf{X}$ *and* $Y_j \in \mathbf{Y}$*,* $a.var(\hat{\beta}_{y_j x_i . \mathbf{x}_{-i}\mathbf{o}}) \leq a.var(\hat{\beta}_{y_j x_i . \mathbf{x}_{-i}\mathbf{z}})$*.*

(b) (RS20 Theorem 2) For every $Y_j \in \mathbf{Y}$ *and pair of vectors* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*,* $a.var(\hat{\Delta}_{y_j \mathbf{x}\mathbf{x}' . \mathbf{o}}) \leq a.var(\hat{\Delta}_{y_j \mathbf{x}\mathbf{x}' . \mathbf{z}})$*.*

In other words, for a given causal linear model, the **O**-set yields the smallest asymptotic variance for the OLS estimator among all valid adjustment sets. If linearity cannot be assumed, the **O**-set yields the smallest variance for regular asymptotically linear estimators. Thus, assume that Figure 1 represents a causal linear model. Proposition 3 then implies that if we estimate the effect of ALN on DET by linearly regressing DET on ALN and the **O**-set {AIS, CDR}, then the estimator will have a smaller asymptotic variance than if we regress DET on ALN and a different valid adjustment set, say the parent set of ALN, which equals {AFF, SAN}. Moreover, when we relax linearity, non-parametric adjustment for the **O**-set {AIS, CDR} is more efficient than non-parametric adjustment for any other valid adjustment set, provided the estimator is in the class of regular asymptotically linear estimators.

3. The O-Set via Forbidden Projection

In this section we provide an alternative, intuitive construction of the **O**-set. For the sake of clarity, we restrict ourselves to DAGs; generalisations to amenable maxPDAGs are given in Appendix C.

To motivate our alternative construction, we posit that a useful adjustment set should be i) valid, ii) easy to compute, and iii) efficient. Consider singleton treatment X and outcome Y, where the latter is not an ancestor of X. The parents of X are easy to determine and guaranteed to be valid (see Pearl, 2009, p. 72f.). However, it is well-known that adjusting for variables strongly associated with treatment tends to reduce the efficiency of OLS and other estimators of the treatment effect. Hence, adjusting for the parents of treatment is typically inefficient compared to other valid adjustment sets. In contrast, it is also well-known that regression adjustment for variables strongly associated with the outcome tends to improve the efficiency of OLS and other estimators. Hence, the parents of the outcome would appear a natural, easy to determine and more efficient alternative for adjustment. However, the parents of Y are not guaranteed to be a valid adjustment set; they may contain forbidden nodes, specifically mediators between treatment and outcome. For example, in Figure 1, FTW is a parent of the outcome DET, but a descendant of the treatment ALN and hence cannot be used for adjustment. Simply omitting such nodes from the parents of Y does not generally lead to a valid adjustment set either. For example, CDR alone does

not form a valid adjustment set in Figure 1, since there are open confounding paths, e.g. $ALN \leftarrow SAN \rightarrow AIS \rightarrow SUS \rightarrow FTW \rightarrow DET$.

Nonetheless, the intuition of using the parents of Y is correct if applied to a modified graph. As we show below, marginalising out, i.e. projecting over, the forbidden nodes results in a graph where the parent set of Y indeed coincides with the **O**-set, and is thus guaranteed to yield an estimator with minimal asymptotic variance in the settings we consider, see Proposition 3. This characterization of the **O**-set thus combines validity, graphical simplicity and efficiency. We will now explain this formally.

Consider again the case of a DAG $\mathcal{D}$ containing sets $\mathbf{X}$ and $\mathbf{Y}$. We first need the concept of latent projection, used to *marginalise* or *collapse over* latent, i.e. unobserved nodes, while preserving the remaining causal relations and (in)dependencies between the observed nodes.

Definition 4 (Latent projection; Verma and Pearl, 1990; Shpitser et al., 2014) *Let $\mathcal{D}$ be a DAG with node set $\mathbf{W} \cup \mathbf{L}$ and $\mathbf{W} \cap \mathbf{L} = \emptyset$. The* latent projection $\mathcal{D}(\mathbf{W})$ over $\mathbf{L}$ on $\mathbf{W}$ *is a graph with node set $\mathbf{W}$ and edges as follows: For distinct nodes $W_i, W_j \in \mathbf{W}$,*

1. *$\mathcal{D}(\mathbf{W})$ contains a directed edge $W_i \rightarrow W_j$ if and only if $\mathcal{D}$ contains a directed path $W_i \rightarrow \cdots \rightarrow W_j$ on which all non-endpoint nodes are in $\mathbf{L}$,*

2. *$\mathcal{D}(\mathbf{W})$ contains a bi-directed edge $W_i \leftrightarrow W_j$ if and only if $\mathcal{D}$ contains a path, with at least one non-endpoint node, of the form $W_i \leftarrow \cdots \rightarrow W_j$ on which all non-endpoint nodes are non-colliders and in $\mathbf{L}$.*

In the latent projection $\mathcal{D}(\mathbf{W})$, two nodes may be connected by a directed and a bi-directed edge at the same time. (In)dependence relations can be read off from a latent projection using the m-separation criterion (Richardson, 2003). For disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subset \mathbf{W}$, $\mathbf{A}$ and $\mathbf{B}$ are d-separated given $\mathbf{C}$ in $\mathcal{D}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are m-separated given $\mathbf{C}$ in $\mathcal{D}(\mathbf{W})$ (Richardson et al., 2017).

For our definition of the **O**-set, we project over the forbidden nodes, save $\mathbf{X}$ and $\mathbf{Y}$, which motivates the following definition:

Definition 5 (Forbidden projection) *Let $\mathcal{D}$ be a DAG with node set $\mathbf{V}$ and let $\mathbf{X}$ and $\mathbf{Y}$ be disjoint subsets of $\mathbf{V}$. We call the graph $\mathcal{D}^{\mathbf{XY}} = \mathcal{D}((\mathbf{V} \setminus \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})) \cup \mathbf{X} \cup \mathbf{Y})$ the* forbidden projection *of $\mathcal{D}$ with respect to $(\mathbf{X}, \mathbf{Y})$.*

Figures 2 and 3 show some examples, where the forbidden nodes are shown as circles. In panels **A** and **B** of Figure 2, the forbidden sets only contain nodes in $\mathbf{X} \cup \mathbf{Y}$, hence nothing is projected over. Panels **E** and **F** show DAGs where the forbidden projection has bi-directed edges, which will become relevant in Proposition 6.

While we primarily introduce the forbidden projection to provide an alternative characterisation of the **O**-set, it is a useful tool in its own right. In particular, as we show next, the forbidden projection of a causal DAG preserves all information relevant to the estimation of a causal effect via adjustment. All proofs are given in Appendix B and generalised to maxPDAGs in Appendix C.

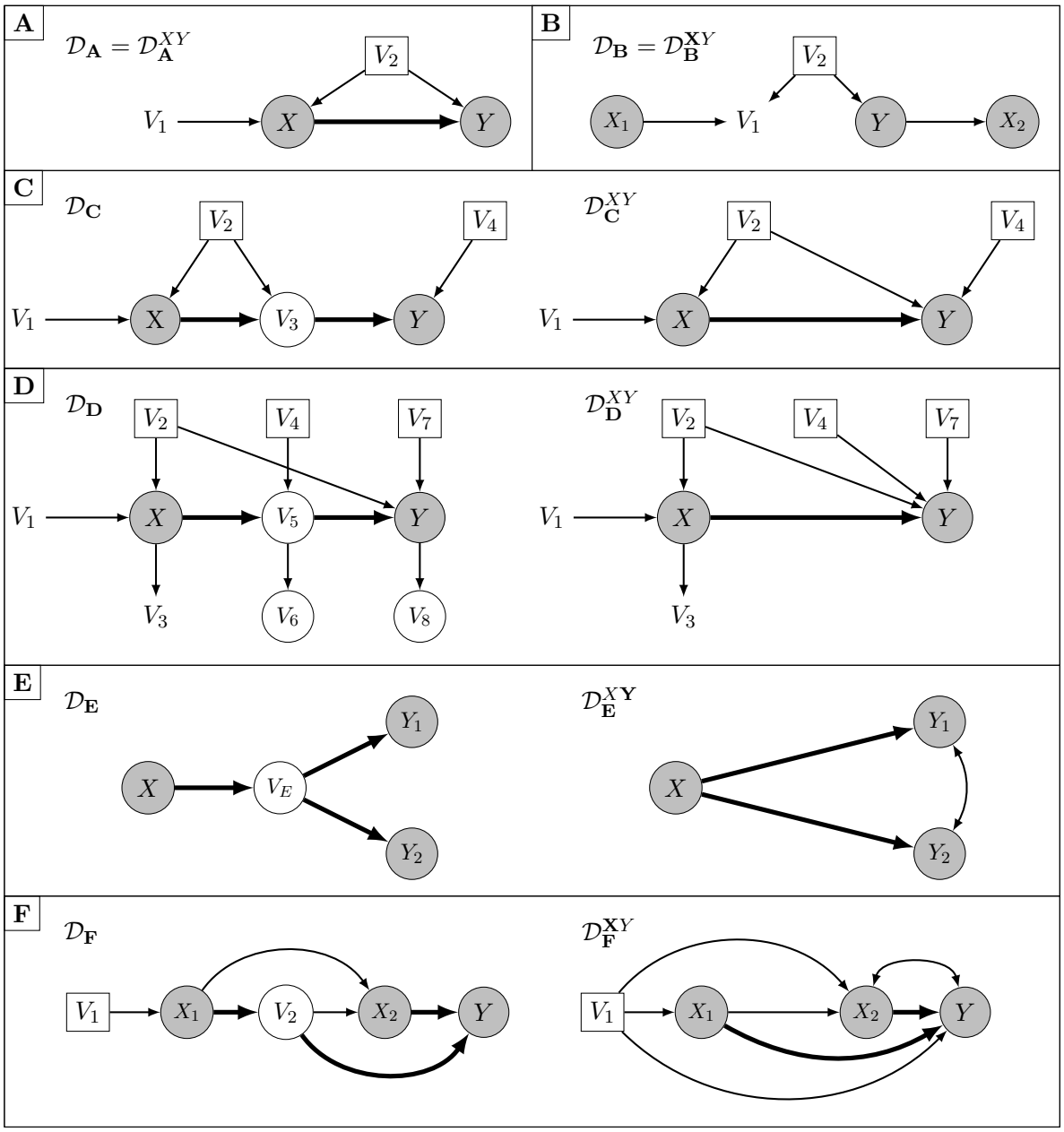


Figure 2: Example DAGs with their forbidden projections. The forbidden nodes are shown as circles, nodes in the **O**-set are shown as boxes. The bold arrows show the causal paths from **X** to **Y**. In panels **A** and **B**, the original DAGs and their forbidden projections are identical. In panel **E**, the empty set is a valid adjustment set and also the **O**-set. In panel **F**, the bi-directed edge between X_2 and Y indicates that the effect of $\mathbf{X} = \{X_1, X_2\}$ on Y is not identified via adjustment.

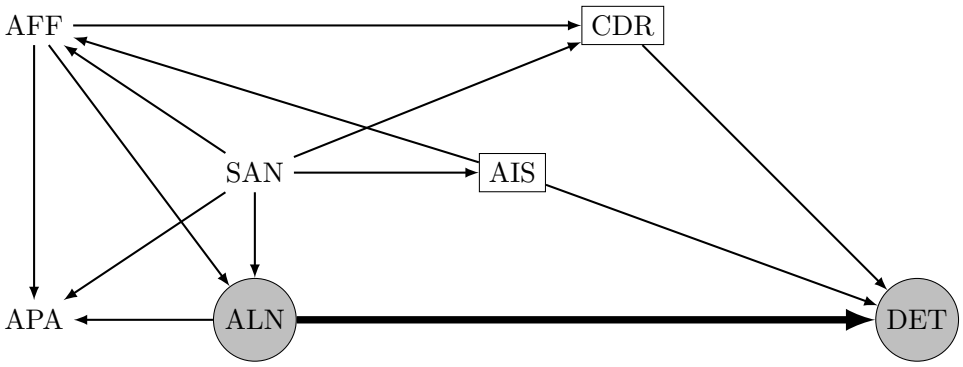


Figure 3: Forbidden projection of the DAG in Figure 1 with respect to $\mathbf{X} = \{\text{ALN}\}$ and $\mathbf{Y} = \{\text{DET}\}$. The forbidden nodes are shown as circles, nodes in the $\mathbf{O}$-set (parents of DET) are shown as boxes. The bold arrow shows the causal path from ALN to DET.

First, the forbidden projection can be used to check whether a valid adjustment set exists relative to given sets of nodes $\mathbf{X}$ and $\mathbf{Y}$:

Proposition 6 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$ *such that* $\mathbf{Y} \subseteq \text{de}(\mathbf{X}, \mathcal{D})$. *Then a valid adjustment set relative to* $(\mathbf{X},\mathbf{Y})$ *in* $\mathcal{D}$ *exists if and only if there is no bi-directed edge between any* $X \in \mathbf{X}$ *and* $Y \in \mathbf{Y}$ *in* $\mathcal{D}^{\mathbf{XY}}$.

In Figure 2, valid adjustment sets with respect to $\mathbf{X}$ and $\mathbf{Y}$ exist in all panels except for panel **F**. The effect of $\mathbf{X} = \{X_1, X_2\}$ on Y in panel **F** is, however, identified e.g. by the more general G-formula (Robins, 1986; Dawid and Didelez, 2010), the algorithm in Tian and Pearl (2003), or the methods in Nandy et al. (2017). See Guo and Perković (2020) and RS20 for results on efficient adjustment in the linear and non-parametric case, respectively. The bi-directed edge between Y_1 and Y_2 in panel **E** has no relevance in defining or determining a valid adjustment set.

For singleton Y such that a valid adjustment set with respect to $(\mathbf{X}, Y)$ exists, the forbidden projection is particularly easy to interpret, as it is itself a causal DAG.

Proposition 7 *Let* $\mathbf{X}$ *and* $\{Y\}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$ *such that* $Y \in \text{de}(\mathbf{X}, \mathcal{D})$. *Then* $\mathcal{D}^{\mathbf{X}Y}$ *is a causal DAG if and only if there exists a valid adjustment set relative to* $(\mathbf{X}, Y)$ *in* $\mathcal{D}$.

Further, an adjustment set that is valid in the original graph is also valid in the forbidden projection and vice versa:

Proposition 8 *Let* $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$. *Then* $\mathbf{Z}$ *is a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$ *if and only if* $\mathbf{Z}$ *is also a valid adjustment set relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}^{\mathbf{XY}}$.

Using the forbidden projection, we now define the $\mathbf{O}^*$-set and prove that it is equal to the $\mathbf{O}$-set.

Definition 9 ($\mathbf{O}^*$-set) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint node sets in a DAG* $\mathcal{D}$*. We define* $\mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D})$ *as:*

$$\mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \mathrm{pa}(\mathbf{Y}, \mathcal{D}^{\mathbf{XY}}) \setminus (\mathbf{X} \cup \mathbf{Y}).$$

In words, the $\mathbf{O}^*$-set is the set of parents of $\mathbf{Y}$ in the forbidden projection $\mathcal{D}^{\mathbf{XY}}$, excluding treatment nodes and outcome nodes. The next proposition states our key result.

Proposition 10 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint subsets of the node set* $\mathbf{V}$ *of a DAG* $\mathcal{D}$ *such that* $\mathbf{Y} \subseteq \mathrm{de}(\mathbf{X}, \mathcal{D})$*. Then* $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D})$*.*

It now follows trivially that the statements about the $\mathbf{O}$-set in Proposition 3 are true for the $\mathbf{O}^*$-set as well.

Again, $\mathbf{Y} \subseteq \mathrm{de}(\mathbf{X}, \mathcal{D})$ in Proposition 10 is not a severe restriction, because if $\mathbf{Y} \not\subseteq \mathrm{de}(\mathbf{X}, \mathcal{D})$, we can instead consider the effect on $\mathbf{Y} \cap \mathrm{de}(\mathbf{X}, \mathcal{D})$, as we know that the effect on $\mathbf{Y} \setminus \mathrm{de}(\mathbf{X}, \mathcal{D})$ is zero.

Figure 3 shows the forbidden projection with respect to ALN and DET of the DAG in Figure 1. The $\mathbf{O}$-set {AIS, CDR} (in boxes) is the parent set of DET. All other valid adjustment sets are less efficient, for example the afore-mentioned sets {AFF, SAN}, {AIS, CDR, AFF} and {AFF, APA, AIS, CDR, SAN}. Due to Proposition 8, the validity of all of these sets can be confirmed by using the generalised adjustment criterion stated above on either the original DAG (Figure 1) or its forbidden projection (Figure 3). See Figure 2 for further examples.

To summarise, the forbidden projection can be used as follows: First, check in the original graph if $\mathbf{Y} \subseteq \mathrm{de}(\mathbf{X}, \mathcal{D})$. Next, construct the forbidden projection $\mathcal{G}^{\mathbf{XY}}$ and check for bi-directed edges. If there is a bi-directed edge between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$, then the causal effect of interest is not identified via adjustment (Proposition 6). Else, $\mathcal{G}^{\mathbf{XY}}$ contains all information necessary to determine a valid adjustment set (Proposition 8), and in particular the $\mathbf{O}$-set, which is then the set of parents of $\mathbf{Y}$ (Definition 9, Proposition 10). If $\mathbf{Y}$ contains only one node, then $\mathcal{G}^{\mathbf{XY}}$ is a causal DAG itself and hence straightforward to interpret (Proposition 7).

4. Optimal Adjustment in the IDA Algorithm

In Sections 2 and 3, we considered optimal adjustment in DAGs, and in Appendix C we generalised the results to amenable maxPDAGs, which include amenable CPDAGs. As a reminder, a maxPDAG is said to be amenable relative to $(\mathbf{X}, \mathbf{Y})$ if every proper possibly causal path from $\mathbf{X}$ to $\mathbf{Y}$ starts with a directed edge out of $\mathbf{X}$. In this section, we consider non-amenable CPDAGs and maxPDAGs.

CPDAGs and maxPDAGs are of interest because they are the output of popular causal search algorithms, i.e. algorithms that attempt to learn a graph from data. Under the linear model with Gaussian error terms, which we focus on in this section, it is generally not possible to learn a unique DAG. Even under the additional assumptions of causal sufficiency and faithfulness (see e.g. Spirtes et al., 2000), one can at best learn a Markov equivalence class of DAGs, uniquely represented by a CPDAG (see e.g. Andersson et al., 1997). Given additional knowledge of some causal relationships between variables, access to interventional data, or other model restrictions, one can obtain a refinement of this class,

uniquely represented by a maxPDAG (Meek, 1995; Perković et al., 2017). For a CPDAG or maxPDAG $\mathcal{G}$, we use $[\mathcal{G}]$ to denote the set of DAGs that it represents. The interpretation of edges in a CPDAG or maxPDAG $\mathcal{G}$ is as follows: A directed edge $A \to B$ means that this edge is present in all DAGs in $[\mathcal{G}]$. An undirected edge $A - B$ means that A and B are adjacent in every DAG in $[\mathcal{G}]$ and there is at least one DAG in $[\mathcal{G}]$ with $A \to B$ and at least one with $A \leftarrow B$.

We suppose in this section that we are interested in a univariate exposure X and a univariate outcome Y. For a given CPDAG or maxPDAG $\mathcal{G}$, the true causal effect of X on Y may differ across the DAGs in $[\mathcal{G}]$. In particular, Perković (2020) (Proposition 4.2) showed that assuming $Y \notin \mathrm{pa}(X, \mathcal{G})$, the true causal effect of X on Y differs across DAGs in $[\mathcal{G}]$ if and only if $\mathcal{G}$ is non-amenable relative to (X, Y), i.e. there is a possibly causal path from X to Y that starts with an undirected edge. Hence, when $\mathcal{G}$ is non-amenable relative to (X, Y), we can at best determine a multiset of possible causal effects $(\tau_{yx}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}$, one for each DAG in $[\mathcal{G}]$. (A multiset $(\tau_{yx}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}$ may contain the same entry multiple times, e.g. if $[\mathcal{G}]$ contains five DAGs, of which three imply an effect of 0 and two imply an effect of 1.2, then $(\tau_{yx}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]} = \{0, 0, 0, 1.2, 1.2\}$.) While obviously less informative than a single number, this multiset of possible causal effects may still yield useful statistics. The minimum absolute value, for example, is a lower bound for the size of the causal effect. However, enumerating all DAGs in $[\mathcal{G}]$ is computationally very expensive even for moderately sized $\mathcal{G}$ when there are many undirected edges.

Maathuis et al. (2009) proposed to reduce the complexity of this problem as follows. Consider two DAGs $\mathcal{D}, \mathcal{D}' \in [\mathcal{G}]$ such that $\mathrm{pa}(X, \mathcal{D}) = \mathrm{pa}(X, \mathcal{D}') = \mathbf{P}$ and $Y \notin \mathbf{P}$. As the parents of X form a valid adjustment set (Pearl, 2009, p. 72f.), $\tau_{yx}(\mathcal{D}) = \tau_{yx}(\mathcal{D}') = \tau_{yx}(\mathbf{P})$, where $\tau_{yx}(\mathbf{P})$ denotes the coefficient of X in the linear regression of Y on X and $\mathbf{P}$, i.e. $\beta_{yx.\mathbf{p}}$. Let $\mathbb{P} = \{\mathrm{pa}(X, \mathcal{D}) \mid \mathcal{D} \in [\mathcal{G}]\}$ denote the set of all possible parent sets of X compatible with $\mathcal{G}$. Then $(\tau_{yx}(\mathbf{P}))_{\mathbf{P}\in\mathbb{P}}$ contains the same distinct values as $(\tau_{yx}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}$, while $|\mathbb{P}| \leq |[\mathcal{G}]|$. Maathuis et al. (2009) showed that it is possible to determine $\mathbb{P}$ locally from the CPDAG $\mathcal{G}$ without enumerating all DAGs in $\mathcal{G}$. They hence proposed a simple local procedure for calculating $(\hat{\tau}_{yx}(\mathbf{P}))_{\mathbf{P}\in\mathbb{P}}$, which is called 'local IDA' (local Intervention Calculus when the DAG is Absent). Perković et al. (2017) proposed a semi-local generalisation to maxPDAGs ('semi-local IDA').

The semi-local IDA algorithm for a maxPDAG is given in Algorithm 1. Let $\mathrm{sib}(X, \mathcal{G})$ denote the set of nodes sharing an undirected edge with X in $\mathcal{G}$. Semi-local IDA loops over all subsets $\mathbf{S} \subseteq \mathrm{sib}(X, \mathcal{G})$. It first constructs a graph $\mathcal{G}'$ such that $\mathrm{pa}(X, \mathcal{G}') = \mathbf{P} = \mathrm{pa}(X, \mathcal{G}) \cup \mathbf{S}$. Here the complexity reduction becomes apparent: only the edges adjacent to X need to be oriented. To verify whether the added orientations are compatible with the original graph $\mathcal{G}$, the algorithm attempts to extend the graph to a maxPDAG by applying Meek's orientation rules (ConstructMaxPDAG algorithm; Meek, 1995; Perković et al., 2017; see Figure 8 in Appendix A). This step is semi-local as edges not adjacent to X need to be oriented. If successful, $\hat{\beta}_{yx.\mathbf{p}}$ is added as a possible causal effect estimate, where $\mathbf{P} = \mathrm{pa}(X, \mathcal{G}') = \mathbf{S} \cup \mathrm{pa}(X, \mathcal{G})$.

Nandy et al. (2017) further generalised semi-local IDA to sets $\mathbf{X}$ and $\mathbf{Y}$. However, this procedure does not use regression adjustment for possible causal effect estimation and is therefore not directly related to our results.

Algorithm 1 Local or semi-local IDA (Maathuis et al., 2009; Perković et al., 2017).
When the input is a CPDAG, line 7 can be simplified and the algorithm becomes fully local.

Require: CPDAG or maxPDAG $\mathcal{G}$ with node set $\mathbf{V} = \{V_1, \ldots, V_p, X, Y\}$, i.i.d. observations for $V_1, \cdots, V_p, X, Y$

Ensure: multiset of estimates $\widehat{\Theta}$

1: $\widehat{\Theta} \leftarrow \emptyset$
2: $\text{sib}(X, \mathcal{G}) \leftarrow \{V \in \mathbf{V} : X - V \text{ in } \mathcal{G}\}$
3: **for all** $\mathbf{S} \subseteq \text{sib}(X, \mathcal{G})$ **do**
4: **LocalBg** $\leftarrow \emptyset$
5: for all $S \in \mathbf{S}$, add $\{S \rightarrow X\}$ to **LocalBg**
6: for all $S \in \text{sib}(X, \mathcal{G}) \setminus \mathbf{S}$, add $\{S \leftarrow X\}$ to **LocalBg**
7: $\mathcal{G}' \leftarrow \text{ConstructMaxPDAG}(\mathcal{G}, \textbf{LocalBg})$
8: **if** $\mathcal{G}' \neq$ "FAIL" **then**
9: **if** $Y \notin \text{pa}(X, \mathcal{G}')$ **then**
10: regress Y on $X \cup \text{pa}(X, \mathcal{G}')$ and add the estimated coefficient of X to $\widehat{\Theta}$
11: **else**
12: add 0 to $\widehat{\Theta}$
13: **end if**
14: **end if**
15: **end for**
16: **return** $\widehat{\Theta}$

Algorithm 2 Optimal IDA.

Require: CPDAG or maxPDAG $\mathcal{G}$ with node set $\mathbf{V} = \{V_1, \ldots, V_p, X, Y\}$, i.i.d. observations for $V_1, \cdots, V_p, X, Y$

Ensure: multiset of estimates $\widehat{\Theta}$

1: $\widehat{\Theta} \leftarrow \emptyset$
2: $\text{sib}(X, \mathcal{G}) \leftarrow \{V \in \mathbf{V} : X - V \text{ in } \mathcal{G}\}$
3: **for all** $\mathbf{S} \subseteq \text{sib}(X, \mathcal{G})$ **do**
4: **LocalBg** $\leftarrow \emptyset$
5: for all $S \in \mathbf{S}$, add $\{S \rightarrow X\}$ to **LocalBg**
6: for all $S \in \text{sib}(X, \mathcal{G}) \setminus \mathbf{S}$, add $\{S \leftarrow X\}$ to **LocalBg**
7: $\mathcal{G}' \leftarrow \text{ConstructMaxPDAG}(\mathcal{G}, \textbf{LocalBg})$
8: **if** $\mathcal{G}' \neq$ "FAIL" **then**
9: **if** $Y \in \text{possde}(X, \mathcal{G}')$ **then**
10: regress Y on $X \cup \mathbf{O}(X, Y, \mathcal{G}')$ and add the estimated coefficient of X to $\widehat{\Theta}$
11: **else**
12: add 0 to $\widehat{\Theta}$
13: **end if**
14: **end if**
15: **end for**
16: **return** $\widehat{\Theta}$

4.1 Optimal IDA

HPM19 established that the parents of X, as used for adjustment by semi-local IDA, form one of the least efficient valid adjustment sets. It therefore seems a good idea to replace $\text{pa}(X, \mathcal{D})$ by the $\mathbf{O}$-set within the IDA algorithm to improve estimation precision. The key question is, however, whether the possible $\mathbf{O}$-sets can still be determined semi-locally. More formally, our aim is to estimate the multiset $(\tau_{yx}(\mathbf{O}))_{\mathbf{O}\in\mathbb{O}}$, $\mathbb{O} = \{\mathbf{O}(X, Y, \mathcal{D}) \,|\, \mathcal{D} \in [\mathcal{G}]\}$, where with a slight abuse of notation we define $\tau_{yx}(\mathbf{O}) = 0$ if $Y \notin \text{possde}(X, \mathcal{D})$. As before, for two DAGs $\mathcal{D}$ and $\mathcal{D}'$ with the same valid $\mathbf{O}$-set $\mathbf{O}(X, Y, \mathcal{D}) = \mathbf{O}(X, Y, \mathcal{D}') = \mathbf{O}$, we have $\tau_{yx}(\mathcal{D}) = \tau_{yx}(\mathcal{D}') = \tau_{yx}(\mathbf{O})$.

At first glance, it appears impossible to determine $\mathbb{O}$ locally or semi-locally, as by Definitions 1 and 9 the causal nodes, their parents and the forbidden nodes, or the forbidden projection, are required to find the $\mathbf{O}$-set. However, it turns out that $\mathbb{O}$ can be determined semi-locally almost in the same manner as $\mathbb{P}$. This is because once the directions of all edges involving X are given, i.e. for given $\mathbf{P}$, application of Meek's rules reveals all descendants of X and, in consequence, all causal nodes, their parents and the forbidden nodes (cf. Lemma 18 in Appendix C). Hence, via Meek's rules there exists a correspondence between possible parent sets and possible $\mathbf{O}$-sets. We therefore propose Algorithm 2, which we call optimal IDA. It is implemented in the `R` package `pcalg` (Kalisch et al., 2012, 2019).

Algorithm 2 does not specify whether $\mathbf{O}(X, Y, \mathcal{G}')$ is determined from $\mathcal{G}'$ or from the forbidden projection. We expect this choice to be of limited relevance to the algorithm's runtime. In our implementation, we determine $\mathbf{O}(X, Y, \mathcal{G}')$ directly from $\mathcal{G}'$. Note also that different possible parent sets can correspond to the same $\mathbf{O}$-set. Hence, optimal IDA could be modified to collect all sets in $\mathbb{O}$ first, remove duplicates, and only then estimate regression coefficients.

In the following, we first state formally what can be said about the efficiency of the estimates output by optimal IDA, showing that it is worthwhile to replace the parents of X by the $\mathbf{O}$-set. Subsequently we compare the computational burden of the two algorithms.

Proposition 11 *Let X and Y be nodes in a causal CPDAG or maxPDAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, such that $\mathbf{V}$ follows a causal linear model compatible with $\mathcal{G}$ with Gaussian errors. Let $\widehat{\Theta}^{\mathbf{P}}$ and $\widehat{\Theta}^{\mathbf{O}}$ be the multisets returned by semi-local IDA and optimal IDA, respectively, applied to X, Y and $\mathcal{G}$, with the subsets of $\text{sib}(X, \mathcal{G})$ considered in the same order for both. Then, for $i \in \{1 \dots, k\}$, with $k = |\widehat{\Theta}^{\mathbf{P}}| = |\widehat{\Theta}^{\mathbf{O}}|$,*

1. $\mathbb{E}[\widehat{\Theta}^{\mathbf{P}}_i] = \mathbb{E}[\widehat{\Theta}^{\mathbf{O}}_i]$ *and*
2. $a.var(\widehat{\Theta}^{\mathbf{P}}_i) \geq a.var(\widehat{\Theta}^{\mathbf{O}}_i)$.

The proof is given in Appendix D. Note that if we do not assume Gaussianity in Proposition 11, then $a.var(\hat{\beta}_{yx.\mathbf{o}}) \leq a.var(\hat{\beta}_{yx.\mathbf{z}})$ can only be guaranteed if (i) $\mathbf{Z}$ is a valid adjustment set in the true DAG, and (ii) $\mathbf{O}$ is the $\mathbf{O}$-set of the true DAG. This is because in a causal linear model with non-Gaussian errors, a variable is only required to be linear in its parents, and is not necessarily linear given another node set (cf. Nandy et al., 2017). However, if we are willing to assume that all errors in the underlying causal model are non-Gaussian, alternative causal search approaches exist which output a DAG instead of an equivalence class, e.g. algorithms such as LiNGAM (Shimizu et al., 2006).

Remark 12 *(1) In terms of the computational burden, semi-local and optimal IDA are very similar for maxPDAGs. The key difference is that optimal IDA adjusts for the* $\mathbf{O}$*-set instead of the parent set of* X *(line 10), where the* $\mathbf{O}$*-set is straightforward to determine from* $\mathcal{G}'$*. However, optimal IDA crucially relies on the construction of the maxPDAG in line 7 to determine the* $\mathbf{O}$*-set, while in semi-local IDA this step can be replaced by a simple local query when the input is known to be a CPDAG. Hence, for the special case of a CPDAG, semi-local IDA can be made fully local by simplifying line 7, whereas optimal IDA cannot.*

(2) A further minor difference between semi-local and optimal IDA is the `if`*-statement in line 9. Semi-local IDA only checks whether* $Y \notin \text{pa}(X, \mathcal{G}')$*, whereas optimal IDA checks the stronger condition* $Y \in \text{possde}(X, \mathcal{G}')$*. Both conditions ensure that the considered adjustment sets* $\text{pa}(X, \mathcal{G}')$ *and* $\mathbf{O}(X, Y, \mathcal{G}')$*, respectively, are valid adjustment sets. Moreover, if* $Y \notin \text{possde}(X, \mathcal{G}')$*, then* $\tau_{yx}(\mathcal{D}) = 0$ *for any* $\mathcal{D} \in [\mathcal{G}']$*. The* 0 *estimate of optimal IDA in this case is therefore the most efficient estimate. Alternatively, we could also insist on* $Y \in \text{possde}(X, \mathcal{G}')$ *in semi-local IDA and return* 0 *otherwise. As discussed in the appendix of Maathuis et al. (2009), this is only recommended if the input graph is thought to be reliable, but can lead to the amplification of errors if the input graph is not accurate.*

Remark 13 *Proposition 11 concerns the asymptotic variance when the true CPDAG or a true maxPDAG is given. When the graph is estimated on the same data as used for IDA, the naive standard errors from the adjusted linear regressions are invalid. Although considerable progress has been made in the area of post-selection inference (e.g. Berk et al., 2013; Belloni et al., 2014; Rinaldo et al., 2019), no method has been proposed specifically for estimating standard errors of causal effect estimates after causal search.*

It is straightforward to extend optimal IDA to situations where $\mathbf{X}$ and $\mathbf{Y}$ are sets. However, as noted earlier, in this case joint causal effect estimation via regression adjustment is not always possible. Optimal IDA will then not return an estimate. The estimation procedures used by joint IDA (Nandy et al., 2017) provide an alternative.

4.2 Illustration

We now illustrate optimal IDA (Algorithm 2) using a toy example. Consider the CPDAG $\mathcal{G}$ shown in Figure 4(a) and suppose we are interested in the causal effect of X on Y. Clearly, $\mathcal{G}$ is not amenable relative to (X, Y) and thus it is sensible to apply optimal IDA.

The set $sib(X, \mathcal{G})$ contains 3 nodes, hence there are 8 potential orientations of the undirected edges with endpoint X. From these 8, 3 imply new v-structures and are thus not compatible with $\mathcal{G}$. The other 5 can be extended to the maxPDAGs shown in Figure 4(b-f), where the bold arrows indicate orientations derived by Meek's rules (see Figure 8 in Appendix A). For example, in 4(b) it follows from $V_1 \to X \to V_3$ that $V_1 \to V_3$ by Meek's Rule 2. By Rule 1, it then follows that $V_3 \to V_5$. The compatibility check and the application of Meek's rules are carried out in line 7 of optimal IDA.

Next, optimal IDA checks for each maxPDAG $\mathcal{G}'$, whether $Y \in \text{possde}(X, \mathcal{G}')$. Here, this is the case for all maxPDAGs except 4(c). For the other four graphs, $\mathbf{O} = \mathbf{O}(X, Y, \mathcal{G}')$ is determined and used to compute $\hat{\beta}_{yx.\mathbf{o}}$. We indicate $\mathbf{O}(X, Y, \mathcal{G}')$ by boxes in the Figures 4(b) and 4(d)-(f). For (c), an effect estimate of zero is returned.

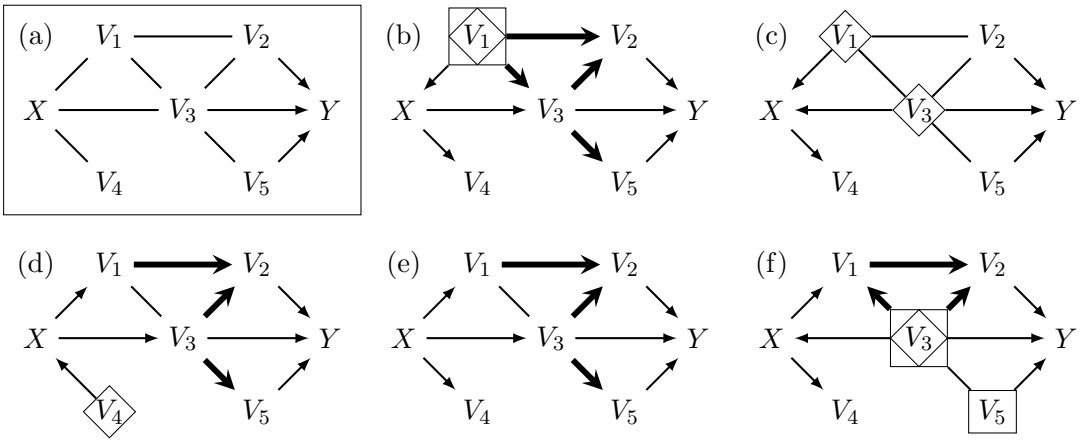


Figure 4: A CPDAG $\mathcal{G}$ (a) and the five maxPDAGs (b-f) corresponding to the five valid orientations of the neighbourhood of X. The bold edges have been obtained by applying Meek's rules. For each maxPDAG $\mathcal{G}'$, the boxes $\square$ indicate $\mathbf{O}(X, Y, \mathcal{G}')$, while the diamonds $\lozenge$ indicate $\text{pa}(X, \mathcal{G})$. In (c), optimal IDA returns 0, as there is no possibly causal path from X to Y.

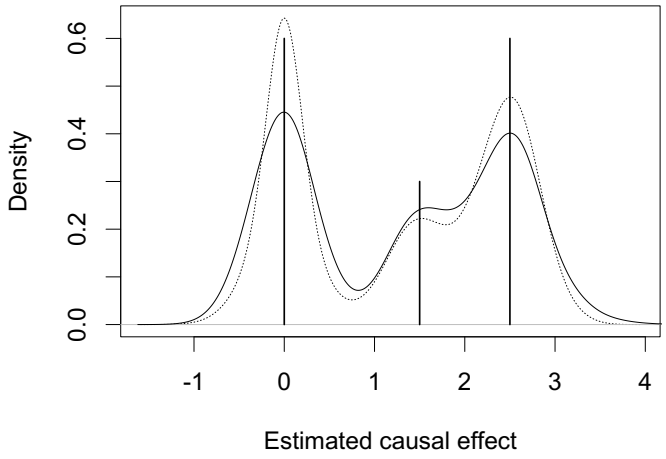


Figure 5: IDA density plot in the style of Maathuis et al. (2009). Shown are density curves for the estimated possible causal effects returned by local IDA (solid) and optimal IDA (dotted). The true possible causal effects are 0, 1.5 and 2.5 (vertical lines; height indicates relative frequency: 0 and 2.5 each occur in two of the five maxPDAGs in Figure 4).

For comparison, the diamonds in Figures 4(b-f) show the adjustment sets in local IDA (Algorithm 1), i.e. $\text{pa}(X, \mathcal{G}')$. In (b) and (e), $\text{pa}(X, \mathcal{G}') = \mathbf{O}(X, Y, \mathcal{G}')$. In (c), optimal IDA returns zero (Algorithm 2, line 12), while local IDA returns $\hat{\beta}_{yx.\mathbf{p}}$ with $\mathbf{P} = \{V_1, V_3\}$, which converges to $\beta_{yx.\mathbf{p}} = 0$. The main advantage of optimal IDA becomes apparent in cases (d) and (f): In (d), $\mathbf{O}(X, Y, \mathcal{G}') = \emptyset$, whereas $\text{pa}(X, \mathcal{G}') = \{V_4\}$ which is guaranteed to reduce efficiency. In (f), $\text{pa}(X, \mathcal{G}') = \{V_3\}$ and $\mathbf{O}(X, Y, \mathcal{G}') = \{V_3, V_5\}$, where the latter improves efficiency.

For further illustration, we carried out a small simulation study in which we generated data according to a causal linear model compatible with Figure 4(b). 1 000 datasets with 40 observations each were generated and given as input to local IDA and optimal IDA, together with the CPDAG in Figure 4(a). The true possible causal effects are 0, 1.5 and 2.5, visualised as vertical lines in Figure 5. The plot shows smoothed density curves for the estimates returned by local IDA (solid) and optimal IDA (dotted). The density plot for optimal IDA is clearly narrower around the values 0 and 2.5. The difference between the algorithms is even more pronounced for graphs with more nodes and longer paths (not shown). The `R`-code (R Core Team, 2019) for reproducing Figure 5 is available in the Online Supplement.

4.3 Simulation

In order to compare the performance of optimal versus local IDA in finite sample settings, we carried out a more extensive simulation study. The design was chosen to reflect a typical situation where IDA is used, i.e. interest lies in the causal effect of X on Y in a (known or estimated) CPDAG $\mathcal{G}$ that is non-amenable relative to (X, Y). Non-amenability implies that the multiset $(\tau_{xy}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}$ of possible causal effects of X on Y compatible with $\mathcal{G}$ contains more than one distinct value (for almost all parameters values of the causal linear model) (Perković, 2020, Proposition 4.2). A useful summary of $(\tau_{xy}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}$ is the minimum absolute value, $\min(\text{abs}((\tau_{xy}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}))$, because when this value is non-zero, we know that X has *some* effect on Y. The aim of our simulation study was to compare how well $\min(\text{abs}((\tau_{xy}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}))$ is estimated by optimal versus local IDA, in terms of the Monte-Carlo mean squared error (MSE).

We investigated 24 scenarios by considering all combinations of the following parameters: number of nodes $p \in \{10, 20, 50, 100\}$, expected number of neighbours per node $d \in \{2, 3, 4\}$, and sample size $n \in \{100, 1\,000\}$. In each scenario, the following was repeated 1 000 times (`R` code for reproducing the simulation study is available in the Online Supplement):

A DAG $\mathcal{D}$, with CPDAG $\mathcal{G}$, with p nodes and d expected neighbours per node was randomly chosen such that $\mathcal{G}$ was non-amenable relative to two randomly chosen nodes (X, Y) and such that $\min(\text{abs}((\tau_{xy}(\mathcal{D}))_{\mathcal{D}\in[\mathcal{G}]}))$ was non-zero. (Note that the DAG with its unique 'true' causal effect was simulated for convenience only. Conceptually, we drew directly from the space of CPDAGs, which is why we consider the whole multiset of possible effects to be 'the truth'.) The following was then repeated 100 times: A dataset with n observations was generated from a linear causal model on $\mathcal{D}$ where the non-zero coefficients were randomly chosen from a uniform distribution on $[-1, -0.1] \cup [0.1, 1]$. Greedy equivalence search (Chickering, 2002) was applied to the data, yielding an estimated CPDAG $\mathcal{G}^*$. Optimal and local IDA were both applied to the true CPDAG $\mathcal{G}$ and the estimated

CPDAG $\mathcal{G}^*$. The four output multisets of estimates were summarised by their minimum absolute values. These were compared on the basis of their Monte-Carlo MSE, i.e. the squared difference between the estimated minimum absolute value and the true minimum absolute value, averaged over the 100 repetitions. Specifically, we calculated the MSEs for the estimated minima using optimal IDA versus local IDA by computing the relative MSE (RMSE), MSE(optimal IDA)/MSE(local IDA). This was done separately for $\mathcal{G}$ and for $\mathcal{G}^*$, and denoted r and r^*, respectively. An RMSE of less than one indicates that optimal IDA is more precise than local IDA in estimating the minimum of the multiset of causal effects. In light of Remark 13, we did not consider estimated standard errors.

In addition to the above 24 scenarios, we investigated the relative performance of optimal IDA in a scenario where the graph is sparse ($d = 1$) and the sample size is moderate ($n = 100$). We considered eight setting where the number of nodes was between $p = 10$ and $p = 1\,000$. Such high-dimensional scenarios occur, for instance, with gene expression data. As greedy equivalence search is slow for large graphs, we reduced the number of replications from 1 000 to 100 and the number of datasets per graph from 100 to 10.

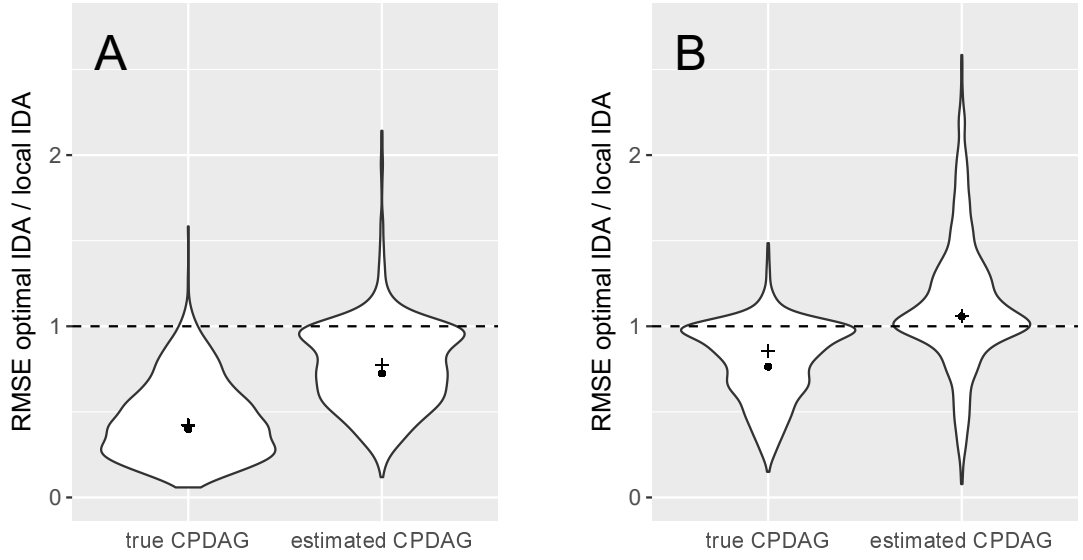


Figure 6: Violin plots of the relative mean squared errors (RMSEs) r and r^* for the true and estimated CPDAGs, respectively. Scenario **A**: $p = 100$ nodes, $d = 4$ expected neighbours per node, sample size $n = 1\,000$. Scenario **B**: $p = 10$, $d = 4$ and $n = 100$. The dots mark the geometric means, the plus signs the medians.

Figure 6 shows violin plots of the RMSEs r and r^* over the 1 000 repetitions, together with the geometric mean and the median. Two scenarios are shown: The one where optimal IDA showed the best overall performance (scenario **A**, $p = 100, d = 4, n = 1\,000$), and the worst one (scenario **B**, $p = 10, d = 4, n = 100$) of all the simulation settings considered. The geometric means and medians for all scenarios are summarised in Tables 1 and 2; the

	$n = 100$			$n = 1\,000$		
	$d = 2$	$d = 3$	$d = 4$	$d = 2$	$d = 3$	$d = 4$
$p = 10$	0.70 (0.76)	0.72 (0.79)	0.76 (0.86)	0.69 (0.78)	0.71 (0.78)	0.75 (0.86)
$p = 20$	0.64 (0.69)	0.63 (0.68)	0.61 (0.66)	0.63 (0.68)	0.60 (0.65)	0.59 (0.65)
$p = 50$	0.60 (0.64)	0.54 (0.57)	0.51 (0.55)	0.55 (0.58)	0.50 (0.54)	0.46 (0.49)
$p = 100$	0.57 (0.61)	0.50 (0.52)	0.44 (0.46)	0.54 (0.58)	0.44 (0.46)	0.40 (0.42)

Table 1: Geometric means (in parentheses: medians) of the relative mean squared errors (RMSEs) r over 1 000 repetitions for scenarios with different numbers of nodes (p), expected number of neighbours per node (d), and sample sizes (n). Optimal and local IDA were applied to the true CPDAG $\mathcal{G}$.

	$n = 100$			$n = 1\,000$		
	$d = 2$	$d = 3$	$d = 4$	$d = 2$	$d = 3$	$d = 4$
$p = 10$	1.06 (1.01)	1.06 (1.04)	1.06 (1.06)	0.95 (0.99)	0.97 (1.00)	1.01 (1.00)
$p = 20$	0.99 (1.00)	0.99 (1.00)	0.96 (1.00)	0.88 (0.96)	0.89 (0.97)	0.94 (0.99)
$p = 50$	0.94 (0.98)	0.89 (0.93)	0.89 (0.93)	0.81 (0.90)	0.79 (0.85)	0.78 (0.86)
$p = 100$	0.97 (1.00)	0.94 (0.97)	0.90 (0.94)	0.81 (0.91)	0.73 (0.80)	0.72 (0.77)

Table 2: Geometric means (in parentheses: medians) of the relative mean squared errors (RMSEs) r^* over 1 000 repetitions for scenarios with different numbers of nodes (p), expected number of neighbours per node (d), and sample sizes (n). Optimal and local IDA were applied to the estimated CPDAG $\mathcal{G}^*$.

complete set of violin plots is shown in Appendix E. Optimal IDA clearly outperformed local IDA, in terms of the geometric mean and median of the RMSE, in all scenarios when applied to the true CPDAG. When the CPDAG was estimated using greedy equivalence search, optimal IDA was still superior in the majority of scenarios, but r^* was notably larger than r in all scenarios, i.e. the relative performance of optimal IDA was worse with an estimated CPDAG than with a known CPDAG. As an estimated graph inevitably contains some errors regarding the presence and direction of edges, this result may indicate that estimation adjusting for the **O**-set suffers more from such errors than adjusting for the set of parents of X.

Small n and small p do not entail much advantage of using optimal IDA: In graphs with only a few nodes, the **O**-set and the set of parents of X are often similar or even coincide, so that the gain in efficiency when using the **O**-set is less pronounced. A smaller sample size leads to more errors in the estimated graph, which affects estimation of the **O**-set more than estimation of the set of parents of X, as we conjectured above. However, optimal IDA seems to have a slight advantage for larger d when p is also larger.

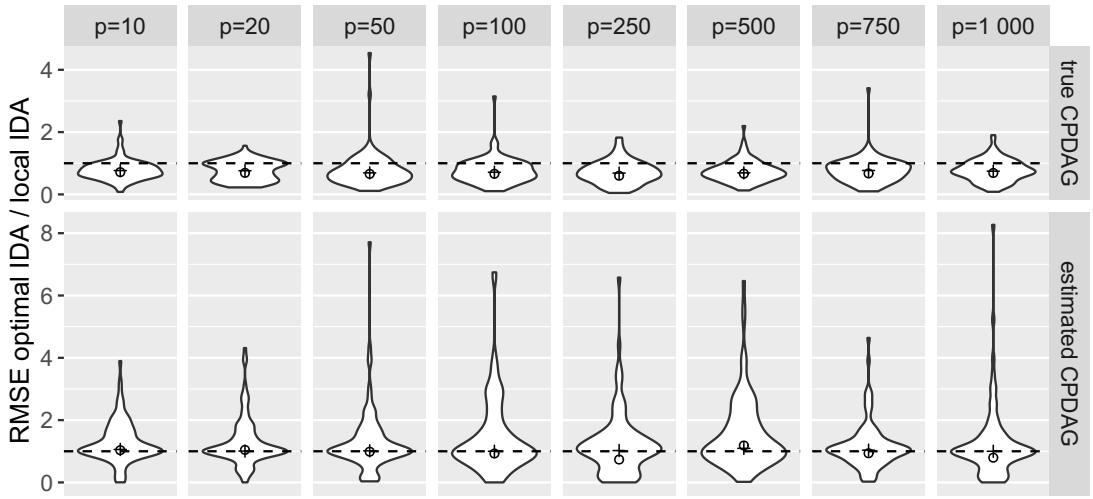


Figure 7: Violin plots of the relative mean squared errors (RMSEs) r and r^* for the true and estimated CPDAGs, respectively. All graphs have $d = 1$ expected neighbour per node and graphs were estimated from $n = 100$ observations; the number of nodes p varies. The dots mark the geometric means, the plus signs the medians.

	$p = 10$	$p = 20$	$p = 50$	$p = 100$
true CPDAG	0.71 (0.77)	0.69 (0.76)	0.66 (0.69)	0.66 (0.71)
estimated CPDAG	1.04 (1.06)	1.04 (1.01)	0.99 (1.02)	0.93 (1.00)
	$p = 250$	$p = 500$	$p = 750$	$n = 1\,000$
true CPDAG	0.60 (0.68)	0.67 (0.69)	0.67 (0.77)	0.69 (0.75)
estimated CPDAG	0.73 (1.02)	1.16 (1.09)	0.94 (1.04)	0.79 (1.00)

Table 3: Geometric means (in parentheses: medians) of the relative mean squared errors (RMSEs) r and r^* for the true and estimated CPDAGs, respectively, over 100 repetitions for scenarios with different numbers of nodes (p), $d = 1$ expected neighbours per node, and sample size $n = 100$.

The additional results for the sparse graphs are shown in Figure 7 and Table 3. Optimal IDA outperformed local IDA regardless of the number of nodes p when the CPDAG was known. When the CPDAG was not known, the median RMSE was about 1 for all p. The geometric mean varied around 1 with no obvious pattern, which may be due to the small number of replications. The results suggest that in the high-dimensional setting, the primary difficulty is learning the graph, limiting the advantage of optimal IDA over local IDA.

In summary, based on the simulation results, we recommend using optimal IDA when there is high confidence in the estimated graph. The advantage over local IDA will be most pronounced when the number of nodes is at least 20, or better 50 or more.

5. The O-Set and Non-Graphical Variable Selection

We now assume that neither the causal DAG $\mathcal{D}$ nor a CPDAG or maxPDAG is known to us, therefore we wish to select a valid adjustment set in a non-graphical manner. We restrict our discussion to the case where we have a univariate treatment X and outcome Y of interest.

In multiple regression analyses, it is common to apply variable selection procedures, e.g. backward selection, to find a set of relevant predictors for an outcome Y. In high-dimensional settings, regularisation methods that combine selection and estimation, such as the Lasso or the Elastic Net, are commonly used (Tibshirani, 1996; Zou and Hastie, 2005). While variable selection for prediction is in general a different task than finding an efficient or optimal adjustment set for causal effect estimation, we will discuss next under what assumptions and modifications these tasks coincide. For a general overview of the relation between variable and confounder selection see Witte and Didelez (2019). A basic assumption for the validity of a selected adjustment set is that the set $\mathbf{Z}$ from which we select the variables must itself be a valid adjustment set, as defined in Section 2. A number of selection procedures can then be used to determine different types of valid adjustment sets as subsets of $\mathbf{Z}$, e.g. a minimal valid adjustment set (Witte and Didelez, 2019; de Luna et al., 2011).

Consider first Algorithm 3, which shows the template for backward regression selection (see e.g. Kleinbaum and Kupper, 1978; Montgomery et al., 2012) with the above basic assumption added at the outset. Under the linear model assumptions with Gaussian errors, $Y \perp\!\!\!\perp Z_i \mid (\mathbf{Z}'_{-i}, X)$ can be tested by comparing the models with regressors $\mathbf{Z}'_{-i} \cup \{X\}$ versus $\mathbf{Z}' \cup \{X\}$, using a t-test with null hypothesis $\beta_{yz_i.x\mathbf{z}'_{-i}} = 0$. 'Pval' in line 6 is a function that outputs the p-value of a test for the null hypothesis specified in the argument. The maximum p-value is compared in line 9 to a threshold α. For a given α, Algorithm 3 implements the classical 'p-value method' (see e.g. Greenland and Pearce, 2015). Denote by $F_{\chi^2_1}(.)$ the distribution function of the χ^2 distribution with one degree of freedom. For a given sample size n, Algorithm 3 with $\alpha = 1 - F_{\chi^2_1}(2)$ or $\alpha = 1 - F_{\chi^2_1}(\log(n))$ is equivalent to backward selection using the AIC or BIC, respectively (e.g. Murtaugh, 2014; Derryberry et al., 2018; although the motivation for using them stems from frameworks other than independence testing, see Akaike, 1974 and Schwarz, 1978). Algorithm 3 can easily be adapted to work with a measure of conditional independence other than the p-value of the

t-test. For example, two non-parametric implementations of Algorithm 3 were proposed by Li et al. (2005).

If the true independence relations are known, Algorithm 3 can be condensed to its oracle version, Algorithm 4. Comparing p-values is then redundant, and every Z_i needs to be visited only once, as it follows from general properties of conditional independence that the ordering $Z_1, Z_2, \ldots, Z_p$ does not matter, provided the joint probability of all variables is strictly positive. Essentially, Algorithm 4 eliminates variables until only the 'direct predictors' of Y are left, i.e. those variables with non-zero coefficients in the oracle regression of Y on X and $\mathbf{Z}$.

Algorithm 4 is the non-graphical version of the pruning algorithm introduced in HPM19 which uses d-separation relationships to prune a valid adjustment set to a subset such that the resultant effect estimator has a smaller asymptotic variance. Assume now that an underlying graph exists. The following Proposition 14 formalises how the **O**-set can be viewed as the target set of backward variable selection algorithms and follows from Proposition 3.6 of HPM19 and Theorem 1 in RS20.

Algorithm 3 Backward regression selection.

Require: i.i.d. observations for variables X, Y and $\mathbf{Z}$, such that $\mathbf{Z}$ is a valid adjustment set relative to (X, Y)

$\mathbf{Z}' \leftarrow \mathbf{Z}$
Pmax $\leftarrow 1$
while Pmax $> \alpha$ **do**
  Plist $\leftarrow$ empty list of length $|\mathbf{Z}'|$
  for all i in 1 to $|\mathbf{Z}'|$ **do**
    Plist$[i] \leftarrow$ Pval$(Y \perp\!\!\!\perp Z_i \mid (X, \mathbf{Z}'_{-i}))$
  end for
  Pmax $\leftarrow$ max(Plist)
  if Pmax $> \alpha$ **then**
    $\mathbf{Z}' \leftarrow \mathbf{Z}'_{-\text{argmax(Plist)}}$
  end if
end while
return $\mathbf{Z}'$

Algorithm 4 Oracle backward regression selection.

Require: independence relations between variables X, Y and $\mathbf{Z}$, such that $\mathbf{Z}$ is a valid adjustment set relative to (X, Y)

$\mathbf{Z}' \leftarrow \mathbf{Z}$
for all i in 1 to $|\mathbf{Z}'|$ **do**
  if $Y \perp\!\!\!\perp Z_i \mid (X, \mathbf{Z}'_{-i})$ **then**
    $\mathbf{Z}' \leftarrow \mathbf{Z}'_{-i}$
  end if
end for
return $\mathbf{Z}'$

Proposition 14 *Let X and Y be nodes in a causal DAG, CPDAG or maxPDAG $\mathcal{G}$ with node set $\mathbf{V}$ and let $\mathbf{V}$ follow a causal model with a joint density faithful to $\mathcal{G}$. Let $\mathbf{Z}$ be a valid adjustment set relative to (X, Y) in $\mathcal{G}$ and let $\mathbf{Z}'$ be the output of Algorithm 4 when applied to $\mathbf{Z}$.*

(a) $\mathbf{Z}'$ is a valid adjustment set and does not depend on the order in which the variables in $\mathbf{Z}$ are considered in Algorithm 4.

(b) If $\mathbf{O}(X, Y, \mathcal{G}) \subseteq \mathbf{Z}$, then $\mathbf{Z}' = \mathbf{O}(X, Y, \mathcal{G})$.

(c) If $\mathbf{V}$ follows a causal linear model, then $a.var(\hat{\beta}_{yx.\mathbf{Z}'}) \leq a.var(\hat{\beta}_{yx.\mathbf{Z}})$.

(d) For every pair of values $x, x' \in \mathcal{X}$, $a.var(\hat{\Delta}_{yxx'.\mathbf{Z}'}) \leq a.var(\hat{\Delta}_{yxx'.\mathbf{Z}})$.

As mentioned earlier, Lasso estimation can also be regarded as variable selection, even though its original motivation and common usage mostly concerns prediction. Under specific assumptions, the Lasso asymptotically selects all and only all the 'direct predictors' of Y with probability 1 (Zhao and Yu, 2006; Lounici, 2008). Thus, although Lasso uses a different principle than backward selection, it follows from Proposition 14 that when the starting set $\mathbf{Z}$ is a valid adjustment set and a superset of the $\mathbf{O}$-set, the $\mathbf{O}$-set can also be viewed as the target set of the Lasso.

We emphasise again that the $\mathbf{O}$-set cannot be determined in a purely data-driven way. Neither the assumption that $\mathbf{Z}$ is valid nor $\mathbf{O}(X, Y, \mathcal{G}) \subseteq \mathbf{Z}$ can be verified empirically. Hence, prior causal knowledge is essential before any variable selection algorithm can be applied (Witte and Didelez, 2019). In contrast to (semi-)local or optimal IDA, however, selection of an adjustment set based on Algorithm 4 allows some latent structures, as long as the assumption that $\mathbf{Z}$ is a valid adjustment set continues to hold. This may be of advantage when only a subset of the variables have been measured.

The guarantees of Proposition 14 for Algorithm 4 do not translate to the finite sample version Algorithm 3. Regression selection in finite samples is known to have several weaknesses (see e.g. Harrell, 2010). Some issues are that the output may only be a local optimum, and that valid post-selection inference is difficult (Leeb and Pötscher, 2008). There is, however, a growing literature on post-selection inference both in the context of OLS-based approaches (e.g. Berk et al., 2013; Rinaldo et al., 2019) and of Lasso-based approaches (e.g. Lockhart et al., 2014; Lee et al., 2016). For causal effect estimation in a non-graphical context, post-selection inference has been considered by Belloni et al. (2014), Dukes and Vansteelandt (2020a), Dukes and Vansteelandt (2020b), Chernozhukov et al. (2018) and others.

6. Conclusions

In this paper, we provided insight into the construction and properties of the $\mathbf{O}$-set introduced by HPM19. We showed that the $\mathbf{O}$-set equals the set of parents of $\mathbf{Y}$ in the latent projection over the forbidden nodes (Proposition 10). This lends formal support to the intuition that adjusting for all direct causes of $\mathbf{Y}$ minimises the residual variance and hence improves precision when estimating the causal effect of $\mathbf{X}$ on $\mathbf{Y}$.

The forbidden projection is a useful tool in its own right when the aim is to estimate a causal effect via adjustment. It displays all variables of interest, while the forbidden variables, which must not be adjusted for, are marginalised out. The forbidden projection thus reduces the complexity of the causal graph while preserving all information relevant for choosing an adjustment set (see Propositions 8 and 7).

We further proposed a new modification of the IDA algorithm, called optimal IDA, which outputs multisets of estimates of possible causal effects by adjusting for the possible **O**-sets. We showed that this increases estimation precision also in cases where the causal structure is a-priori unknown and needs to be estimated. Moreover, this extends the applicability of optimal adjustment to non-amenable CPDAGs/maxPDAGs. Optimal IDA has been implemented in the `R` package `pcalg`. While causal search methods in general have some well-known shortcomings, IDA has proved to be a valuable tool for instance for screening purposes in large datasets (Le et al., 2013; Engelmann et al., 2015; Luo et al., 2018). The 'optimal' version can further improve its performance.

Finally, we detailed the prerequisites and assumptions under which non-graphical algorithms for backward variable selection can be viewed as aiming at selecting the **O**-set. Essentially, we need to assume that the set of variables to select from consists of all nodes in the forbidden projection, or a suitable subset thereof. The algorithm then determines the parents / direct causes of Y based on detected conditional independencies. If the input contains forbidden nodes, however, or lacks certain confounders, the algorithm might select an invalid adjustment set. To avoid the latter, sufficient prior knowledge on the set of variables corresponding to forbidden nodes is therefore a key prerequisite when automated variable selection is to be used for causal inference. While this prerequisite may not require full knowledge of the underlying causal DAG, it is important to recognise that such prior knowledge cannot be established in a purely data-driven way (Witte and Didelez, 2019).

Much research on variable selection in causal graphs has focussed on finding small or minimal adjustment sets (de Luna et al., 2011; Textor and Liśkiewicz, 2011; Knüppel and Stang, 2010). Small adjustment sets are useful during study planning, for instance when data collection is expensive and costs are to be minimised. Moreover, they entail desirable statistical properties e.g. for matching estimators, because suitable matches are more easily found when matching on a few variables only. In general, the **O**-set is not minimal, but instead entails optimality of causal effect estimation by regression adjustment in linear causal models and non-parametric settings. Simulation results further indicate that the optimality of the **O**-set extends to other parametric settings and estimation methods, e.g. estimation of the marginal odds ratio via standardised logistic regression (Witte and Didelez, 2019). Combining the benefits of small and optimal adjustment sets, RS20 show that the optimal minimal set, i.e. the set among all minimal adjustment sets yielding the most precise estimation in the class of regular asymptotically linear estimators, must be a subset of the **O**-set, underlining its relevance and importance.

We note that adjustment is only one of several possible ways of identifying causal effects. While adjusting for the **O**-set is asymptotically more efficient than adjusting for any other valid adjustment set, it is possible that an even smaller asymptotic variance can be obtained by using an alternative identification strategy, e.g. the front-door strategy (Pearl, 2009). This is further investigated in RS20 and in Guo and Perković (2020).

Finally, we would like to discuss some avenues for future research. First, given the results by RS20, a natural question is whether a non-parametric version of optimal IDA is feasible. Those aspects of IDA that relate to finding different possible valid adjustment sets are obviously not limited to the causal linear model, and estimators such as in RS20 could also be employed for any given X, Y and adjustment set. The simplifications for graph search algorithms and IDA under linearity, however, are considerable. For instance, greedy equivalence search with a Gaussian score has been shown to be consistent (Chickering, 2002), and has the advantage of always returning a CPDAG. Non-parametric graph search algorithms exist, but often come with large computational burdens and/or a low power to detect edges (Shah and Peters, 2020; Ramsey, 2014). Further, under a causal linear model, the causal effect of X on Y is a single value, see Section 2; and the marginal and conditional causal effects for different valid adjustment sets are all identical. For the non-parametric case, instead, the causal effect of X on Y is an unspecified function, and issues of non-collapsibility might also come into play. Solving these conceptual hurdles for non-parametric optimal IDA remains an open question.

Second, we assumed throughout this paper that all variables are observed. HPM19 have shown that in the presence of hidden variables, an asymptotically optimal set may not exist. Smucler et al. (2020), however, gave a sufficient condition under which an optimal adjustment set exists when the underlying DAG includes hidden variables, and showed that an optimal minimal adjustment set always exists. A necessary and sufficient condition for the existence of an optimal adjustment set, however, has not yet been formulated.

Acknowledgments

We thank the reviewers and the editor for their valuable comments and suggestions. We gratefully acknowledge financial support by the German Research Foundation (DFG—Project DI 2372/1-1).

Appendix A. Terminology

The following terminology is used throughout this paper. It is consistent with, and extends HPM19 where needed.

Graphs. A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ consists of a node set $\mathbf{V}$ and a set of edges $\mathbf{E}$. We consider three types of edges: *directed* ($\to$), *bi-directed* ($\leftrightarrow$) and *undirected* ($-$). There can be more than one edge between a given pair of nodes. We only consider loop-free graphs, i.e. an edge between a node and itself is not allowed. A loop-free graph where there is at most one edge between a given pair of nodes is called a *simple graph*. Two nodes joined by at least one edge are called *endpoints of the edge* and *adjacent*. A directed edge $A \to B$ is said to be *out of* A and *into* B. A graph $\mathcal{G}' = (\mathbf{V}', \mathbf{E}')$ is the *induced subgraph* of $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with respect to $\mathbf{V}'$ if $\mathbf{V}' \subseteq \mathbf{V}$ and $\mathbf{E}'$ includes all edges in $\mathbf{E}$ that are between nodes in $\mathbf{V}'$.

Paths. A *path* is a sequence of nodes and edges $(V_0, e_1, V_1, \ldots, e_K, V_K)$, $K \geq 1$, such that every node occurs only once and for $k = 1, \ldots, K$, e_k has endpoints V_{k-1} and V_k. In a simple graph, the path $(V_0, e_1, V_1, \ldots, e_K, V_K)$ can unambiguously be identified by the sequence of nodes $(V_0, V_1, \ldots, V_K)$ alone. V_0 and V_K are called *endpoints of the path* $(V_0, e_1, V_1, \ldots, e_K, V_K)$ and the path is said to be *between* V_0 *and* V_K or *from* V_0 *to* V_K, irrespective of the direction of the edges. For sets of nodes $\mathbf{A}$ and $\mathbf{B}$, a path is said to be between $\mathbf{A}$ and $\mathbf{B}$ or from $\mathbf{A}$ to $\mathbf{B}$ if its first node is in $\mathbf{A}$ and the last node is in $\mathbf{B}$. A path from $\mathbf{A}$ to $\mathbf{B}$ is *proper* if only its first node V_0 is in $\mathbf{A}$. Let $p = (V_0, e_1, V_1, \ldots, e_K, V_K)$ and $k = 1, \ldots, K$. Then an edge $V_k \leftarrow V_{k+1}$ on p is said to point towards $V_0, \ldots, V_k$, while an edge $V_k \to V_{k+1}$ on p is said to point towards $V_{k+1}, \ldots, V_K$. A path is *directed* from V_0 to V_K if all edges in the sequence are directed and point towards V_K. A path p is *possibly directed* from V_0 to V_K if all edges on p are either directed or undirected and there are no i, j, $1 \leq i < j \leq K$, such that $V_i \leftarrow V_j$ (cf. Perković et al. (2017); this definition of a possibly directed path is non-standard as V_i and V_j are not necessarily adjacent nodes on the path, which is required for maxPDAGs later). We define the concatenation of two paths $p = (V_0, e_1, V_1, \ldots, e_K, V_K)$ and $q = (V_K, e_{K+1}, V_{K+1}, \ldots, e_{K+L}, V_{K+L})$ as $p \oplus q = (V_0, e_1, V_1, \ldots, e_{K+L}, V_{K+L})$, where we require that the nodes $V_0, \ldots, V_{K+L}$ are distinct.

Ancestry. If there is a directed path from A to B, or if $A = B$, then A is an *ancestor* of B and B is a *descendant* of A. If there is a possibly directed path from A to B, or if $A = B$, then A is a *possible ancestor* of B and B is a *possible descendant* of A. If there is an edge $A \to B$, then A is a *parent* of B and B is a *child* of A. If there is an edge $A - B$, A and B are *siblings*. Note that in our terminology, a node is a (possible) ancestor and (possible) descendant of itself, but not a parent/child/sibling of itself. For a node V in a simple graph $\mathcal{G}$, we denote the set of all ancestors, possible ancestors, descendants, possible descendants, parents, children and siblings of V in $\mathcal{G}$ as $\mathrm{an}(V, \mathcal{G})$, $\mathrm{possan}(V, \mathcal{G})$, $\mathrm{de}(V, \mathcal{G})$, $\mathrm{possde}(V, \mathcal{G})$, $\mathrm{pa}(V, \mathcal{G})$, $\mathrm{ch}(V, \mathcal{G})$, $\mathrm{sib}(V, \mathcal{G})$, respectively. For a set of nodes $\mathbf{W}$, the set $\mathrm{an}(\mathbf{W}, \mathcal{G})$ is defined as $\bigcup_{W \in \mathbf{W}} \mathrm{an}(W, \mathcal{G})$, with analogous definitions for $\mathrm{possan}(\mathbf{W}, \mathcal{G})$, $\mathrm{de}(\mathbf{W}, \mathcal{G})$, $\mathrm{possde}(\mathbf{W}, \mathcal{G})$, $\mathrm{pa}(\mathbf{W}, \mathcal{G})$, $\mathrm{ch}(\mathbf{W}, \mathcal{G})$ and $\mathrm{sib}(\mathbf{W}, \mathcal{G})$.

Colliders, definite-status paths and v-structures. A non-endpoint node V is a *collider* on a path p if both edges adjoining V on p have arrowheads at V, i.e. $\to V \leftarrow$, $\leftrightarrow V \leftarrow$, $\to V \leftrightarrow$, $\leftrightarrow V \leftrightarrow$. A non-endpoint node V is a *non-collider* on a path p if at least one of the edges adjoining V on p is out of V, i.e. $\to V \to$, $- V \to$, $\leftrightarrow V \to$, $\leftarrow V \to$, $\leftarrow V \leftrightarrow$, $\leftarrow V -$, $\leftarrow V \leftarrow$, or if both edges adjoining V on p are undirected edges and the

two nodes adjacent to V on p are not adjacent to each other. A *definite-status path* is a path on which every non-endpoint is either a collider or a non-collider. In a DAG or an ADMG, all paths are of definite status. Three nodes A, B and C form a *v-structure* in a graph $\mathcal{G}$ if $A \to B \leftarrow C$ is the induced subgraph $\mathcal{G}'$ on $\{A, B, C\}$.

ADMGs, DAGs and PDAGs. A directed path from A to B, together with an edge $A \leftarrow B$ forms a *directed cycle*. A graph with only directed and bi-directed edges and without directed cycles is called an *acyclic directed mixed graph* (ADMG). A simple graph with only directed edges and without directed cycles is called a *directed acyclic graph* (DAG). A simple graph with only directed and undirected edges containing no directed cycles is called a *partially directed acyclic graph* (PDAG).

Blocking and separation. (Richardson, 2003; Maathuis and Colombo, 2015; Pearl, 2009) A definite-status path p in an ADMG or PDAG $\mathcal{G}$ is *blocked* by a node set $\mathbf{C}$ if (i) p contains a non-collider in $\mathbf{C}$ or (ii) p contains a collider that is not in $\text{an}(\mathbf{C}, \mathcal{G})$. Otherwise the path p is *open* given $\mathbf{C}$. Node sets $\mathbf{A}$ and $\mathbf{B}$ are said to be *m-separated* given a set $\mathbf{C}$ if every path between an $A \in \mathbf{A}$ and a $B \in \mathbf{B}$ is blocked by $\mathbf{C}$. We then write $\mathbf{A} \perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C}$. In DAGs, m-separation is called d-separation.

Markov equivalence and CPDAGs. (Andersson et al., 1997) The *(Markov) equivalence class* of a DAG $\mathcal{D}$ is the set of DAGs that imply the same d-separation relationships as $\mathcal{D}$. These are all DAGs with the same adjacencies and v-structures Verma and Pearl (1990). Markov equivalence classes can be represented as *completed partially directed acyclic graphs* (CPDAGs), which are simple graphs with directed or undirected edges, without directed cycles and with certain restrictions regarding the patterns of edges that can occur. The equivalence class represented by a CPDAG $\mathcal{G}$ is denoted by $[\mathcal{G}]$. A directed edge $A \to B$ in $\mathcal{G}$ means that this edge is present in all DAGs in the equivalence class $[\mathcal{G}]$. An undirected edge $A - B$ in $\mathcal{G}$ means that A and B are adjacent in every DAG in $[\mathcal{G}]$ and there is at least one DAG in $[\mathcal{G}]$ with $A \to B$ and at least one with $A \leftarrow B$.

Meek's rules and maxPDAGs. (Perković et al., 2017) Certain subsets of equivalence classes of DAGs can be represented by maximally oriented PDAGs (maxPDAGs), which are PDAGs with edge orientations that are closed under the orientation rules in Figure 8 (*Meek's rules*, Meek (1995)). The set of DAGs represented by a maxPDAG $\mathcal{G}$ is denoted by $[\mathcal{G}]$. The edges in maxPDAGs have the same interpretation as in CPDAGs. DAGs and CPDAGs are special cases of maxPDAGs.

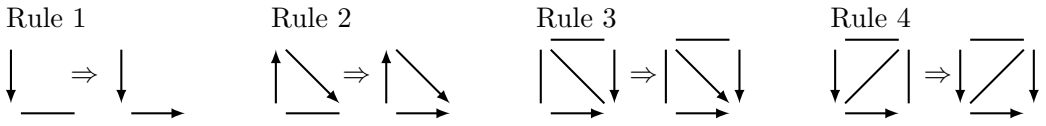


Figure 8: Meek's orientation rules. Let $\mathcal{G}$ be a simple graph with only directed and undirected edges and without directed cycles. If the graph on the left is an induced subgraph of $\mathcal{G}$, then orient the undirected edges in $\mathcal{G}$ according to the graph on the right (Meek, 1995). The rules prevent directed cycles and new v-structures.

Partial topological ordering. Let $\mathcal{D}$ be a DAG with node set $\mathbf{V}$ and let $\mathbf{V}_1, \dots, \mathbf{V}_p$ be a partition of $\mathbf{V}$. Then $\mathbf{V}_1 < \cdots < \mathbf{V}_p$ is a *partial topological ordering* of $\mathbf{V}$ if for every $i > j$, there are no directed edges from $\mathbf{V}_i$ to $\mathbf{V}_j$.

Independence and faithfulness. For sets of random variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, if $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, we write $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$. A joint density $f(\mathbf{v})$ over a set of random variables $\mathbf{V}$ is *Markov* with respect to a DAG $\mathcal{D}$ with node set $\mathbf{V}$ if for disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, $\mathbf{X} \perp_{\mathcal{D}} \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$; the density $f(\mathbf{v})$ is *faithful* to $\mathcal{D}$ if also $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp_{\mathcal{D}} \mathbf{Y} \mid \mathbf{Z}$.

Causal DAGs, CPDAGs, maxPDAGs and ADMGs. Intuitively, a *causal DAG* is a DAG where an edge $A \to B$ means that A is a direct cause of B (relative to the variables included). This can be formalised using the intervention operator, denoted by $do(\cdot)$ in Pearl (2009). For random variables $\mathbf{V}$ and $\mathbf{X} \subseteq \mathbf{V}$, the *post-intervention density* $f(\mathbf{v} \mid do(\mathbf{x}'))$ is the joint density of $\mathbf{V}$ in a (hypothetical) experiment that fixes $\mathbf{X}$ to $\mathbf{x}'$ for everyone in the population by an external intervention. A joint density $f(\mathbf{v})$ is *compatible with a causal DAG* $\mathcal{D} = (\mathbf{V}, \mathbf{E})$ if for all $\mathbf{X} \subseteq \mathbf{V}$, the post-intervention density $f(\mathbf{v} \mid do(\mathbf{x}'))$ can be written as

$$f(\mathbf{v} \mid do(\mathbf{x}')) = \mathbf{1}(\mathbf{x} = \mathbf{x}') \prod_{V \in \mathbf{V} \setminus \mathbf{X}} f(v \mid \mathrm{pa}(V, \mathcal{D})),$$

where $\mathbf{1}(\mathbf{x} = \mathbf{x}')$ is the indicator function that is 1 if $\mathbf{x} = \mathbf{x}'$ and 0 otherwise. This is known as the truncated factorisation formula (Spirtes et al., 2000; Pearl, 2009). A CPDAG or maxPDAG $\mathcal{G}$ is called a *causal CPDAG* or *causal maxPDAG* if $[\mathcal{G}]$ contains a causal DAG. A *causal ADMG* is an ADMG that has been obtained by subjecting a causal DAG to a latent projection, see Definition 4.

(Possibly) causal nodes and forbidden nodes. See Section 2.

Valid adjustment sets and amenability. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be disjoint sets of random variables, where $\mathbf{Z}$ is possibly empty. Then $\mathbf{Z}$ is a *valid adjustment set* relative to $(\mathbf{X}, \mathbf{Y})$ if we have

$$f(\mathbf{y} \mid do(\mathbf{x})) = \begin{cases} f(\mathbf{y} \mid \mathbf{x}) & \text{if } \mathbf{Z} = \emptyset, \\ \int_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) f(\mathbf{z}) d\mathbf{z} & \text{otherwise.} \end{cases} \tag{1}$$

Relative to a causal DAG, CPDAG, maxPDAG or ADMG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, a valid adjustment set is defined as follows: Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be disjoint subsets of $\mathbf{V}$, where $\mathbf{Z}$ is possibly empty. Then $\mathbf{Z}$ is a *valid adjustment set* relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if equation (1) holds for every joint density $f(\mathbf{v})$ compatible with $\mathcal{G}$ (Perković et al., 2018). Further, $\mathcal{G}$ is said to be *amenable* for adjustment relative to $(\mathbf{X}, \mathbf{Y})$ if every proper possibly causal path from $\mathbf{X}$ to $\mathbf{Y}$ starts with a directed edge out of $\mathbf{X}$ (Perković et al., 2018).

Generalised adjustment criterion. (Perković et al., 2017, 2018; Shpitser et al., 2010) Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be disjoint node sets in a causal DAG, CPDAG, maxPDAG or ADMG $\mathcal{G}$. Then $\mathbf{Z}$ is a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if and only if the following three conditions hold:

(a) $\mathcal{G}$ is amenable relative to $(\mathbf{X}, \mathbf{Y})$,

(b) $\mathbf{Z} \cap \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \emptyset$,

(c) all proper non-causal definite-status paths from $\mathbf{X}$ to $\mathbf{Y}$ are blocked by $\mathbf{Z}$.

Causal linear model. Let $\mathcal{D}$ be a causal DAG with node set $\mathbf{V} = (V_1, \cdots, V_p)$. Then $\mathbf{V}$ is said to follow a *causal linear model* compatible with $\mathcal{D}$ if the distribution of each $V_i \in \mathbf{V}$ can be described by an equation of the form

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

with $\alpha_{ij} \in \mathbb{R}$ and ϵ_{v_i} a random variable with mean 0 and finite variance such that $\epsilon_{v_1}, \ldots, \epsilon_{v_p}$ are jointly independent. For a causal CPDAG or maxPDAG $\mathcal{G}$, $\mathbf{V}$ is said to follow a causal linear model compatible with $\mathcal{G}$ if $\mathbf{V}$ follows a causal linear model compatible with a DAG in $[\mathcal{G}]$.

Partial variance notation. Consider a random variable S and a random vector $\mathbf{T}$. We denote the covariance matrix of $\mathbf{T}$ by $\Sigma_{\mathbf{tt}}$ and the row vector of covariances between S and $\mathbf{T}$ by $\Sigma_{s\mathbf{t}}$. The partial variance of S given $\mathbf{T}$ is defined as $\sigma_{ss.\mathbf{t}} = \text{Var}(S) - \Sigma_{s\mathbf{t}}\Sigma_{\mathbf{tt}}^{-1}\Sigma_{s\mathbf{t}}^{-1}$.

Asymptotic variance. Consider a sequence of estimators $(\hat{\beta}_n)_{n\in\mathbb{N}}$ such that $\sqrt{n}(\hat{\beta}_n - \beta)$ converges in distribution to $\mathcal{N}(0, v)$. We call v the asymptotic variance of $\hat{\beta}$ and write $a.var(\hat{\beta}) = v$.

Appendix B. Proofs for Section 3

In this appendix, we prove our claims about the forbidden projection made in Section 3.

Proposition 6 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$ *such that* $\mathbf{Y} \subseteq \text{de}(\mathbf{X}, \mathcal{D})$. *Then a valid adjustment set relative to* $(\mathbf{X},\mathbf{Y})$ *in* $\mathcal{D}$ *exists if and only if there is no bi-directed edge between any* $X \in \mathbf{X}$ *and* $Y \in \mathbf{Y}$ *in* $\mathcal{D}^{\mathbf{XY}}$.

Proof We show that a valid adjustment set relative to $(\mathbf{X},\mathbf{Y})$ in $\mathcal{D}$ *cannot* exist if and only if there is a bi-directed edge between a $X \in \mathbf{X}$ and a $Y \in \mathbf{Y}$ in the forbidden projection $\mathcal{D}^{\mathbf{XY}}$.

Assume first that there is a bi-directed edge in $\mathcal{D}^{\mathbf{XY}}$ between some $X \in \mathbf{X}$ and a $Y \in \mathbf{Y}$. Then according to Definition 4 of the latent projection there is a path in $\mathcal{D}$ between X and Y on which all nodes are non-colliders and contained in the forbidden set. This constitutes a non-causal path that cannot be blocked by any sets of nodes that are not forbidden. Hence no valid adjustment set relative to $(\mathbf{X},\mathbf{Y})$ and $\mathcal{D}$ exists.

Assume now that there is no valid adjustment set relative to $(\mathbf{X},\mathbf{Y})$ in $\mathcal{D}$. Then Lemma 15 implies $\mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D}), \mathcal{D}) \neq \emptyset$. Let $X^* \in \mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D}), \mathcal{D})$. Then there must exist a node $C^* \in \text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$ and a node $Y^* \in \mathbf{Y}$ such that there is a path of the form $X^* \leftarrow \cdots \leftarrow C^* \rightarrow \cdots \rightarrow Y^*$ where all non-endpoints are non-colliders on the path and in the forbidden set. It follows from Definition 4 of the latent projection that $\mathcal{D}^{\mathbf{XY}}$ contains a bi-directed edge $X^* \leftrightarrow Y^*$. ■

Lemma 15 (Corollary 27 in Perković et al., 2018) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$ *such that* $\mathbf{Y} \subseteq \text{de}(\mathbf{X}, \mathcal{D})$. *Then a valid adjustment set relative to* $(\mathbf{X},\mathbf{Y})$ *in* $\mathcal{D}$ *exists if and only if* $\mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D}), \mathcal{D}) = \emptyset$.

Proposition 7 *Let* $\mathbf{X}$ *and* $\{Y\}$ *be disjoint node sets in a causal DAG* $\mathcal{D}$ *such that* $Y \in \mathrm{de}(\mathbf{X}, \mathcal{D})$. *Then* $\mathcal{D}^{\mathbf{X}Y}$ *is a causal DAG if and only if there exists a valid adjustment set relative to* $(\mathbf{X}, Y)$ *in* $\mathcal{D}$.

Proof First assume that $\mathcal{D}^{\mathbf{X}Y}$ is a causal DAG. Then by Proposition 6, a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{D}$ exists.

Now assume that a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{D}$ exists. We show that (1) $\mathcal{D}^{\mathbf{X}Y}$ is a DAG syntactically, i.e. a directed graph without cycles, (2) semantically, applying the d-separation criterion to sets of nodes in $\mathcal{D}^{\mathbf{X}Y}$ yields the same separations as applying the d-separation criterion to the same node sets in $\mathcal{D}$, and (3) $\mathcal{D}^{\mathbf{X}Y}$ is a causal DAG for $(\mathbf{V} \setminus \mathrm{forb}(\mathbf{X}, Y, \mathcal{D})) \cup \mathbf{X} \cup \{Y\}$.

(1) As we assume that a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{D}$ exists, it follows from Proposition 6 together with Lemma 16 that $\mathcal{D}^{\mathbf{X}Y}$ does not contain bi-directed edges. Acyclicity of latent projections is guaranteed by property 1 of Definition 4 of the latent projection: every directed edge in $\mathcal{D}(\mathbf{W})$ corresponds to a directed path in $\mathcal{D}$, hence if $\mathcal{D}^{\mathbf{X}Y}$ had a directed cycle then so would $\mathcal{D}$. It follows that $\mathcal{D}^{\mathbf{X}Y}$ is acyclic.

(2) The m-separations in a latent projection $\mathcal{D}(\mathbf{W})$ correspond to the d-separations between nodes in $\mathbf{W}$ in the original DAG $\mathcal{D}$ (Richardson et al., 2017). In our case, $\mathcal{D}(\mathbf{W}) = \mathcal{D}^{\mathbf{X}Y}$ is itself a DAG syntactically, and for DAGs m-separation and d-separation are equivalent.

(3) Since $\mathcal{D}$ is a causal DAG for the random variables $\mathbf{V}$, the truncated factorisation derived from $\mathcal{D}$ holds for all interventions $do(\mathbf{T} = \mathbf{t}')$ with $\mathbf{T} \subseteq \mathbf{V}$:

$$f(\mathbf{v} \mid do(\mathbf{t}')) = \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{V \in \mathbf{V} \setminus \mathbf{T}} f(v \mid \mathrm{pa}(V, \mathcal{D})). \tag{2}$$

We need to show that the truncated factorisation implied by $\mathcal{D}^{\mathbf{X}Y}$ holds for the joint marginal distribution of $(\mathbf{V} \setminus \mathrm{forb}(\mathbf{X}, Y, \mathcal{D})) \cup \mathbf{X} \cup \{Y\}$. We distinguish two cases. In the first case, $Y \notin \mathrm{de}(\mathbf{X}, \mathcal{D})$. This case is trivial, as the forbidden set is then empty and $\mathcal{D} = \mathcal{D}^{\mathbf{X}Y}$. For the second case, $Y \in \mathrm{de}(\mathbf{X}, \mathcal{D})$ we define the following sets: $\mathbf{A} = (\mathbf{V} \setminus \mathrm{forb}(\mathbf{X}, Y, \mathcal{D})) \cup \mathbf{X}$ is the node set of $\mathcal{D}^{\mathbf{X}Y}$ without Y, $\mathbf{C} = \mathrm{cn}(\mathbf{X}, Y, \mathcal{D}) \setminus (\mathbf{X} \cup \{Y\})$ is the set of forbidden nodes that are ancestors of Y, excluding $\mathbf{X}$ and Y, and $\mathbf{M} = \mathrm{forb}(\mathbf{X}, Y, \mathcal{D}) \setminus (\mathbf{X} \cup \{Y\} \cup \mathbf{C})$ is the forbidden set excluding $\mathbf{X}$, Y and $\mathbf{C}$, so that $\mathbf{C} \cup \mathbf{M} = \mathrm{forb}(\mathbf{X}, Y, \mathcal{D}) \setminus (\mathbf{X} \cup \{Y\})$ is the set over which we marginalise when subjecting $\mathcal{D}$ to the forbidden projection. Then the following partial topological ordering holds: $\mathbf{A} < \mathbf{C} < Y < \mathbf{M}$. (Note that Y cannot have descendants in $\mathbf{X}$, as otherwise no valid adjustment set would exist by Lemma 15.)

We can now rewrite equation (2) as

$$f(\mathbf{v} \mid do(\mathbf{t}')) = \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{A \in \mathbf{A} \setminus \mathbf{T}} f(a \mid \mathrm{pa}(A, \mathcal{D})) \prod_{C \in \mathbf{C} \setminus \mathbf{T}} f(c \mid \mathrm{pa}(C, \mathcal{D})) \\ f(y \mid \mathrm{pa}(Y, \mathcal{D}))^{\mathbf{1}(Y \notin \mathbf{T})} \prod_{M \in \mathbf{M} \setminus \mathbf{T}} f(m \mid \mathrm{pa}(M, \mathcal{D})).$$

Consider now interventions only in nodes $\mathbf{T} \subseteq (\mathbf{V} \setminus \mathrm{forb}(\mathbf{X}, Y, \mathcal{D})) \cup \mathbf{X} \cup \{Y\}$, then $\mathbf{C} \setminus \mathbf{T} = \mathbf{C}$ and $\mathbf{M} \setminus \mathbf{T} = \mathbf{M}$. Upon marginalising the above intervention distribution over

$\mathbf{M}$ the last term in the product vanishes but the remaining terms do not change:

$$f(\mathbf{a}, y, \mathbf{c} \mid do(\mathbf{t}')) = \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{A \in \mathbf{A} \setminus \mathbf{T}} f(a \mid \mathrm{pa}(A, \mathcal{D})) \prod_{C \in \mathbf{C}} f(c \mid \mathrm{pa}(C, \mathcal{D})) \\ f(y \mid \mathrm{pa}(Y, \mathcal{D}))^{\mathbf{1}(Y \notin \mathbf{T})}.$$

Further marginalising over $\mathbf{C}$, the partial topological order guarantees that the variables in $\mathbf{A}$ do not have parents in $\mathbf{C}$. This yields

$$f(\mathbf{a}, y \mid do(\mathbf{t}')) = \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{A \in \mathbf{A} \setminus \mathbf{T}} f(a \mid \mathrm{pa}(A, \mathcal{D})) \int_{\mathbf{c}} \prod_{C \in \mathbf{C}} f(c \mid \mathrm{pa}(C, \mathcal{D})) \\ f(y \mid \mathrm{pa}(Y, \mathcal{D}))^{\mathbf{1}(Y \notin \mathbf{T})} d\mathbf{c}.$$

A variable is conditionally independent of its non-descendants given its parents. All variables in $\mathbf{A} \cup \mathbf{C}$ are non-descendants of Y, hence $Y \perp\!\!\!\perp \mathbf{A} \cup \mathbf{C} \mid \mathrm{pa}(Y, \mathcal{D})$ and $f(y \mid \mathrm{pa}(Y, \mathcal{D})) = f(y \mid \mathrm{pa}(Y, \mathcal{D}) \cup \mathbf{a} \cup \mathbf{c}) = f(y \mid \mathbf{a} \cup \mathbf{c})$. The second equality holds because the parents of Y, if there are any, form a subset of $\mathbf{A} \cup \mathbf{C}$. Similarly, all variables in $\mathbf{A}$ are non-descendants of all variables in $\mathbf{C}$, hence $f(c \mid \mathrm{pa}(C, \mathcal{D})) = f(c \mid \mathrm{pa}(C, \mathcal{D}) \cup \mathbf{a})$. Further, all parents of variables in $\mathbf{C}$ are in $\mathbf{A} \cup \mathbf{C}$, hence $\prod_{C \in \mathbf{C}} f(c \mid \mathrm{pa}(C, \mathcal{D}) \cup \mathbf{a}) = f(\mathbf{c} \mid \mathbf{a})$. We obtain

$$\begin{aligned} f(\mathbf{a}, y \mid do(\mathbf{t}')) &= \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{A \in \mathbf{A} \setminus \mathbf{T}} f(a \mid \mathrm{pa}(A, \mathcal{D})) \int_{\mathbf{c}} f(\mathbf{c} \mid \mathbf{a}) f(y \mid \mathbf{a} \cup \mathbf{c})^{\mathbf{1}(Y \notin \mathbf{T})} d\mathbf{c} \\ &= \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{A \in \mathbf{A} \setminus \mathbf{T}} f(a \mid \mathrm{pa}(A, \mathcal{D})) \int_{\mathbf{c}} f(\mathbf{c}, y \mid \mathbf{a})^{\mathbf{1}(Y \notin \mathbf{T})} f(\mathbf{c} \mid \mathbf{a})^{\mathbf{1}(Y \in \mathbf{T})} d\mathbf{c} \\ &= \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{A \in \mathbf{A} \setminus \mathbf{T}} f(a \mid \mathrm{pa}(A, \mathcal{D})) f(y \mid \mathbf{a})^{\mathbf{1}(Y \notin \mathbf{T})}. \end{aligned}$$

Two things remain to be shown. First, for every $A \in \mathbf{A}$, $\mathrm{pa}(A, \mathcal{D}) = \mathrm{pa}(A, \mathcal{D}^{\mathbf{X}Y})$ because A does not have parents in the node set over which we marginalised in the projection. Second, $f(y \mid \mathbf{a}) = f(y \mid \mathrm{pa}(Y, \mathcal{D}^{\mathbf{X}Y}))$, which follows from the fact that all conditional independencies between variables in $\mathcal{D}^{\mathbf{X}Y}$ can be read off $\mathcal{D}^{\mathbf{X}Y}$ using the d-separation criterion, as we showed in part (2) of this proof. Hence, we have

$$\begin{aligned} f(\mathbf{a}, y \mid do(\mathbf{t}')) &= \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{A \in \mathbf{A} \setminus \mathbf{T}} f(a \mid \mathrm{pa}(A, \mathcal{D}^{\mathbf{X}Y})) f(y \mid \mathrm{pa}(Y, \mathcal{D}^{\mathbf{X}Y}))^{\mathbf{1}(Y \notin \mathbf{T})} \\ &= \mathbf{1}(\mathbf{t} = \mathbf{t}') \prod_{V \in (\mathbf{A} \cup \{Y\}) \setminus \mathbf{T}} f(v \mid \mathrm{pa}(V, \mathcal{D}^{\mathbf{X}Y})), \end{aligned}$$

which is exactly the truncated factorisation formula implied by $\mathcal{D}^{\mathbf{X}Y}$. Hence, $\mathcal{D}^{\mathbf{X}Y}$ is a causal DAG for the random variables $\mathbf{A} \cup \{Y\} = (\mathbf{V} \setminus \mathrm{forb}(\mathbf{X}, Y, \mathcal{D})) \cup \mathbf{X} \cup \{Y\}$. ■

Lemma 16 *Let $\mathcal{D}$ be a DAG with node set $\mathbf{V}$ and let $\mathbf{X} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{X}$ such that a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{D}$ exists. Then any edge that is in $\mathcal{D}^{\mathbf{X}Y}$ but not in $\mathcal{D}$ is a directed edge into Y.*

Proof We only consider the case where $Y \in \mathrm{de}(\mathbf{X}, \mathcal{D})$, as otherwise $\mathcal{D} = \mathcal{D}^{\mathbf{X}Y}$ and our statement follows trivially. Define $\mathbf{F} = \mathrm{forb}(\mathbf{X}, Y, \mathcal{D}) \setminus (\mathbf{X} \cup \{Y\})$.

By Definition 5, an edge present in $\mathcal{D}^{\mathbf{X}Y}$ but not in $\mathcal{D}$ only occurs if $\mathcal{D}$ contains a node $W_j \in \mathbf{V} \setminus \mathbf{F}$ that has an ancestor in $\mathbf{F}$. We show that the only node in $\mathbf{V} \setminus \mathbf{F}$ that can have an ancestor in $\mathbf{F}$ is Y.

Consider first a $W \in \mathbf{V} \setminus \mathrm{forb}(\mathbf{X}, Y, \mathcal{D})$. W does not have ancestors in $\mathbf{F}$, as otherwise W would be a forbidden node itself. Consider next a node $X \in \mathbf{X}$. X does not have ancestors in $\mathbf{F}$ either, as every node in $\mathbf{F}$ is a descendant of $\mathrm{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$, but we assume that a valid adjustment set exists relative to $(\mathbf{X}, Y)$ in $\mathcal{D}$, implying $\mathbf{X} \cap \mathrm{de}(\mathrm{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D}), \mathcal{D}) = \emptyset$ by Lemma 15. Hence, Y is the only node in $\mathbf{V} \setminus \mathbf{F}$ that can have an ancestor in $\mathbf{F}$. ■

Proposition 8 *Let $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ be disjoint node sets in a causal DAG $\mathcal{D}$. Then $\mathbf{Z}$ is a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}$ if and only if $\mathbf{Z}$ is also a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}^{\mathbf{XY}}$.*

Proof Throughout, let $\mathbf{F} = \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \setminus (\mathbf{X} \cup \mathbf{Y})$.

We first suppose that $\mathbf{Z}$ is a valid adjustment set in $\mathcal{D}$ and show that this implies that it is also a valid adjustment set in $\mathcal{D}^{\mathbf{XY}}$. Hence, $\mathbf{Z} \cap \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \emptyset$ and $\mathbf{Z} \cap \mathbf{Y} = \emptyset$, so that every node in $\mathbf{Z}$ is also a node of $\mathcal{D}^{\mathbf{XY}}$. Further, $\mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}^{\mathbf{XY}}) \subseteq \mathbf{X} \cup \mathbf{Y}$ and hence $\mathbf{Z} \cap \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}^{\mathbf{XY}}) = \emptyset$. Amenability trivially holds in both $\mathcal{D}$ and $\mathcal{D}^{\mathbf{XY}}$ by assumption.

It remains to show that every proper non-causal path from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}^{\mathbf{XY}}$ is blocked by $\mathbf{Z}$, which we do by contradiction. So suppose that $\mathbf{Z}$ is a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}$ and that there exists a proper non-causal path $p = (V_0, e_1, V_1, \ldots, e_K, V_K)$ from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}^{\mathbf{XY}}$ that is open given $\mathbf{Z}$. We denote $\mathbf{Y}^F = \mathbf{Y} \cap \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}^{\mathbf{XY}}) = \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}^{\mathbf{XY}}) \setminus \mathbf{X}$ and note that $\mathrm{de}(\mathbf{Y}^F, \mathcal{D}^{\mathbf{XY}}) \subseteq \mathbf{Y}^F$.

Let $V_L \in \mathbf{Y}$ be the first node on p that is in $\mathbf{Y}$ and consider the path segment $p' = (V_0, e_1, V_1, \ldots, e_L, V_L)$. Suppose that $L < K$ and that $V_L \in \mathbf{Y}^{\mathbf{F}}$. If p' is causal, then p must either be causal or contain a collider in $\mathbf{Y}^F$, contradicting our assumption that it is open given $\mathbf{Z}$. If $V_L \in \mathbf{Y} \setminus \mathbf{Y}^F$ then p' cannot be causal. Hence, we can suppose that $L = K$ or replace p with p' without loss of generality.

Consider now the case that $V_K \in \mathbf{Y} \setminus \mathbf{Y}^F$. This implies that all nodes on p except V_0 are not in $\mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}^{\mathbf{XY}})$. Since $\mathrm{de}(\mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) \setminus \mathbf{X}, \mathcal{D}) \subseteq \mathrm{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$ and by definition of latent projections, this implies that p is also a path in $\mathcal{D}$. As for any node V in $\mathcal{D}^{\mathbf{XY}}$, $\mathrm{de}(V, \mathcal{D}) \subseteq \mathrm{de}(V, \mathcal{D}^{\mathbf{XY}}) \cup \mathbf{F}$, and $\mathbf{Z} \cap \mathbf{F} = \emptyset$, it follows that p is also open given $\mathbf{Z}$ in $\mathcal{D}$.

Consider now the case that $V_K \in \mathbf{Y}^F$. The path p cannot be a one-edge path, as the two possible such paths would require the existence of paths in $\mathcal{D}$ implying that no valid adjustment sets relative to $(\mathbf{X}, \mathbf{Y})$ exist in $\mathcal{D}$. By the fact that $\mathrm{de}(\mathbf{Y}^F, \mathcal{D}) \subseteq \mathbf{Y}^F$, the last edge of p must be of the form $p'' = V_{K-1} \to V_K$. By the same argument and the definition of the forbidden projection, the segment $p' = (V_0, e_1, V_1, \ldots, V_{K-1})$ is also a path in $\mathcal{D}$, which by definition of p and p'' must be non-causal. The path p'' corresponds to a causal path q'' in $\mathcal{D}$, such that all nodes except for V_{K-1} on q'' are forbidden.

The path $q = p' \oplus q''$ is a proper non-causal path from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}$; we now show that it is open given $\mathbf{Z}$. Since $\mathrm{de}(V, \mathcal{D}) \subseteq \mathrm{de}(V, \mathcal{D}^{\mathbf{XY}}) \cup \mathbf{F}$, for any node $V \notin \mathbf{F}$ in $\mathcal{D}$, it follows from the fact that p' is open given $\mathbf{Z}$ in $\mathcal{D}^{\mathbf{XY}}$ that it is also open given $\mathbf{Z}$ in $\mathcal{D}$. Since $\mathbf{F} \cap \mathbf{Z} = \emptyset$,

q'' is also open given $\mathbf{Z}$. The node V_{K-1} is a non-collider on p and hence by the assumption that p is open given $\mathbf{Z}$ it follows that $V_K \notin \mathbf{Z}$. Since V_{K-1} is also a non-collider on q it follows that q is open given $\mathbf{Z}$ in $\mathcal{D}$.

We now turn to the second part of the proof showing that if a set $\mathbf{Z}$ containing no nodes in $\mathbf{F}$ is not a valid adjustment relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}$ then it is also not a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}^{\mathbf{XY}}$.

Suppose that $\mathbf{Z} \cap \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \neq \emptyset$. Since $\mathbf{Z} \cap \mathbf{F} = \emptyset$ it follows that $\mathbf{Z} \cap (\mathbf{X} \cup \mathbf{Y}) \neq \emptyset$; but this clearly implies that $\mathbf{Z}$ cannot be a valid adjustment set in $\mathcal{D}^{\mathbf{XY}}$.

Suppose now that $\mathbf{Z} \cap (\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \cup \mathbf{Y}) = \emptyset$ and that $\mathbf{Z}$ is not a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}$. This implies the existence of a proper non-causal path $p = (V_0, e_1, V_1, \ldots, e_K, V_K)$ from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}$ that is open given $\mathbf{Z}$.

Consider first the case that p contains no nodes in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. Then p also exists in $\mathcal{D}^{\mathbf{XY}}$ and by the fact that $\text{de}(V, \mathcal{D}^{\mathbf{XY}}) = \text{de}(V, \mathcal{D}) \setminus \mathbf{F}$ it follows that p is also open given $\mathbf{Z}$ in $\mathcal{D}^{\mathbf{XY}}$. Suppose now that p contains at least one node in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. Since p is open given $\mathbf{Z}$, it cannot contain a collider in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. If we suppose that all nodes on p are in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$, then its existence implies that no valid adjustment exists in $\mathcal{D}$, while the corresponding edge in the forbidden projection would imply the same for $\mathcal{D}^{\mathbf{XY}}$. Hence, we can suppose that p contains at least one node not in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$ and let V_L be the last such node. Let $p' = (V_0, e_1, V_1, \ldots, e_L, V_L)$ and $p'' = (V_L, e_{L+1}, V_1, \ldots, e_K, V_K)$. By construction, $V_{L+1} \in \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$ and since $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \setminus \mathbf{X} \subseteq \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$, it follows that p'' is causal. Thus the forbidden projection will map p'' to the path $q'' = V_L \to V_K$. This also implies that p' is non-causal.

Suppose first that $V_0 \in \mathbf{X} \cap \text{de}(\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \setminus \mathbf{X}, \mathcal{D})$. Then no valid adjustment set exists in $\mathcal{D}$. Further, there must be a bi-directed edge from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}^{\mathbf{XY}}$ and hence that no valid adjustment set exists in $\mathcal{D}^{\mathbf{XY}}$ either. We can hence suppose that $V_0 \notin \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \setminus \mathbf{X}$. This implies that all nodes on p' except V_0 are not in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. Since this implies that no node on p' is in $\text{de}(\mathbf{F}, \mathcal{D})$ it follows that p' is also a path in $\mathcal{D}^{\mathbf{XY}}$. The path $q = p' \oplus q''$ is a proper non-causal path from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{D}^{\mathbf{XY}}$. By the usual argument p' is also open given $\mathbf{Z}$ in $\mathcal{D}^{\mathbf{XY}}$ and trivially, this is also true for q''. Further, $V_L \notin \mathbf{Z}$ is a non-collider on q'' and hence, q is open given $\mathbf{Z}$ in $\mathcal{D}^{\mathbf{XY}}$. ■

Proposition 10 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint subsets of the node set* $\mathbf{V}$ *of a DAG* $\mathcal{D}$ *such that* $\mathbf{Y} \subseteq \text{de}(\mathbf{X}, \mathcal{D})$. *Then* $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D})$.

Proof We first show that $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \subseteq \mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. Let $Z \in \mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. By Definition 1, $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \cap \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \emptyset$ and hence Z is a node in $\mathcal{D}^{\mathbf{XY}}$. Furthermore, since $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \subseteq \text{pa}(\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D}), \mathcal{D})$, and $\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \subseteq \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$, there is a node $Y \in \mathbf{Y}$ such that $\mathcal{D}$ contains a directed path $Z \to \cdots \to Y$ on which all non-endpoint nodes are in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. Due to property 1 of Definition 4, this corresponds to an edge $Z \to Y$ in $\mathcal{D}^{\mathbf{XY}}$, hence $Z \in \mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D})$.

Next, we show that $\mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \subseteq \mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. Let $Z^* \in \mathbf{O}^*(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. By Definition 5, this implies that $Z^* \in \mathbf{V} \setminus (\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \cup \mathbf{X} \cup \mathbf{Y})$. Moreover, by Definition 9, there is an edge $Z^* \to Y^*$ in $\mathcal{D}^{\mathbf{XY}}$ with $Y^* \in \mathbf{Y}$. In $\mathcal{D}$, this corresponds to a directed path $Z^* \to \cdots \to Y^*$ on which all non-endpoint nodes are in $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$, and

$Z^* \notin \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. Denote the path by p. There are two cases: In the first case, p has no non-endpoint nodes, i.e. $\mathcal{D}$ contains the edge $Z^* \to Y^*$. Since we assume $\mathbf{Y} \subseteq \text{de}(\mathbf{X}, \mathcal{D})$, Y^* must be in $\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$, hence $Z^* \in \mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. In the second case, p has at least one non-endpoint node. This means that $Z^* \in \text{pa}(W, \mathcal{D})$, where $W \in \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) \setminus (\mathbf{X} \cup \mathbf{Y})$ and $W \in \text{an}(Y^*, \mathcal{D})$. Since in a DAG, all forbidden nodes are descendants of $\mathbf{X}$, we also have $W \in \text{de}(\mathbf{X}, \mathcal{D})$, and hence $W \in \text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. It follows that $Z^* \in \mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$. ■

Appendix C. Generalisation of the Forbidden Projection and the O*-Set to Amenable MaxPDAGs

In this appendix, we generalise the forbidden projection (Definition 5) and the $\mathbf{O}^*$-set (Definition 9) to amenable maxPDAGs and show that Propositions similar to 6, 8, 7 and 10 still hold for the more general definitions.

The latent projection in general (Definition 4) cannot be generalised to (amenable) maxPDAGs as marginalising does not generally result in an ADMG. As an example, consider the maxPDAG $W_1 - L \to W_2$ with latent node L. It is not clear how the projection should be constructed in this case: $W_1 \to W_2$ would give the wrong impression that W_1 is an ancestor of W_2 (instead of a *possible* ancestor), while $W_1 - W_2$ would imply that W_2 is a possible ancestor of W_1. As we will show in the following propositions, however, the latent projection can be meaningfully generalised to amenable maxPDAGs when projecting over the special case of a forbidden set.

Definition 17 (Forbidden projection for amenable maxPDAGs) *Let $\mathcal{G}$ be a maxPDAG with node set $\mathbf{V}$, and let $\mathbf{X} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{X}$ such that $\mathcal{G}$ is amenable relative to $(\mathbf{X}, Y)$. Define $\mathbf{F} = \text{forb}(\mathbf{X}, Y, \mathcal{G}) \setminus (\mathbf{X} \cup \{Y\})$. The* forbidden projection *$\mathcal{G}^{\mathbf{X}Y}$ of $\mathcal{G}$ is a graph with node set $\mathbf{V} \setminus \mathbf{F}$ and edges as follows: For distinct nodes $W_i, W_j \in \mathbf{V} \setminus \mathbf{F}$,*

1. *$\mathcal{G}^{\mathbf{X}Y}$ contains a directed edge $W_i \to W_j$ if and only if $\mathcal{G}$ contains a directed path $W_i \to \cdots \to W_j$ on which all non-endpoint nodes are in $\mathbf{F}$,*
2. *$\mathcal{G}^{\mathbf{X}Y}$ contains a bi-directed edge $W_i \leftrightarrow W_j$ if and only if $\mathcal{G}$ contains a path, with at least one non-endpoint node, of the form $W_i \leftarrow \cdots \to W_j$ on which all non-endpoints are non-colliders and in $\mathbf{F}$,*
3. *$\mathcal{G}^{\mathbf{X}Y}$ contains an undirected edge $W_i - W_j$ if and only if $\mathcal{G}$ contains $W_i - W_j$.*

Note that we restrict the definition to singleton Y. This is because with a set $\mathbf{Y}$, we run into similar construction/interpretation problems as described above. Consider, for example, an amenable maxPDAG with node set $\{X, F, Y_1, Y_2\}$ and edges $X \to Y_1 - F \to Y_2$ as well as $X \to F$. None of $Y_1 \to Y_2$, $Y_1 \leftrightarrow Y_2$ or $Y_1 - Y_2$ are correct representations of the marginal distribution.

Before generalising the $\mathbf{O}^*$-set, we now describe the properties of the forbidden projection for maxPDAGs. A key property is that if a valid adjustment set exists, the forbidden projection of a maxPDAG is itself a maxPDAG (Proposition 22). This is analogous to Proposition 7 for DAGs. Proposition 19, in analogy to Proposition 6, states that if $\mathbf{X}$ has

a causal effect on Y, then the forbidden projection can be used to check whether a valid adjustment set exists. In Proposition 25, we show that a set $\mathbf{Z}$ is a valid adjustment set in the forbidden projection if and only if it is a valid adjustment set in the original graph, which is analogous to Proposition 8.

We begin with a lemma that will allow us to use possde$(\mathbf{X}, \mathcal{G})$ and de$(\mathbf{X}, \mathcal{G})$ interchangeably when $\mathcal{G}$ is an amenable maxPDAG.

Lemma 18 *Let $p = (V_1, V_2, \ldots, V_K)$ be a possibly directed path in a maxPDAG $\mathcal{G}$ such that no node on p shares an undirected edge with V_1. Then a subsequence of p forms a directed path from V_1 to V_K in $\mathcal{G}$.*

Proof We show this by induction. By assumption, (V_1, V_2) is a subsequence of p and forms a directed path from V_1 to V_2. Now assume that a subsequence of p forms a directed path from V_1 to V_{k-1}, for $2 < k \leq K$. Denote this subsequence by $(V_1 = W_1, W_2, \ldots, W_Q = V_{k-1})$. Clearly, if $V_{k-1} \to V_k$ then $(V_1 = W_1, W_2, \ldots, W_Q = V_{k-1}, V_k)$ is a subsequence of p and forms a directed path from V_1 to V_k in $\mathcal{G}$, which is what we wanted to show. If, on the other hand, if $V_{k-1} - V_k$ then there are four cases, three of which lead to a contradiction:

(1) The induced subgraph of $\mathcal{G}$ on $\{W_{Q-1}, W_Q = W_{k-1}, V_k\}$ is Graph 1 in Figure 9. Then $\mathcal{G}$ is not closed under Meek's Rule 1 (see Figure 8), which is a contradiction.

(2) The induced subgraph of $\mathcal{G}$ on $\{W_{Q-1}, W_Q = W_{k-1}, V_k\}$ is Graph 2 in Figure 9. Then $\mathcal{G}$ is not closed under Meek's Rule 2, which is a contradiction.

(3) The induced subgraph of $\mathcal{G}$ on $\{W_{Q-1}, W_Q = W_{k-1}, V_k\}$ is Graph 3 in Figure 9. This implies the induced subgraph of $\mathcal{G}$ on $\{W_{Q-2}, W_{Q-1}, V_k\}$ would also be graph 3 (with the same reasons as above excluding graphs 1 and 2). Repeating the argument for $W_{Q-3}, W_{Q-4}, \ldots$ implies an undirected edge between $W_1 = V_1$ and V_k, which is a contradiction.

(4) The induced subgraph of $\mathcal{G}$ on $\{W_{Q-1}, W_Q = W_{k-1}, V_k\}$ is Graph 4 in Figure 9. Then $(V_1 = W_1, W_2, \ldots, W_{Q-1}, V_k)$ is a subsequence of p and forms a directed path from V_1 to V_k, which is what we wanted to show. ■

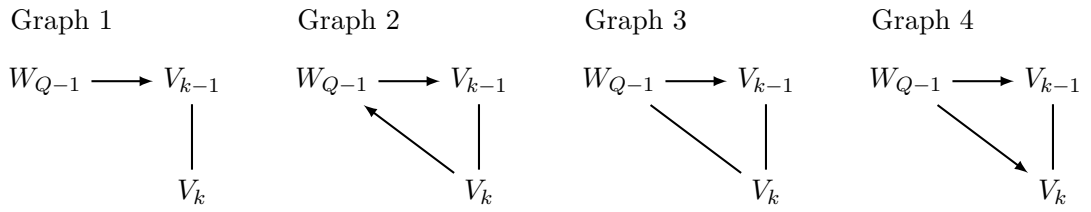


Figure 9: Graphs for the proof of Lemma 18.

Proposition 19 *Let $\mathcal{G}$ be a causal maxPDAG with node set $\mathbf{V}$, and let $\mathbf{X} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{X}$ such that $Y \in$ possde$(\mathbf{X}, \mathcal{G})$ and $\mathcal{G}$ is amenable relative to $(\mathbf{X}, Y)$. Then a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{G}$ exists if and only if there is no bi-directed edge between any nodes in the forbidden projection $\mathcal{G}^{\mathbf{X}Y}$.*

Proof By Lemma 18, $Y \in \text{possde}(\mathbf{X}, \mathcal{G})$ implies $Y \in \text{de}(\mathbf{X}, \mathcal{G})$. The proof is now analogous to the proof of Proposition 6, with Lemma 15 replaced by Lemma 20. ■

Lemma 20 *Let $\mathcal{G}$ be a causal maxPDAG with node set $\mathbf{V}$, and let $\mathbf{X} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{X}$ such that $Y \in \text{de}(\mathbf{X}, \mathcal{G})$ and $\mathcal{G}$ is amenable relative to $(\mathbf{X}, Y)$. Then a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{G}$ exists if and only if $\mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, Y, \mathcal{G}), \mathcal{G}) = \emptyset$.*

Proof We show that *no* valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{G}$ exists if and only if $\mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, Y, \mathcal{G}), \mathcal{G}) \neq \emptyset$. The proof is similar to the proof of Corollary 27 in Perković et al. (2018).

Assume first that no valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{G}$ exists, then by Lemma 21, there is a proper non-causal definite-status path from some $X \in \mathbf{X}$ to Y that is open given $\text{adjust}(\mathbf{X}, Y, \mathcal{G}) = \text{possan}(\mathbf{X} \cup \{Y\}, \mathcal{G}) \setminus (\mathbf{X} \cup \text{forb}(\mathbf{X}, Y, \mathcal{G}))$. Denote one such path by p. Assume for contradiction that p contains a collider and denote the collider by C. As p is open given $\text{adjust}(\mathbf{X}, Y, \mathcal{G})$, a descendant of C is in $\text{adjust}(\mathbf{X}, Y, \mathcal{G})$. This implies that $\text{an}(C, \mathcal{G}) \cap \text{forb}(\mathbf{X}, Y, \mathcal{G}) = \emptyset$, as otherwise all descendants of C would be in $\text{forb}(\mathbf{X}, Y, \mathcal{G})$ and could not be in $\text{adjust}(\mathbf{X}, Y, \mathcal{G})$. As $Y \in \text{forb}(\mathbf{X}, Y, \mathcal{G})$, it follows that at least one of the nodes adjacent to C on p must be a non-endpoint non-collider on p. Denote one such node by B. As $B \in \text{pa}(C, \mathcal{G})$, $B \notin \text{forb}(\mathbf{X}, Y, \mathcal{G})$ and $B \in \text{adjust}(\mathbf{X}, Y, \mathcal{G})$. But then p is not open given $\text{adjust}(\mathbf{X}, Y, \mathcal{G})$, which is a contradiction. Hence, p does not contain a collider. As p is non-causal, p cannot be directed towards Y, and as we assume that $Y \in \text{de}(\mathbf{X}, \mathcal{G})$, p cannot be directed towards X. Hence, p is a path of the form $X \leftarrow \cdots \leftarrow A \rightarrow \cdots \rightarrow Y$, where every non-endpoint is a non-collider not in $\text{adjust}(\mathbf{X}, Y, \mathcal{G})$. It follows that A is in $\text{forb}(\mathbf{X}, Y, \mathcal{G})$ and thus is a descendant of $\mathbf{X}$, which implies that $X \in \text{de}(\text{cn}(\mathbf{X}, Y, \mathcal{G}), \mathcal{G})$ and hence $\mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, Y, \mathcal{G}), \mathcal{G}) \neq \emptyset$.

Assume now that $\mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, Y, \mathcal{G}), \mathcal{G}) \neq \emptyset$. Pick a node from $\mathbf{X} \cap \text{de}(\text{cn}(\mathbf{X}, Y, \mathcal{G}), \mathcal{G})$ and denote it by X^*. Then there must exist a node $C* \in \text{cn}(\mathbf{X}, Y, \mathcal{G})$ such that there is a path of the form $X^* \leftarrow \cdots \leftarrow C^* \rightarrow \cdots \rightarrow$ where that all non-endpoint non-colliders on the path are in the forbidden set. This path cannot be blocked by any set of non-forbidden nodes. ■

Lemma 21 (Theorem 5.6 in Perković et al., 2017) *Let $\mathbf{X}$ and $\mathbf{Y}$ be disjoint node sets in a causal maxPDAG $\mathcal{G}$ such that $\mathcal{G}$ is amenable relative to $(\mathbf{X}, \mathbf{Y})$, and let $\text{adjust}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \text{possan}(\mathbf{X} \cup \mathbf{Y}, \mathcal{G}) \setminus (\mathbf{X} \cup \mathbf{Y} \cup \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}))$. Then a valid adjustment set relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{D}$ exists if and only if all proper non-causal definite-status paths from $\mathbf{X}$ to $\mathbf{Y}$ are blocked by $\text{adjust}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$ in $\mathcal{G}$.*

Proposition 22 *Let $\mathcal{G}$ be a causal maxPDAG with node set $\mathbf{V}$, and let $\mathbf{X} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{X}$ such that $\mathcal{G}$ is amenable relative to $(\mathbf{X}, Y)$ and there exists a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{G}$. Denote the set of DAGs represented by $\mathcal{G}$ by $[\mathcal{G}] = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$. Then the forbidden projection $\mathcal{G}^{\mathbf{X}Y}$ is the causal maxPDAG representing the DAGs in $\{\mathcal{D}_1^{\mathbf{X}Y}, \mathcal{D}_2^{\mathbf{X}Y}, \ldots, \mathcal{D}_M^{\mathbf{X}Y}\}$.*

Proof We only consider the case that $Y \in \text{possde}(\mathbf{X}, \mathcal{G})$, as otherwise the proposition follows trivially from the fact that $\mathcal{G}^{\mathbf{X}Y} = \mathcal{G}$. By Lemma 18, $Y \in \text{possde}(\mathbf{X}, \mathcal{G})$ implies $Y \in \text{de}(\mathbf{X}, \mathcal{G})$. We know from Propositions 6 and 19 that none of $\mathcal{G}^{\mathbf{X}Y}, \mathcal{D}_1^{\mathbf{X}Y}, \mathcal{D}_2^{\mathbf{X}Y}, \ldots, \mathcal{D}_M^{\mathbf{X}Y}$ contain any bi-directed edges. Consider edges present in the latent projections but not in the original graphs: For the maxPDAG $\mathcal{G}$, denote the set of edges in $\mathcal{G}^{\mathbf{X}Y}$ but not in $\mathcal{G}$ by $\mathbf{e}(\mathcal{G})$, and define analogous sets $\mathbf{e}(\mathcal{D}_1), \mathbf{e}(\mathcal{D}_2), \ldots, \mathbf{e}(\mathcal{D}_M)$ for the DAGs $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M$. None of $\mathbf{e}(\mathcal{G}), \mathbf{e}(\mathcal{D}_1), \mathbf{e}(\mathcal{D}_2), \ldots, \mathbf{e}(\mathcal{D}_M)$ contain any undirected edges. Further, any directed edges in any of $\mathbf{e}(\mathcal{G}), \mathbf{e}(\mathcal{D}_1), \mathbf{e}(\mathcal{D}_2), \ldots, \mathbf{e}(\mathcal{D}_M)$ are into Y. This is because for every $\mathcal{G}' \in \{\mathcal{G}, \mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$, an edge in $\mathbf{e}(\mathcal{G}')$ corresponds to a directed path with at least one forbidden non-endpoint node in $\mathcal{G}'$. If an edge in $\mathbf{e}(\mathcal{G}')$ was into a node $V \in \mathbf{V} \setminus (\mathbf{X} \cup \{Y\})$, then V would itself be forbidden, which is a contradiction. If an edge in $\mathbf{e}(\mathcal{G}')$ was into a node $X \in \mathbf{X}$, then X would be in $\text{de}(\text{cn}(\mathbf{X}, Y, \mathcal{G}'), \mathcal{G}')$, which by Lemma 20 contradicts our assumption that a valid adjustment set exists relative to $(\mathbf{X}, Y)$ in $\mathcal{G}'$. Hence, all edges in all of $\mathbf{e}(\mathcal{G}), \mathbf{e}(\mathcal{D}_1), \mathbf{e}(\mathcal{D}_2), \ldots, \mathbf{e}(\mathcal{D}_M)$ are into Y. In fact, by Lemma 23 below, $\mathbf{e}(\mathcal{G}) = \mathbf{e}(\mathcal{D}_1) = \mathbf{e}(\mathcal{D}_2) = \cdots = \mathbf{e}(\mathcal{D}_M) = \mathbf{e}$. The graphs $\mathcal{G}^{\mathbf{X}Y}, \mathcal{D}_1^{\mathbf{X}Y}, \mathcal{D}_2^{\mathbf{X}Y}, \ldots, \mathcal{D}_M^{\mathbf{X}Y}$ can thus be constructed by copying the induced subgraphs of $\mathcal{G}, \mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M$ with respect to $(\mathbf{V} \setminus \text{forb}(\mathbf{X}, Y, \mathcal{G})) \cup (\mathbf{X}, Y)$ and adding the edges in $\mathbf{e}$. Hence, $\mathcal{G}^{\mathbf{X}Y}$ represents the DAGs in $\{\mathcal{D}_1^{\mathbf{X}Y}, \mathcal{D}_2^{\mathbf{X}Y}, \ldots, \mathcal{D}_M^{\mathbf{X}Y}\}$ in the sense that every directed edge in $\mathcal{G}^{\mathbf{X}Y}$ is also present in all DAGs in $\{\mathcal{D}_1^{\mathbf{X}Y}, \mathcal{D}_2^{\mathbf{X}Y}, \ldots, \mathcal{D}_M^{\mathbf{X}Y}\}$, and for every undirected edge $V_i - V_j$ in $\mathcal{G}^{\mathbf{X}Y}$, there is at least one DAG in $\{\mathcal{D}_1^{\mathbf{X}Y}, \mathcal{D}_2^{\mathbf{X}Y}, \ldots, \mathcal{D}_M^{\mathbf{X}Y}\}$ with $V_i \to V_j$ and at least one with $V_i \leftarrow V_j$.

In order show that $\mathcal{G}^{\mathbf{X}Y}$ has all the characteristics of a maxPDAG, we show that $\mathcal{G}^{\mathbf{X}Y}$ is closed under Meek's rules. Referring to Figure 8, we argue that the graphs on the left-hand sides of Rules 1 – 4 cannot be induced subgraphs of $\mathcal{G}^{\mathbf{X}Y}$. Assume for contradiction that the left-hand graph of Rule 1, $\to -$, was an induced subgraph of $\mathcal{G}^{\mathbf{X}Y}$. As this graph is not an induced subgraph of $\mathcal{G}$ by assumption, and all of $\mathbf{e}(\mathcal{G}), \mathbf{e}(\mathcal{D}_1), \mathbf{e}(\mathcal{D}_2), \ldots, \mathbf{e}(\mathcal{D}_M)$ consist of only directed edges into Y, we can conclude that the directed edge in $\to -$ is into Y, i.e. $\to Y-$. Hence, Y shares an undirected edge with some node V in $\mathcal{G}$, but this means that V is a forbidden node in some $\mathcal{D} \in [\mathcal{G}]$, which is not allowed according to Lemma 24. By similar arguments, none of the graphs on the left-hand sides of Rules 1 – 4 in Figure 8 is an induced subgraph of $\mathcal{G}^{\mathbf{X}Y}$. ■

Lemma 23 *Let $\mathcal{G}$ be a causal maxPDAG with node set $\mathbf{V}$, and let $\mathbf{X} \subset \mathbf{V}$ and $Y \in \mathbf{V} \setminus \mathbf{X}$ such that $\mathcal{G}$ is amenable relative to $(\mathbf{X}, Y)$ and there exists a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{G}$. Define $\mathbf{F} = \text{forb}(\mathbf{X}, Y, \mathcal{G}) \setminus (\mathbf{X} \cup \{Y\})$ and pick a node $V_1 \in \mathbf{V} \setminus \mathbf{F}$. Then the following two statements are equivalent:*

(i) A DAG $\mathcal{D} \in [\mathcal{G}]$ contains a directed path $p = (V_1, V_2, \ldots, V_K = Y)$ such that all non-endpoint nodes on p are in $\mathbf{F}$.

(ii) The maxPDAG $\mathcal{G}$ contains a directed path $q = (V_1 = W_1, W_2, \ldots, W_Q = Y)$ such that all non-endpoint nodes on q are in $\mathbf{F}$.

Proof Statement (ii) implies that the directed path p is present in all DAGs in $[\mathcal{G}]$ by the defining properties of a maxPDAG. Hence, we only show that (i) implies (ii). Again by the properties of maxPDAGs, the sequence of nodes $(V_1, V_2, \ldots, V_K = Y)$ forms a possibly

directed path from V_1 to Y in $\mathcal{G}$. We first show that no node in $\{V_2, \dots, V_K = Y\}$ shares an undirected edge with V_1. Suppose, for contradiction, that node $V_k, 2 \leq k \leq K$ shares an undirected edge with V_1 and distinguish two cases: (1) $V_1 \in \mathbf{X}$, (2) $V_1 \in \mathbf{V} \setminus \mathbf{X}$. The first case contradicts our assumption that $\mathcal{G}$ is amenable relative to $(\mathbf{X}, Y)$. The second case implies that V_1, as a possible descendant of V_k, is in $\mathbf{F}$, but we chose V_1 such that $V_1 \in \mathbf{V} \setminus \mathbf{F}$. Hence, no node in $\{V_2, \dots, V_K = Y\}$ shares an undirected edge with V_1. We can thus apply Lemma 18 and conclude that a subsequence of $(V_1, V_2, \dots, V_K = Y)$ forms a directed path from V_1 to Y in $\mathcal{G}$, which implies that statement (ii) holds. ■

Lemma 24 (Lemma E.8 in HPM19) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint node sets in a maxPDAG* $\mathcal{G}$*, such that* $\mathcal{G}$ *is amenable relative to* $(\mathbf{X}, \mathbf{Y})$*, and let* $\mathcal{D}$ *be a DAG in* $[\mathcal{G}]$*. Then* $\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G}) = \text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{D})$*.*

Proposition 25 *Let* $\mathcal{G}$ *be a causal maxPDAG with node set* $\mathbf{V}$ *and let* $\mathbf{X} \subset \mathbf{V}$ *and* $Y \in \mathbf{V} \setminus \mathbf{X}$ *such that* $\mathcal{G}$ *is amenable relative to* $(\mathbf{X}, Y)$*. Then a set* $\mathbf{Z}$ *is a valid adjustment set relative to* $(\mathbf{X}, Y)$ *in* $\mathcal{G}$ *if and only if it is a valid adjustment set relative to* $(\mathbf{X}, Y)$ *in the forbidden projection* $\mathcal{G}^{\mathbf{X}Y}$*.*

Proof Let $\mathcal{D} \in [\mathcal{G}]$ and $D^{\mathbf{X}Y}$ its forbidden projection. By Proposition 8, a set $\mathbf{Z}$ is a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{D}$ if and only if it is a valid adjustment set relative to $(\mathbf{X}, Y)$ in $\mathcal{D}^{\mathbf{X}Y}$. By Proposition 22, the set $[\mathcal{G}^{\mathbf{X}Y}]$ contains exactly the forbidden projections of all DAGs in $[\mathcal{G}]$. Hence if $\mathbf{Z}$ is a valid adjustment set in all $\mathcal{D} \in [\mathcal{G}]$, then it is a valid adjustment set in all $\mathcal{D}^{\mathbf{X}Y} \in [\mathcal{G}^{\mathbf{X}Y}]$ and vice versa. ■

To summarise, the forbidden projection for amenable causal maxPDAGs has very similar properties as the forbidden projection for causal DAGs, as long as we consider a singleton outcome node Y: Bi-directed edges in the projection indicate the lack of a valid adjustment set; if a valid set exists, the forbidden projection is a maxPDAG itself, preserving all the information relevant to causal effect identification via adjustment; in particular, all valid sets can be read off the forbidden projection as well as the original graph.

Finally, we now generalise Definition 9 of the $\mathbf{O}^*$-set and its optimality property in Proposition 10 to amenable maxPDAGs with singleton Y.

Definition 26 *Let* $\mathcal{G}$ *be a causal maxPDAG with node set* $\mathbf{V}$*, let* $\mathbf{X} \subset \mathbf{V}$ *and* $Y \in \mathbf{V} \setminus \mathbf{X}$ *such that* $\mathcal{G}$ *is amenable relative to* $(\mathbf{X}, Y)$*, and let* $\mathcal{G}^{\mathbf{X}Y}$ *be the corresponding forbidden projection. We define* $\mathbf{O}^*(\mathbf{X}, Y, \mathcal{G})$ *as:*

$$\mathbf{O}^*(\mathbf{X}, Y, \mathcal{G}) = \text{pa}(Y, \mathcal{G}^{\mathbf{XY}}) \setminus \mathbf{X}.$$

Proposition 27 *Let* $\mathcal{G}$ *be a causal maxPDAG with node set* $\mathbf{V}$*, let* $\mathbf{X} \subset \mathbf{V}$ *and* $Y \in \mathbf{V} \setminus \mathbf{X}$ *such that* $Y \in \text{possde}(\mathbf{X}, \mathcal{D})$ *and* $\mathcal{G}$ *is amenable relative to* $(\mathbf{X}, Y)$*, let* $\mathbf{Z}$ *be a valid adjustment set relative to* $(\mathbf{X}, Y)$ *in* $\mathcal{G}$ *and let* $\mathbf{O}^* = \mathbf{O}^*(\mathbf{X}, Y, \mathcal{G})$*. Then* $\mathbf{O}^*$ *is a valid adjustment set relative to* $(\mathbf{X}, Y)$ *in* $\mathcal{G}$ *and if the variables* $\mathbf{V}$ *follow a linear causal model compatible with* $\mathcal{G}$*, then, for every* $X_i \in \mathbf{X}$*,* $a.var(\hat{\beta}_{yx_i.\mathbf{X}_{-i}\mathbf{O}^*}) \leq a.var(\hat{\beta}_{yx_i.\mathbf{X}_{-i}\mathbf{Z}})$*.*

Proof We prove this by showing that $\mathbf{O}(\mathbf{X}, Y, \mathcal{G})$ and $\mathbf{O}^*(\mathbf{X}, Y, \mathcal{G})$ are equal and invoking Propositions 2 and 3. By Lemma 18, $Y \in \text{possde}(\mathbf{X}, \mathcal{G})$ implies $Y \in \text{de}(\mathbf{X}, \mathcal{G})$. Then the equivalence follow directly from Lemma 28 in combination with Proposition 10: For every DAG $\mathcal{D} \in [\mathcal{G}]$, $\mathbf{O}(\mathbf{X}, Y, \mathcal{G}) = \mathbf{O}(\mathbf{X}, Y, \mathcal{D}) = \mathbf{O}^*(\mathbf{X}, Y, \mathcal{D}) = \mathbf{O}^*(\mathbf{X}, Y, \mathcal{G})$. ■

Lemma 28 (Lemma E.7 in HPM19) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint node sets in a maxPDAG* $\mathcal{G}$ *such that* $\mathcal{G}$ *is amenable relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$*, let* $\mathbf{Y} \subseteq \text{possde}(\mathbf{X}, \mathcal{G})$*, and let* $\mathcal{D}$ *be a DAG in* $[\mathcal{G}]$*. Then* $\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{D}) = \mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$.

Appendix D. Proof for Section 4

Proposition 11 *Let* X *and* Y *be nodes in a causal CPDAG or maxPDAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$*, such that* $\mathbf{V}$ *follows a causal linear model compatible with* $\mathcal{G}$ *with Gaussian errors. Let* $\widehat{\Theta}^{\mathbf{P}}$ *and* $\widehat{\Theta}^{\mathbf{O}}$ *be the multisets returned by semi-local IDA and optimal IDA respectively, applied to* X*,* Y *and* $\mathcal{G}$*, with the subsets of* $\text{sib}(X, \mathcal{G})$ *considered in the same order for both. Then, for* $i \in \{1 \dots, k\}$*, with* $k = |\widehat{\Theta}^{\mathbf{P}}| = |\widehat{\Theta}^{\mathbf{O}}|$*,*

1. $\mathbb{E}[\widehat{\Theta}_i^{\mathbf{P}}] = \mathbb{E}[\widehat{\Theta}_i^{\mathbf{O}}]$ *and*
2. $a.var(\widehat{\Theta}_i^{\mathbf{P}}) \geq a.var(\widehat{\Theta}_i^{\mathbf{O}})$.

Proof Consider any set $\mathbf{S}_i \subseteq \text{sib}(X, \mathcal{G})$. Perković et al. (2017) showed that there exists a DAG $\mathcal{D} \in [\mathcal{G}]$, such that $\mathbf{P}_i = \text{pa}(X, \mathcal{D}) = \mathbf{S}_i \cup \text{pa}(X, \mathcal{G})$, if and only if directing the edges in the neighbourhood of X according to $\mathbf{P}_i$ and applying Meek's orientation rules results in a valid maxPDAG $\mathcal{G}_i'$. If this is not the case for $\mathbf{S}_i$, both algorithms discard $\mathbf{S}_i$ at their respective line 8. We can hence suppose that there exists a DAG $\mathcal{D} \in [\mathcal{G}]$, such that $\mathbf{P}_i = \text{pa}(X, \mathcal{D}) = \mathbf{S}_i \cup \text{pa}(X, \mathcal{G})$.

Suppose that $Y \in \text{possde}(X, \mathcal{G}_i')$. In this case $\widehat{\Theta}_i^{\mathbf{O}} = \hat{\beta}_{yx.\mathbf{o}_i}$, where $\mathbf{O}_i = \mathbf{O}(X, Y, \mathcal{G}_i')$. As $\mathcal{G}_i'$ is amenable by construction, it follows from Lemma 28 in Appendix C that $\mathbf{O}_i = \mathbf{O}(X, Y, \mathcal{D})$. Further, $Y \in \text{possde}(X, \mathcal{G}_i')$ implies that $Y \notin \text{pa}(X, \mathcal{G}_i')$ and thus $\widehat{\Theta}_i^{\mathbf{P}} = \hat{\beta}_{yx.\mathbf{p}_i}$. By Proposition 2, $\mathbf{O}_i$ is a valid adjustment set relative to (X, Y) in $\mathcal{D}$, and clearly the same holds for $\mathbf{P}_i$. Since we suppose multivariate Gaussianity, this implies that both $\hat{\beta}_{yx.\mathbf{o}_i}$ and $\hat{\beta}_{yx.\mathbf{p}_i}$ are consistent estimators of $\tau_{yx}(\mathcal{D})$, and $\mathbb{E}[\hat{\beta}_{yx.\mathbf{o}_i}] = \mathbb{E}[\hat{\beta}_{yx.\mathbf{p}_i}] = \tau_{yx}(\mathcal{D})$.

Further, by Lemmas E.4 and E.5 of the Supplement of HPM19, $\mathbf{P}_i \setminus \mathbf{O}_i$ is conditionally independent of Y given $\{X\} \cup \mathbf{P}_i$, and $\mathbf{O}_i \setminus \mathbf{P}_i$ is conditionally independent of X given $\mathbf{P}_i$, respectively. These two independencies allow us to invoke Lemma C.2 of HPM19 and conclude that $\sigma_{yy.x\mathbf{o}_i} \leq \sigma_{yy.x\mathbf{p}_i}$ as well as $\sigma_{xx.\mathbf{o}_i} \geq \sigma_{xx.\mathbf{p}_i}$. As we assume a multivariate Gaussian distribution, it follows that

$$a.var(\hat{\beta}_{yx.\mathbf{o}_i}) = \frac{\sigma_{yy.x\mathbf{o}_i}}{\sigma_{xx.\mathbf{o}_i}} \leq \frac{\sigma_{yy.x\mathbf{p}_i}}{\sigma_{xx.\mathbf{p}_i}} = a.var(\hat{\beta}_{yx.\mathbf{p}_i}).$$

Suppose now that $Y \notin \text{possde}(X, \mathcal{G}_i')$. Then $Y \notin \text{de}(X, \mathcal{D})$, hence $\tau_{yx}(\mathcal{D}) = 0$. As $Y \notin \text{possde}(X, \mathcal{G}_i')$, $\widehat{\Theta}_i^{\mathbf{O}} = 0$ and as a result $a.var(\widehat{\Theta}_i^{\mathbf{O}}) = 0$. If $Y \in \text{pa}(X, \mathcal{G}_i')$, then $\widehat{\Theta}^{\mathbf{P}_i} = 0$ and as a result $a.var(\widehat{\Theta}_i^{\mathbf{P}}) = 0$. If $Y \notin \text{possde}(X, \mathcal{G}_i') \cup \text{pa}(X, \mathcal{G}_i')$, then $\widehat{\Theta}_i^{\mathbf{P}} = \hat{\beta}_{yx.\mathbf{p}_i}$ and by nature of parent sets $\mathbb{E}[\hat{\beta}_{yx.\mathbf{p}_i}] = 0$. Clearly, $a.var(\widehat{\Theta}_i^{\mathbf{P}}) > 0$ in this case. ■

Appendix E. Violin Plots

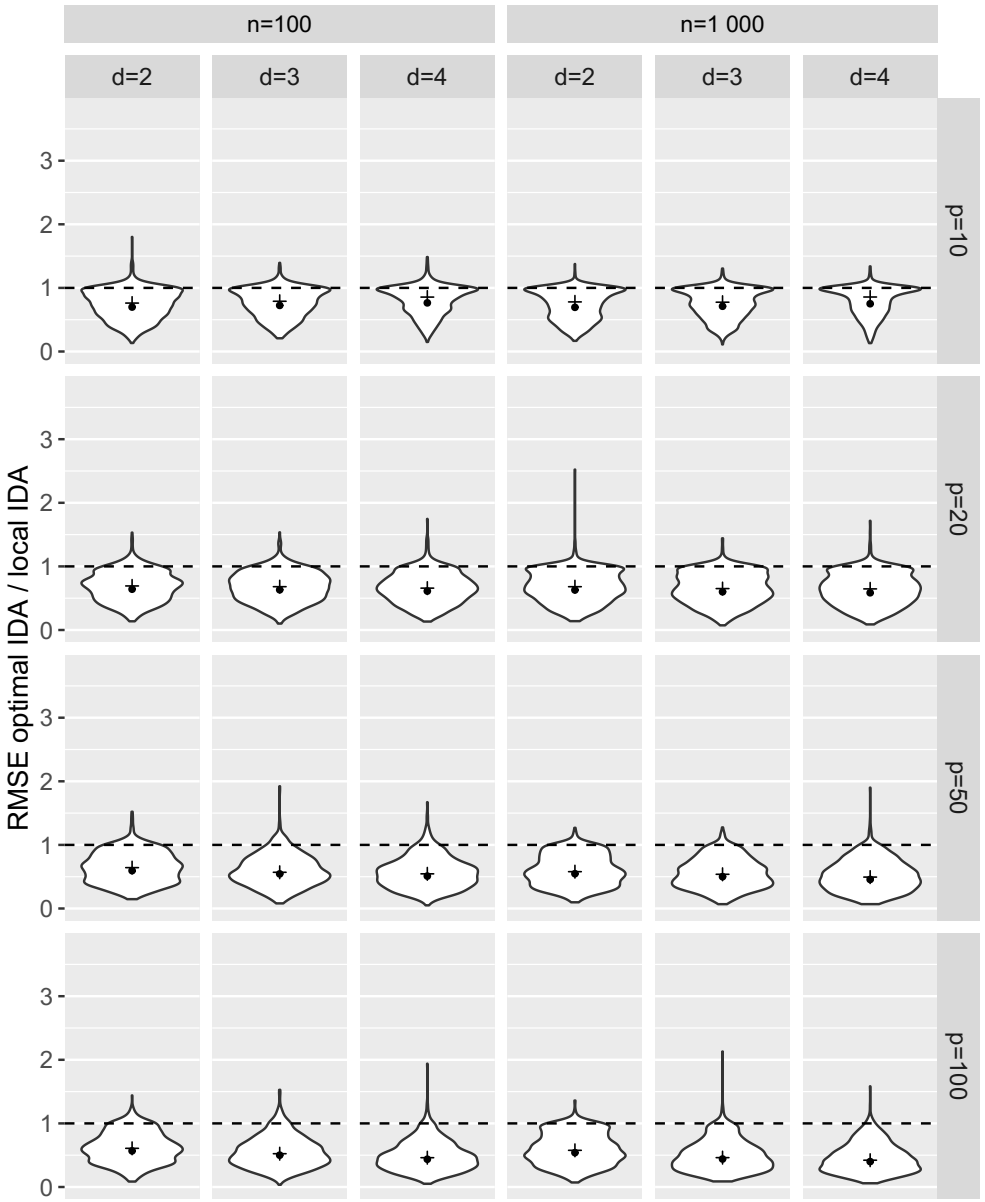


Figure 10: Violin plots of the relative mean squared error (RMSE) r over 1000 repetitions for scenarios with different numbers of nodes (p), expected number of neighbours per node (d), and sample sizes (n). Optimal and semi-local IDA were applied to the true CPDAG $\mathcal{G}$. The dots mark the geometric means, the plus signs the medians.

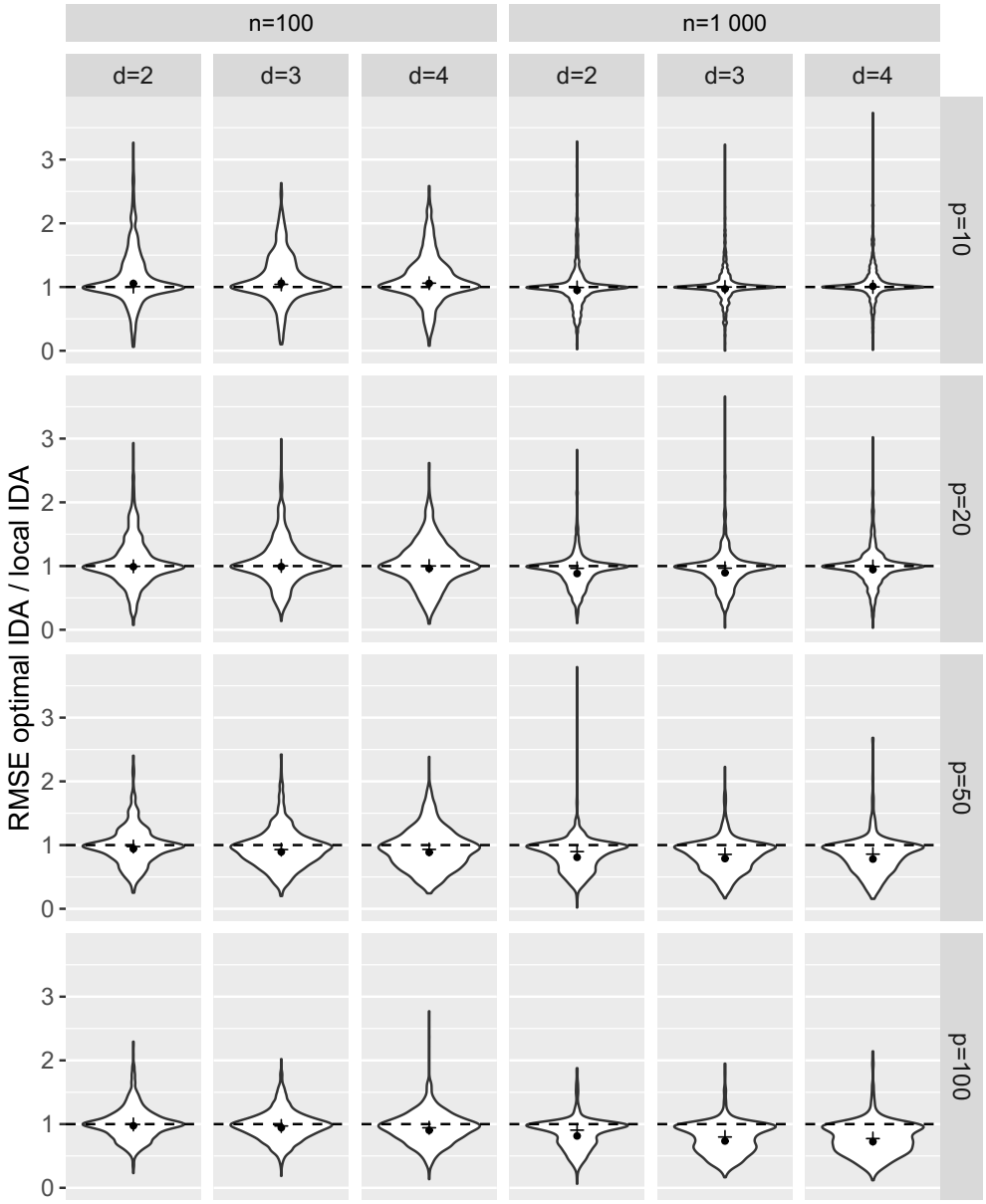


Figure 11: Violin plots of the relative mean squared error (RMSE) r^* over 1000 repetitions for scenarios with different numbers of nodes (p), expected number of neighbours per node (d), and sample sizes (n). Optimal and semi-local IDA were applied to the estimated CPDAG $\mathcal{G}^*$. The dots mark the geometric means, the plus signs the medians.

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

M. Alan Brookhart, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156, 2006.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

A. Philip Dawid and Vanessa Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010.

Xavier de Luna, Ingeborg Waernbaum, and Thomas S Richardson. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, 2011.

DeWayne Derryberry, Ken Aho, John Edwards, and Teri Peterson. Model selection and regression t-statistics. *The American Statistician*, 72(4):379–381, 2018.

Oliver Dukes and Stijn Vansteelandt. How to obtain valid tests and confidence intervals after propensity score variable selection? *Statistical Methods in Medical Research*, 29(3): 677–694, 2020a.

Oliver Dukes and Stijn Vansteelandt. Inference on treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 2020b.

Julia C. Engelmann, Thomas Amann, Birgitta Ott-Rötzer, Margit Nützel, Yvonne Reinders, Jörg Reinders, Wolfgang E. Thasler, Theresa Kristl, Andreas Teufel, Christian G. Huber, Peter J. Oefner, Rainer Spang, and Claus Hellerbrand. Causal modeling of cancer-stromal communication identifies PAPPA as a novel stroma-secreted factor activating NFκB signaling in hepatocellular carcinoma. *PLoS Computational Biology*, 11(5):e1004293, 2015.

Sander Greenland and Neil Pearce. Statistical foundations for model-based adjustments. *Annual Review of Public Health*, 36:89–108, 2015.

F. Richard Guo and Emilija Perković. Efficient least squares for estimating total effects under linearity and causal sufficiency. *arXiv preprint arXiv:2008.03481v2*, 2020.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331, 1998.

Frank E. Harrell, Jr. *Regression modeling strategies*. Springer, 5th edition, 2010.

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.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

Markus Kalisch, Alain Hauser, Martin Maechler, Diego Colombo, Doris Entner, Patrik Hoyer, Antti Hyttinen, Jonas Peters, Nicoletta Andri, Emilija Perkovic, Preetam Nandy, Philipp Ruetimann, Daniel Stekhoven, Manuel Schuerch, and Marco Eigenmann. *pcalg: Methods for Graphical Models and Causal Inference*, 2019. URL `https://CRAN.R-project.org/package=pcalg`. R package version 2.6-6.

David G. Kleinbaum and Lawrence L. Kupper. *Applied regression analysis and other multivariable methods*. Duxbury Press, 1978.

Sven Knüppel and Andreas Stang. DAG program: Identifying minimal sufficient adjustment sets. *Epidemiology*, 21(1):159, 2010.

Manabu Kuroki and Zhihong Cai. Selection of identifiability criteria for total effects by using path diagrams. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI-04)*, 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.

Thuc Duy Le, Lin Liu, Anna Tsykin, Gregory J Goodall, Bing Liu, Bing-Yu Sun, and Jiuyong Li. Inferring microRNA–mRNA causal regulatory relationships from expression data. *Bioinformatics*, 29(6):765–771, 2013.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Hannes Leeb and Benedikt M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008.

Lexin Li, R. Dennis Cook, and Christopher J. Nachtsheim. Model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):285–299, 2005.

Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42(2):413–468, 2014.

Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

Jared K. Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

Jiawei Luo, Wei Huang, and Buwen Cao. A novel approach to identify the miRNA-mRNA causal regulatory modules in cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(1):309–315, 2018.

Marloes H. Maathuis and Diego Colombo. A generalised back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. 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(4):247–248, 2010.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410, 1995.

Douglas C. Montgomery, Elizabeth A. Peck, and C. Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 5th edition, 2012.

Paul A. Murtaugh. In defense of p values. *Ecology*, 95(3):611–617, 2014.

Preetam Nandy, Marloes H. Maathuis, and Thomas S. Richardson. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674, 2017.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 411–420, 2001.

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

Emilija Perković. Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI-20)*, page ID: 229, 2020.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. A complete generalized adjustment criterion. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pages 682–691, 2015.

Emilija Perković, Markus Kalisch, and Maloes H. Maathuis. Interpreting and using CPDAGs with background knowledge. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI-17)*, page ID: 120, 2017.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

Joseph D. Ramsey. A scalable conditional independence test for nonlinear, non-gaussian data. *arXiv preprint arXiv:1401.5031v2*, 2014.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, art. arXiv:1701.06686, 2017.

Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469, 2019.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12):1393–1512, 1986.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

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

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7: 2003–2030, 2006.

Susan M. Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

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 (UAI-10)*, pages 527–536, 2010.

Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Ezequiel Smucler, Andrea Rotnitzky, and James M. Robins. A unifying approach for doubly-robust l_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737v3*, 2019.

Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:1912.00306*, 2020.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Johannes Textor and Maciej Liśkiewicz. Adjustment criteria in causal diagrams: An algorithmic perspective. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 681–688, 2011.

Jin Tian and Judea Pearl. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University, University of California, Los Angeles, 2003.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Dirk van Kampen. The Schizotypic Syndrome Questionnaire (SSQ): Psychometrics, validation and norms. *Schizophrenia Research*, 84(2–3):305–322, 2006.

Dirk van Kampen. The SSQ model of schizophrenic prodromal unfolding revised: An analysis of its causal chains based on the language of directed graphs. *European Psychiatry*, 29(7):437–448, 2014.

TS Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University, University of California, Los Angeles, 1990.

Janine Witte and Vanessa Didelez. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*, 61(5):1270–1289, 2019.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

5 Causal discovery with cohort data

This chapter is about causal discovery from cohort data, where repeated measurements have been taken in the same individuals. Using cohort data for causal discovery comes with certain advantages, but also involves challenges (Andrews, Foraita, Didelez and Witte, 2021). On the one hand, the number of subjects and the number of variables included is usually high. The former ensures that conditional independencies can be discovered with a higher power, while the latter makes the causal sufficiency assumption more plausible. Further, as cohort data are collected in waves, they have a natural time structure, which can be exploited during edge orientation. On the other hand, the variables are often measured on mixed scales (continuous, unordered categorical, ordered categorical, count, etc.). This is in contrast to e.g. gene expression data, where all variables are continuous. In addition, cohort data typically contain missing values, as participants may skip appointments, choose not to answer parts of a questionnaire or drop out of the study, i.e. contribute data up to a certain time point and then leave.

This chapter is organised as follows. In Section 5.1, I propose the tPC-algorithm, a modified version of LMPC-stable that accounts for the partial time-ordering of cohort data. I prove that tPC is sound and complete and index-stable, and point out why repeated measurements are both a curse and a blessing for constraint-based algorithms. In Section 5.2, I provide an overview about methods for handling incomplete data in the context of causal discovery, including test-wise deletion and multiple imputation. These two methods are further investigated in this chapter's paper, *Witte, Foraita and Didelez (2021)*, contained in Section 5.3. In this preprint, we present theoretical results on the recoverability of causal structures under test-wise deletion, new insights into multiple imputation for conditional independence testing, an extensive empirical comparison and an application to real data. The chapter concludes with a general discussion of challenges in causal discovery in Section 5.4.

5.1 Incorporating temporal information

As cohort data are collected in waves, they have a natural time structure. Typically, variables measured at baseline can be assumed to 'happen before' variables measured at the first follow-up examination, which in turn 'happen before' variables measured at the second follow-up, etc. More formally, the set of measured variables $\mathbf{V}$ can be partitioned into subsets $\mathbf{V}^1, \ldots, \mathbf{V}^P$ such that a smaller superscript indicates an earlier time point or *tier*, and it is assumed that variables in a given tier cannot causally influence variables in earlier tiers. If an underlying causal DAG is assumed, this corresponds to a *partial topological (node) ordering*, defined as follows:

Definition 49 (Partial topological (node) ordering)
A partial topological (node) ordering *of a DAG $\mathcal{D}$ with node set $\mathbf{V}$ is an ordered partition $\mathbf{V}^1 < \cdots < \mathbf{V}^P$ of $\mathbf{V}$ such that for $p \in \{1, \ldots, P\}$, $\mathbf{V}^p \subseteq \mathrm{nde}(\bigcup_{i=p+1}^{P} \mathbf{V}^i, \mathcal{D})$.*

In other words, edges are not directed from a later tier to an earlier tier. There are situations in which a partial topological ordering can be assumed even though the data have not been collected in waves. For examples, variables describing the environment in which the study data were collected (e.g. weather, weekday, examiner, lab) can usually be assumed to be in an earlier tier than the remaining variables. The same holds for variables measuring constant attributes of the study subjects (e.g. data of birth / age, biological sex, country of origin). Mooij et al. (2020) called such variables 'context variables'.

If background knowledge in the form of a partial topological ordering is available, it should be exploited to make the causal discovery analysis more reliable (Li and Beek, 2018; de Campos et al., 2019; Wang and Michailidis, 2019). Two general strategies are conceivable: One is to run the causal discovery algorithm as usual, and to use the background knowledge to rate the plausibility of the estimated CPDAG. The other is to modify the causal discovery algorithm to only return estimated graphs that are in agreement with the background knowledge. In this thesis, I focus on the second strategy. In particular, I investigate how to best incorporate a known partial topological ordering into the LMPC-stable algorithm (see Section 2.6.1).

Consider first the following procedure:

1. Run LMPC-stable as usual.

2. In the estimated CPDAG, find all undirected edges $V_i - V_j$ such that V_i is in an earlier tier than V_j according to the specified topological ordering, and orient them as $V_i \to V_j$.

3. Apply Meek's rules.

This procedure is sound and complete in the sense that if the input is oracle conditional independence information and a correct partial topological ordering, then the output is the maximally informative MPDAG given this input, which follows immediately from the soundness and completeness of LMPC-stable (Colombo and Maathuis, 2014) and Meek's rules (Meek, 1995a). However, when the procedure is applied to finite data, it can happen that some of the directed edges in the graph estimated in step 1 contradict the specified partial topological ordering, such that edge orientations need to be overwritten in step 2, which is inefficient. Further, the number of conditional independence tests performed is larger than necessary, as will be shown in Section 5.1.1.

Hence, I propose the tPC-algorithm, which incorporates the specified partial topological ordering as early in the algorithm as possible, and reduces the number of conditional independence tests and thus the number of opportunities for type I and type II errors to occur. The tPC-algorithm is based on LMPC-stable, with the following modifications:

1. Conditional independence testing is generally restricted such that when testing $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$ for an ordered pair (V_i, V_j), the variables in $\mathbf{S}$ must not be in a later tier than V_i. This modification has been suggested by Spirtes et al. (2000), p. 93, and similar ideas have been put forward e.g. in de Campos and Castellano (2007), Asvatourian et al. (2020) and Petersen et al. (2021), but I am not aware of a proof in the literature showing that these modifications are valid for either PC or LMPC-stable.

2. In the v-structure-detection phase (see Section 2.6.1), only those triples $V_i - V_j - V_k$ are considered as candidate v-structures where V_j is in the same tier as V_i or in the same tier as V_k.

3. Between the v-structure-orientation phase and the application of Meek's rules, undirected edges between nodes in different tiers are oriented according to the specified partial topological ordering.

Pseudocode for tPC is included in Algorithm 4 in Appendix A.3. The 't' in 'tPC' may stand for 'temporal', 'time-ordered', 'tiers' or '(partial) topological ordering'.

In the following sections, I prove that tPC is sound and complete and index-stable.

I implemented tPC in the `R`-package `tpc`, which is available on GitHub (`www.github.com/bips-hb/tpc`). The functionalities of the package are illustrated in Andrews, Foraita, Didelez and Witte, 2021. Other software packages allowing the user to specify a partial topological ordering are the Java application TETRAD (Scheines et al., 1998), the `R` package `bnlearn` (Scutari, 2010) and the `R` package `causalDisco` (Petersen et al., 2021). However, modification 1 is not implemented in `bnlearn` and only partly implemented in `causalDisco`, and in TETRAD it only affects the skeleton phase, but not the v-structure phase.

The tPC-algorithm is applied to data from the IDEFICS/I.Family cohort study (Ahrens et al., 2017) in Foraita, Witte et al. (2021).

5.1.1 tPC is sound and complete

I now show that tPC is sound and complete in the sense that it returns the maximally informative MPDAG given oracle independence information and a known, true partial topological ordering. In the proof, I use $t(V) = i$ to denote that node V is in the i-th tier, i.e. $V \in \mathbf{V}^i$.

The proof uses that a distribution is Markov to a DAG $\mathcal{D}$ if and only if the *local Markov property* holds, which is defined as follows (Lauritzen, 1996, p. 51):

Proposition 50 (Local Markov property)
A distribution P over a set of random variables $\mathbf{V}$ has the local Markov property *relative to a DAG $\mathcal{D}$ with node set $\mathbf{V}$ if for all $V \in \mathbf{V}$,*

$$V \perp\!\!\!\perp \text{nde}(V, \mathcal{D}) \setminus \text{pa}(V, \mathcal{D}) \mid \text{pa}(V, \mathcal{D}).$$

Proposition 51
Let $\mathcal{D}$ be a causal DAG with node set $\mathbf{V}$ such that the distribution of $\mathbf{V}$ is faithful to $\mathcal{D}$, and let $\mathbf{V}^1, \dots, \mathbf{V}^P$ be a partition of $\mathbf{V}$ such that $\mathbf{V}^1 < \cdots < \mathbf{V}^P$ is a partial topological ordering of $\mathcal{D}$. Then oracle tPC (Algorithm 4 in Appendix A.3) using $\mathbf{V}^1 < \cdots < \mathbf{V}^P$ as input returns the maximally informative MPDAG representing $\mathcal{D}$.

Proof. I first prove that the estimated skeleton $\mathcal{C}'$ is the true skeleton of $\mathcal{D}$ by showing that two nodes V_i and V_j are non-adjacent in $\mathcal{C}'$ if and only if they are non-adjacent in $\mathcal{D}$:

(i) Suppose first that V_i and V_j are non-adjacent in $\mathcal{C}'$. This implies that $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$ for some $\mathbf{S} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$, as otherwise the edge between V_i and V_j would not have been deleted by tPC. By faithfulness, $V_i \perp\!\!\!\perp V_j \perp_{\mathcal{D}} \mathbf{S}$, hence V_i and V_j are non-adjacent in $\mathcal{D}$.

(ii) Suppose now that V_i and V_j are non-adjacent in $\mathcal{D}$, and assume without loss of generality that $V_i \in \text{nde}(V_j, \mathcal{D})$. By the local Markov property, $V_i \perp\!\!\!\perp V_j \mid \text{pa}(V_i, \mathcal{D})$. Since tPC does not erroneously delete edges, see (i), $\text{pa}(V_i, \mathcal{D}) \subseteq \text{sib}(V_i, \mathcal{C}^*)$ for every intermediate graph estimate $\mathcal{C}^*$ in the skeleton phase. Further, for every node $P \in \text{pa}(V_i, \mathcal{D})$, $t(P) \leq t(V_i)$. This implies that the conditional independence $V_i \perp\!\!\!\perp V_j \mid \text{pa}(V_i, \mathcal{D})$ is tested by tPC, which means that the edge between V_i and V_j is deleted. Hence, V_i and V_j are non-adjacent in $\mathcal{C}'$.

Consider now phase II (detection of v-structures). Colombo and Maathuis (2014) showed that phase II of LMPC-stable is sound and complete in the sense that it detects all and only the true v-structures. Phase II of tPC differs from phase II of LMPC-stable in two aspects:

(i) Fewer node triples are considered; in particular, a triple (V_i, V_j, V_k) is not considered if (1) $t(V_j) > \max(t(V_i), t(V_k))$ or (2) $t(V_j) < t(V_i)$ or $t(V_j) < t(V_k)$, see line 22 of Algorithm 3 and Algorithm 4. However, in case (1), if the triple (V_i, V_j, V_k) forms a v-structure, then this will be oriented in phase III of tPC, and in case (2), (V_i, V_j, V_k) cannot form a v-structure, as at least one of the edges would then be directed against the partial topological ordering. Hence, no v-structures are missed by tPC by ignoring these types of triples.

(ii) Fewer conditional independence tests are performed; in particular, given the ordered pair (V_i, V_j), the test for $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$ is skipped if for any $S \in \mathbf{S}$, $t(S) > t(V_i)$, see line 24 of Algorithm 3 and lines 23, 24 and 26 of Algorithm 4. However, this does not prevent tPC from testing $V_i \perp\!\!\!\perp V_k \mid \text{pa}(V_i, \mathcal{D})$ and $V_i \perp\!\!\!\perp V_k \mid \text{pa}(V_k, \mathcal{D})$, one of which holds due to the local Markov property. Since no conflicts occur in oracle tPC, one detected conditional independence is enough to correctly identify whether the triple (V_i, V_j, V_k) forms a v-structure.

In phase III of tPC, all and only those edge orientations are added that represent the partial topological ordering. The resulting graph $\mathcal{C}'''$ is thus a partially directed graph in which all v-structures have been oriented, and additional edges have been oriented according to the partial temporal ordering.

Finally, Meek's rules are applied in phase IV. Meek (1995a) showed that the rules are sound and complete in the sense that given a partially oriented graph with

the same skeleton and v-structures as $\mathcal{D}$ and some additional edge orientations, the output is the maximally informative maxPDAG representing $\mathcal{D}$. As argued in Colombo and Maathuis (2014), the order in which Meek's rules are applied does not matter, as with oracle input there will be no conflicts. Hence, oracle tPC in its entirety is sound and complete. □

5.1.2 tPC is stable

Next, I show that the tPC-algorithm is index-stable in the sense of Colombo and Maathuis (2014).

Proposition 52
The tPC-algorithm (Algorithm 4 in Appendix A.3) is index-stable.

Proof. Consider first phase I of tPC, the skeleton phase. Colombo and Maathuis (2014) showed that the skeleton phase of LMPC-stable is index-stable. The only difference between the two skeleton phases is that in tPC, certain conditioning sets are skipped in line 12 of Algorithm 4, regardless of the sequence in which the variables appear in the dataset. Hence, the skeleton phase of tPC is order-independent as well.

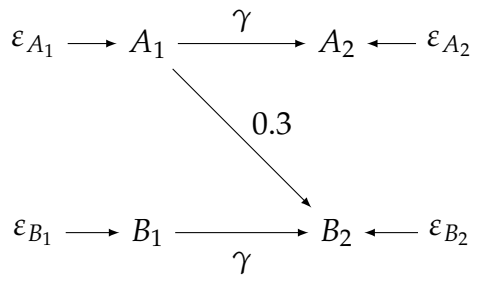
Consider now phase II of tPC, the v-structure identification phase. Colombo and Maathuis (2014) show that the v-structure phase of LMPC-stable is index-stable. The only difference between the two v-structure phases is that tPC in Algorithm 4 ignores certain triples and again skips certain conditional independence tests. As this does not depend on the sequence of the variables in the dataset, the v-structure phase of tPC is order-independent.

Next, consider phase III of tPC, the between-block edge phase. Obviously, this does not depend on the sequence of the variables either.

Finally, consider phase IV of tPC, application of Meek's rules. This is identical to the application of Meek's rules in LMPC, and was shown to be index-stable by Colombo and Maathuis (2014). Thus, tPC in its entirety is index-stable. □

5.1.3 Repeated measurements and tPC

In typical cohort studies, variables are repeatedly measured in order to obtain information about their development over time. On the one hand, the background



$$
\begin{aligned}
A_1 &\leftarrow \varepsilon_{A_1} \\
A_2 &\leftarrow \gamma \cdot A_1 + \varepsilon_{A_2} \\
B_1 &\leftarrow \varepsilon_{B_1} \\
B_2 &\leftarrow \gamma \cdot B_1 + 0.3 \cdot A_2 + \varepsilon_{B_2}
\end{aligned}
$$

Figure 13: Structure equation model for the repeated measurements simulation experiment.

knowledge provided by the natural time ordering can be exploited by causal discovery algorithms, as discussed above. On the other hand, repeated measures can also make their performance worse, as illustrated next.

Consider the structural equation model in Figure 13, where the error terms are taken to be normally distributed with mean zero and variances such that the marginal variances of A_1, A_2, B_1 and B_2 are all equal to 1. Here, A_1 and B_1 stand for two variables measured at time 1, and A_2 and B_2 denote the same variables measured at time 2. The autocorrelation between A_1 and A_2 as well as between B_1 and B_2 is determined by the edge coefficient γ. The edge coefficient of the edge between A_1 and B_2 is 0.3.

I simulated two times 10 000 datasets with 1 000 observations each from this structural equation model, setting γ to 0.3 or 0.9, respectively. The data were then analysed using (i) LMPC-stable and (ii) tPC with the specified partial topological ordering $\{A_1, B_1\} < \{A_2, B_2\}$. Both algorithms used Fisher's z-test with $\alpha = 0.05$. Table 1 contains the proportions of simulation runs in which an edge between A_1 and B_2 was present in the estimated MPDAG, irrespective of its orientation.

Table 1 shows that PC and tPC retained the edge between A_1 and B_2 in more than 80 % of simulation runs when γ was set to 0.3. In contrast, when γ was set to 0.9, the edge was erroneously removed in the majority of simulation runs. This is because the conditional correlation between A_1 and B_2 given A_2 or given B_1 (or both) is smaller when the conditioning variable is strongly correlated with either A_1 or B_2. Hence, the tests for $A_1 \perp\!\!\!\perp B_2 \mid A_2$ and $A_1 \perp\!\!\!\perp B_2 \mid B_1$ have less power when γ is large. The problem can be slightly alleviated when using tPC instead of PC, as the test for $A_1 \perp\!\!\!\perp B_2 \mid A_2$ is then skipped.

The problem of strong correlations masking weaker correlations in the context of causal discovery has been described in Steck and Tresp (1999). They propose to modify the PC-algorithm such that it obeys the 'necessary path condition', which

Table 1: Results from the repeated measures simulation experiment. Shown are the proportions of simulation runs in which an edge between A_1 and B_2 was present in the estimated MPDAG, irrespective of its orientation.

	$\gamma = 0.3$	$\gamma = 0.9$
LMPC-stable	0.81	0.26
tPC	0.83	0.31

essentially states that two nodes that are found to be marginally dependent must be connected by at least one open path in the final graph estimate. It would be interesting to know whether this modification can improve the performance of causal discovery in the presence of strong autocorrelation. An example where it *cannot* be expected to be helpful is the graph in Figure 13 but with an additional edge between A_1 and B_1 (or between A_2 and B_2), as the necessary path condition is then fulfilled even if the edge between A_1 and B_2 is removed.

5.2 Missing data in causal discovery

Missing values can be dealt with rather naturally in a maximum likelihood framework (Friedman, 1997; Didelez and Pigeot, 1998), and there is thus a large literature on missing data handling in score-based causal discovery (see Scutari, 2020, for an overview). In the context of constraint-based causal discovery, the problem of incomplete data was often ignored in the past or simple ad-hoc method were used. For example, Alekseyenko et al. (2011) removed incomplete data columns from the analysis, and Quintana (2020) removed incomplete rows. Since deleting observations is not efficient and can lead to selection bias, a number of alternative methods have been proposed. Franzin et al. (2017) suggested to impute missing values using a k-means algorithm. While this may yield more accurate results than e.g. imputing the column mean, it bears the risk of overconfident test decisions (i.e. increased type I error rates), as the uncertainty due to the missing values is not taken into account when values are singly imputed. A more advanced technique is *multiple imputation*, where multiple imputed datasets are created and used to assess the variability in the imputations (White et al., 2011; van Buuren, 2018). This was first combined with causal discovery in Foraita et al. (2020) for normally distributed data. Sokolova et al. (2017) suggested to first estimate the covariance matrix using an expectation maximisation procedure, before performing causal discovery based on the estimated covariances. Their method assumes that the variables

follow a non-paranormal distribution, which excludes e.g. discrete variables with more than two categories. A general strategy applicable to any type of data is *test-wise deletion* (Tu et al., 2019, 2020), where incomplete rows are deleted on a test-by-test basis. This is more efficient than the analysis-wide deletion of data rows. Tu et al. (2019) stated sufficient conditions under which the PC-algorithm applying test-wise deletion discovers the correct CPDAG. Gain and Shpitser (2018) studied the recoverability of graphical structures when the missingness mechanism is known.

5.3 Paper 3: *Witte, Foraita and Didelez (2021)*

In Paper 3, we further investigate test-wise deletion and multiple imputation for causal discovery. Building on Tu et al. (2019), we derive sufficient and necessary conditions for the recoverability of CPDAGs under test-wise deletion. We further extend the multiple imputation approach of Foraita et al. (2020) to categorical and mixed variables and compare different strategies for dealing with missing values using simulated and real data. Due to the additional complications arising when the data contain repeated measurements (see Section 5.1.3), the paper focusses on missing data in cross-sections and leaves drop-out to future research.

The notation and terminology in Witte et al. (2021) mostly agree with those used in this frame text, except that graphical siblings are called *neighbours*.

Own contributions

The theoretical analyses contained in the manuscript are my own work. I extended the R package `micd`, designed and implemented the simulation experiments and performed the real data analysis. I wrote the first draft of the manuscript and led the revision process.

Multiple imputation and test-wise deletion for causal discovery with incomplete cohort data

August 31, 2021

Janine Witte[1,2], **Ronja Foraita**[1], **Vanessa Didelez**[1,2]

[1] Leibniz Institute for Prevention Research and Epidemiology—BIPS
[2] University of Bremen

ABSTRACT

Causal discovery algorithms estimate causal graphs from observational data. This can provide a valuable complement to analyses focussing on the causal relation between individual treatment-outcome pairs. Constraint-based causal discovery algorithms rely on conditional independence testing when building the graph. Until recently, these algorithms have been unable to handle missing values. In this paper, we investigate two alternative solutions: Test-wise deletion and multiple imputation. We establish necessary and sufficient conditions for the recoverability of causal structures under test-wise deletion, and argue that multiple imputation is more challenging in the context of causal discovery than for estimation. We conduct an extensive comparison by simulating from benchmark causal graphs: As one might expect, we find that test-wise deletion and multiple imputation both clearly outperform list-wise deletion and single imputation. Crucially, our results further suggest that multiple imputation is especially useful in settings with a small number of either Gaussian or discrete variables, but when the dataset contains a mix of both neither method is uniformly best. The methods we compare include random forest imputation and a hybrid procedure combining test-wise deletion and multiple imputation. An application to data from the IDEFICS cohort study on diet- and lifestyle-related diseases in European children serves as an illustrating example.

Keywords: causal search, causal inference, MICE, missing values, PC-algorithm, structure learning

1 Introduction

Causal graphs have become very popular in epidemiology and other disciplines as a means to represent the causal structure among random variables (Greenland et al., 1999; Tennant et al., 2021; Morgan and Winship, 2014; Cunningham, 2021). A causal graph drawn based on background knowledge helps communicating causal assumptions, and can guide variable selection when estimating a causal effect (Didelez, 2018). In contrast, the aim of *causal discovery* is to infer a plausible graph or set of graphs from data when the causal structure is not known a priori. The estimated graphs can be used to support or challenge existing theories, to generate new hypotheses, or to estimate possible causal effects consistent with the data (Maathuis et al., 2009). Since its introduction in the 1980s, causal discovery has been applied in a variety of fields including epidemiology (Moffa et al., 2017), medical imaging (Ray et al., 2015), genome-wide association studies (Alekseyenko et al., 2011), education research (Rau and Scheines, 2012), stock market research (Bessler and Yang, 2003), linguistics (Roberts and Winters, 2013) and climate research (Ebert-Uphoff and Deng, 2012).

Popular causal discovery methods are constraint-based algorithms, which search for conditional independencies between the variables and reconstruct the causal structure so as to satisfy the constraints imposed by these independencies. A main advantage of the constraint-based approach is its flexibility. As the algorithms mainly rely on conditional independence testing, they can be applied to any type of data (continuous, categorical, ordinal, mixed etc.), as long as suitable tests are available. Moreover, constraint-based algorithms can in principle be applied even in the presence of latent variables (Spirtes et al., 2000; Zhang, 2008).

Most software implementations of constraint-based causal discovery require fully observed data as an input. Simple ways of dealing with incomplete data lead to unsatisfactory results: Under list-wise deletion, also called complete-case analysis, all incomplete records are deleted, which can severely reduce the sample size and induce selection bias. Single imputation usually leads to underestimation of standard errors. Recently, two promising new strategies have been suggested for constraint-based causal discovery with missing values: (i) test-wise deletion (Strobl et al., 2018; Tu et al., 2019, 2020), where each conditional independence test is performed using the subset of records containing complete data for all variables involved in that particular test, and (ii) multiple imputation for Gaussian data (Foraita et al., 2020).

In this paper, we formally investigate, generalise and compare test-wise deletion and multiple imputation in the context of causal discovery. Building on Tu et al. (2019), we establish necessary and sufficient conditions for the recoverability of causal graphs under test-wise deletion. Further, we extend the multiple imputation approach by Foraita et al. (2020) to discrete and mixed variables, characterise situations in which multiple imputation is expected to outperform test-wise deletion, and discuss why selecting the imputation model is challenging in causal discovery. The performance of list-wise deletion, test-wise deletion, single imputation and multiple imputation is compared on simulated and real data. Our findings are not only useful for causal discovery; they also provide insights into the general problem of conditional independence testing with missing values, e.g. necessary and sufficient conditions for identification of (in)dependencies.

1.1 Motivating example: the IDEFICS study

Our work was motivated by IDEFICS (Identification and prevention of dietary and lifestyle-induced health effects in children and infants study), a prospective cohort study including 16 229 children from eight European countries. The children were first examined in 2007/2008, and a follow-up examination took place two years later. The cohort was later extended by the I.Family study (Ahrens et al., 2017).

Designed to identify factors relating to childhood obesity and other non-communicable health conditions, the IDEFICS study included measurements on diet, lifestyle, living environment, socio-economic background and mental and physical health. Even though these factors are known to interact in a complex manner (Lee et al., 2017; Vandenbroeck et al., 2017), analyses of the IDEFICS data often focus on individual exposures and/or individual outcomes (e.g. Börnhorst et al., 2016; Hebestreit et al., 2016; Pohlabeln et al., 2017). A causal discovery analysis would therefore be a valuable addition to the analyses conducted so far.

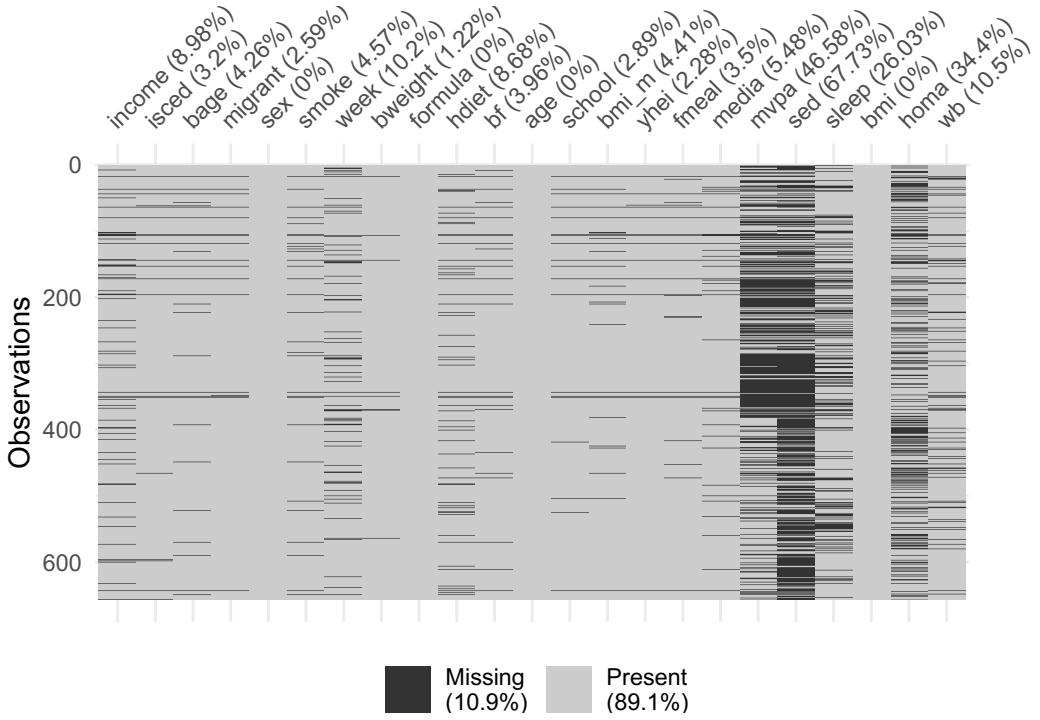


Figure 1: Missingness pattern in selected IDEFICS variables. The numbers in parentheses indicate the missingness percentage per variable.

Like most observational datasets, the IDEFICS data contain missing values. Figure 1 visualises the missingness pattern in a subsample of the IDEFICS data containing 657 children from Germany. The choice of variables roughly follows Foraita et al. (2021), who performed causal discovery on a larger subsample of the IDEFICS and I.Family data including more children and time points. See Table 1 for details on the variables used in the present paper. The overall proportion of missing data points is 10.9 %. Only 78 data rows (11.9 %) are completely observed, hence list-wise deletion would reduce the sample size by almost 90 %. It is therefore clear that a more efficient method for dealing with the missing values is needed.

1.2 Outline

The paper is organised as follows: We start with a brief review of causal graphs and constraint-based causal discovery in Section 2. In Section 3, we contrast Rubin's classification of missingness with the newer concept of the missingness graph. Section 4 contains the theoretical results on test-wise deletion and multiple imputation, and a comparison of their performance on data simulated using simple graphical structures. A comprehensive simulation study for benchmark settings is described in Section 5. Section 6 contains an application to the IDEFICS data. We conclude with a discussion in Section 7. All new methods are implemented in the `R` package `micd` available on GitHub (`www.github.com/bips-hb/micd`).

Table 1: Baseline variables of IDEFICS. The data were log-transformed as indicated in order to reduce skewness of the marginal distributions.

income	Household income (three categories)
isced	Parent's education (three categories)
bage	Mother's age in years when the child was born (continuous)
migrant	Migration status of child (binary)
sex	Sex of the child (binary)
smoke	Mother smoked during pregnancy (binary)
week	Completed weeks of pregnancy (continuous)
bweight	Birthweight in g (continuous)
formula	Child received formula milk (binary)
hdiet	Months until child was integrated into household diet (continuous, log-transformed)
bf	Total duration of breastfeeding in months (continuous, log-transformed)
age	Age in years of the child upon inclusion in the study (continuous)
school	Child visits kindergarten or school (three categories)
bmi_m	Mother's BMI (continuous)
yhei	Child's youth healthy eating score (continuous)
fmeal	Child eats breakfast at home 7 days a week (binary)
media	Child's audiovisual media consumption in hours/day (continuous)
mvpa	Child's physical activity in hours/week (continuous, log-transformed)
sed	Child's sedentary behaviour in hours/week (continuous)
sleep	Child's sleep duration in hours (continuous)
bmi	Child's BMI z-score (continuous)
homa	Child's HOMA insulin resistance index (continuous)
wb	Child's well-being score (continuous)

2 Background on causal discovery

In this section, we review causal discovery with complete data.

2.1 (Causal) graphs

We start by defining the required graphical terminology.

Nodes, edges and cycles. A *graph* consists of a set of nodes $\mathbf{V}$ and a set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. Here, graphs have at most one edge between a given pair of nodes, and edges are either directed ($\rightarrow$) or undirected ($-$). An edge from a node to itself is not allowed. Two nodes connected by an edge are *adjacent*. If $V_i \rightarrow V_j$, then V_i is a *parent* of V_j and V_j is a *child* of V_i. If $V_i - V_j$, then V_i and V_j are *neighbours*. A sequence of nodes $(V_1, \ldots, V_P)$ with $V_1 = V_P$ such that for $1 \leq i < P$, there is a directed edge $V_i \rightarrow V_{i+1}$, is called a *directed cycle*. A *directed acyclic graph* (*DAG*) is a graph with only directed edges and without directed cycles. The *skeleton* of a DAG $\mathcal{D}$ has the same nodes and adjacencies as $\mathcal{D}$, but only undirected edges.

Paths. A sequence of distinct nodes $(V_1, \ldots, V_P)$ such that for $1 \leq i < P$, V_i and V_{i+1} are adjacent, is called a *path* between V_1 and V_P. If in addition for $1 \leq i < P$, $V_i \rightarrow V_{i+1}$, then the path is *directed* from V_1 to V_P. A node V_i is a *descendant* of a node V_j if either $V_i = V_j$ or there is a directed path from V_j to V_i.

Colliders and d-separation. Consider a path $p = (V_1, \ldots, V_P)$ in a DAG $\mathcal{D}$ with node set $\mathbf{V}$. For $1 < i < P$, the node V_i is a *collider* on p if $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$; otherwise, V_i is a *non-collider* on p. The path p is *open* given a set of nodes $\mathbf{Z} \subseteq \mathbf{V}$ if (i) no non-collider on p is in $\mathbf{Z}$ and (ii) every collider on p has a descendant in $\mathbf{Z}$. Otherwise, p is *blocked* given $\mathbf{Z}$. For disjoint sets of nodes $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$, $\mathbf{X}$ and $\mathbf{Y}$ are *d-separated* by $\mathbf{Z}$ in $\mathcal{D}$ if every path between a node $X \in \mathbf{X}$ and a node $Y \in \mathbf{Y}$ is blocked given $\mathbf{Z}$. This is denoted as $\mathbf{X} \perp_{\mathcal{D}} \mathbf{Y} \mid \mathbf{Z}$.

Consider a set of random variables $\mathbf{V} = \{V_1, \ldots, V_K\}$, which can be continuous or discrete or a mix thereof. We assume that the causal structure among the variables in $\mathbf{V}$ can be represented by a *causal DAG* $\mathcal{D}$ with node set $\mathbf{V}$. In particular, we assume that the joint density $f(\mathbf{v}) = f(v_1, \ldots, v_K)$ of $\mathbf{V}$ is *Markov* and *faithful* to $\mathcal{D}$. The Markov assumption requires that $f(\mathbf{v})$ factorises as $f(\mathbf{v}) = \prod_{k=1}^{K} f(v_k \mid \mathrm{pa}(V_k, \mathcal{D}))$, where $\mathrm{pa}(V_k, \mathcal{D})$ denotes the set of parents of the node V_k in $\mathcal{D}$. Under this assumption, every d-separation $\mathbf{X} \perp_{\mathcal{D}} \mathbf{Y} \mid \mathbf{Z}$ in the graph corresponds to a conditional independence $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ in the distribution, where the latter is read as '$\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$'. The faithfulness assumption requires the reverse to be true as well, i.e. every conditional independence in the distribution corresponds to a d-separation in the graph.

In order to give the DAG $\mathcal{D}$ a causal interpretation, we additionally assume that if we intervened in the physical system underlying the random variables and fixed the value of a variable V_j in $\mathcal{D}$ to v_j, then the resulting distribution of the remaining variables would still factorise as $f(v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_K) = \prod_{k \in \{1,\ldots,K\} \setminus j} f(v_k \mid \mathrm{pa}(V_k, \mathcal{D}))$. This is plausible only if there are no latent variables, i.e. variables not in the graph, representing common causes of two or more variables in the graph. Hence, we assume the absence of such variables. This assumption in particular is known as *causal sufficiency*.

2.2 Causal discovery

The core idea of constraint-based causal discovery is to search for conditional independencies in the data, and use them to reconstruct the graph. However, as several DAGs can imply the same set of d-separations and hence conditional independencies, it is not possible in general to infer a single DAG from observational data alone, even if all above assumptions hold and the sample is infinitely large. The set of all DAGs implying a given set of d-separations is called a *(Markov) equivalence class* and can uniquely be represented by a so-called *completed partially directed acyclic graph* (*CPDAG*) with directed and undirected edges. An undirected edge in a CPDAG means that both orientations occur within the equivalence class. Without further background knowledge or parametric assumptions, constraint-based causal discovery can at best recover the true CPDAG.

In this paper, we consider the most popular constraint-based causal discovery algorithm, which is the PC-algorithm[1] (Spirtes et al., 2000). PC starts with a fully connected undirected graph and proceeds in three steps. First, a series of conditional independence tests is performed for each pair of variables (X, Y). If X and Y are found to be conditionally independent for some conditioning set, the edge between them is deleted. In order to keep the number of performed tests small, the conditioning sets are always chosen from among the nodes adjacent to X or the nodes adjacent to Y in the current graph. The resulting undirected graph is the estimated skeleton. Second, PC searches for triples of variables (X, Y, Z) such that (i) the estimated skeleton contains a path $X - Y - Z$, (ii) X and Z are not adjacent in the estimated skeleton, and (iii) X and Z are conditionally independent given a set (or several sets, see Colombo and Maathuis, 2014) of variables not containing Y. The path is then oriented as $X \rightarrow Y \leftarrow Z$. Third, additional edges are oriented according to logical rules (Meek, 1995). It can be shown that PC recovers the true CPDAG if the above assumptions of faithfulness and causal sufficiency hold and correct conditional (in)dependence information is provided (Spirtes et al., 2000). If background knowledge is available, e.g. in the form of a partial node ordering, the output of PC can be a graph with more directed edges than the CPDAG (Meek, 1995).

2.3 Conditional independence testing

Conditional independence tests commonly used for the PC-algorithm are Fisher's z-test for continuous data and the G^2-test for categorical data. Briefly, Fisher's z-test tests for a zero conditional correlation, assuming that the variables in the test follow a multivariate Gaussian distribution. The G^2-test is a non-parametric conditional independence test for contingency tables. It can also be viewed as a likelihood-ratio test under a saturated multinomial model. If a dataset contains both continuous and categorical data, common strategies are to either discretise the continuous variables, or to treat the categorical variables as continuous. For the case that the variables in the test jointly follow a Conditional Gaussian (CG) distribution (Lauritzen and Wermuth, 1989), Andrews et al. (2018) described a likelihood-ratio test, which we call the 'CG-test'. More details on Fisher's z-test, the G^2-test and the CG-test are given in Appendix A.

The significance level α for the conditional independence tests performed within the PC-algorithm has the role of a tuning parameter, where a smaller value leads to a

[1] PC was named after its inventors, Peter Spirtes and Clark Glymour.

sparser graph.

3 Missingness mechanisms and missingness graphs

Assume that a subset $\mathbf{V}^* \subseteq \mathbf{V}$ of the variables may contain missing values. For each $V \in \mathbf{V}^*$, we define a response indicator R_V that is 1 if V is observed, and 0 if V is missing. The response indicators are themselves binary random variables. We denote as $\mathbf{R}(\mathbf{V}) = \{R_V : V \in \mathbf{V}^*\}$ the set of all variable-wise response indicators. Further, for a subset $\mathbf{A} \subseteq \mathbf{V}$, we define $R^{\mathbf{A}}$ to be 1 if all variables in $\mathbf{A}$ are observed, and 0 otherwise.

In line with the literature, we assume that the missing values are not known, but exist. We refer to the distribution of the variables had all values been measured as the *full-data distribution*.

Next, we discuss two ways of describing the relation between the substantive variables $\mathbf{V}$ and the response indicators $\mathbf{R}(\mathbf{V})$, i.e. the *missingness mechanism*. The traditional classification according to Rubin (1976) (see Section 3.1) is relevant for multiple imputation, which requires the data to be missing at random. In more recent work (e.g. Mohan and Pearl, 2021), assumptions about the missingness mechanism are encoded in a causal graph over $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$ (see Section 3.2). Due to its graphical nature, this alternative framework combines well with the concept of causal discovery. In particular, it can be used to assess the identifiability of conditional (in)dependencies under test-wise deletion, see Tu et al. (2019) and Section 4.1 below.

3.1 Rubin's classification of missingness

Three classes of missingness mechanisms are often distinguished in the literature (Rubin, 1976): Values are said to be *missing completely at random* (MCAR) if $f(\mathbf{r} \mid \mathbf{v}) = f(\mathbf{r})$, i.e. missingness is independent of the substantive variables. Values are said to be *missing at random* (MAR) if, for each individual i in the dataset, $f(\mathbf{r}_i \mid \mathbf{v}_i) = f(\mathbf{r}_i \mid \mathbf{v}_i^O)$, where $\mathbf{V}_i^O$ is the set of variables that is observed for individual i. MAR thus expresses that for each individual, missingness may be associated with the observed variables, but is conditionally independent of the unobserved variables. If values are not MCAR or MAR, they are said to be *missing not at random* (MNAR). Whether values in a given dataset are MAR cannot be determined empirically. Note that the conditioning set $\mathbf{V}_i^O$ in the MAR equation may contain different variables for each individual, hence the equation corresponds to conditional independence between 'events', not between random variables (Seaman et al., 2013; Mealli and Rubin, 2015; Doretti et al., 2018). This can make the MAR assumption difficult to justify in practice. As an example, consider two incompletely observed variables *BMI* and *well-being*, where *BMI* is MAR given *well-being*. This implies that the missingness of *BMI* may depend on the value of *well-being* only in those individuals for whom *well-being* is observed, while for the other individuals, missingness of *BMI* and *well-being* must be independent.

Rubin's categories of missingness mechanisms were derived in the context of likelihood inference. For instance, under MAR, regression parameters and their standard errors can consistently be estimated in the presence of missing data using multiple imputation as discussed below.

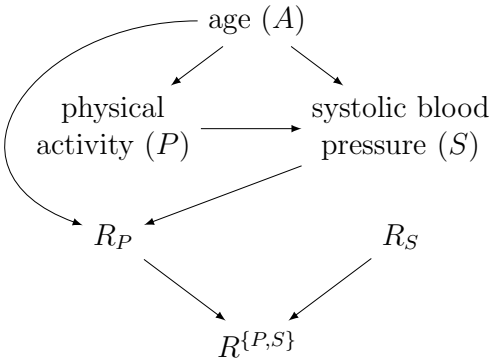


Figure 2: Example missingness DAG.

3.2 Missingness graphs

A recent line of work uses *missingness graphs* to encode assumptions about the missingness mechanism (Daniel et al., 2012; Westreich, 2012; Mohan et al., 2013; Moreno-Betancur et al., 2018; Mohan and Pearl, 2021). These graphs include both the substantive variables $\mathbf{V}$ as well as the response indicators $\mathbf{R}(\mathbf{V})$ as nodes, where it is assumed that the response indicators do not cause the substantive variables (i.e. there are no directed edges from nodes in $\mathbf{R}(\mathbf{V})$ to nodes in $\mathbf{V}$). A set-wise missingness indicator $R^{\mathbf{A}}$ as defined above is represented as a child of all nodes in $\{R_A : A \in \mathbf{A}\}$. The usual rules of d-separation can then be used to determine whether aspects of the full-data distribution are identified from the observed data. In this paper, we only consider missingness graphs that are DAGs and call them *missingness DAGs*.

Consider the missingness DAG in Figure 2 as an example. It shows three substantive variables, *age* (A), *physical activity* (P) and *systolic blood pressure* (S), together with their response indicators R_P and R_S. As *age* is assumed to be fully observed, its response indicator is omitted. According to this graph, the missingness of *physical activity* depends on *age* and *systolic blood pressure*. Using the rules of do-calculation as described in Mohan et al. (2013), it can be established e.g. that the full-data joint density $f(a, p, s)$ of the substantive variables can be identified from the incompletely observed variables as $f(a, p, s) = f(p \mid a, s, R_P = 1, R_S = 1)f(s \mid a, R_S = 1)f(a)$. Note that under a causal interpretation, the graph depicts the assumption that the nodes in the graph (including the response indicators) do not share common causes except where shown in the graph; for example, we assume that *age* is the only common cause of *physical activity* and *systolic blood pressure*.

Missingness graphs can only represent dependence relations between variables, not between events. Therefore, Rubin's MAR assumption cannot be depicted in a missingness graph. For example, it cannot be determined from the graph in Figure 2 whether *physical activity* is MAR or MNAR, since its missingness could depend, in some individuals, on *systolic blood pressure* values that are themselves missing. Mohan et al. (2013) proposed an alternative, variable-based definition of MAR, which is, however, not immediately relevant for the present paper.

4 Test-wise deletion and multiple imputation for constraint-based causal discovery

In this section, we investigate the assumptions under which conditional (in)dependencies are identified under multiple imputation or test-wise deletion, and discuss how different aspects affect the power of the conditional independence tests. As we will see, the answers are not necessarily the same as for the estimation of regression coefficients, which has been the primary focus of missing data methods.

4.1 Test-wise deletion

Consider $X \in \mathbf{V}$, $Y \in \mathbf{V} \setminus \{X\}$ and $\mathbf{Z} \subset \mathbf{V} \setminus \{X, Y\}$. Test-wise deletion means that the conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ is tested in the subsample of the data where X, Y and $\mathbf{Z}$ are fully observed (irrespective of missing values in other variables). Formally, this implies testing $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1)$, where we defined $R^{XY\mathbf{Z}} = R^{\{X,Y\}\cup\mathbf{Z}}$ for better readability. We say that a *conditional independence* in the full-data distribution is *identified under test-wise deletion* if

$$X \perp\!\!\!\perp Y \mid \mathbf{Z} \Rightarrow X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1).$$

Vice versa, we say that a *conditional dependence is identified under test-wise deletion* if

$$X \not\perp\!\!\!\perp Y \mid \mathbf{Z} \Rightarrow X \not\perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1).$$

Assume that the distribution of $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$ is faithful to a missingness DAG. Tu et al. (2019) showed that in this setting, conditional dependencies (but not independencies) are identified under test-wise deletion under an additional assumption they termed *faithful observability*:

$$X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1) \Leftrightarrow X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 0).$$

In words, an independence in the distribution underlying the data used in the test must also be present in the distribution underlying the (partially) unobserved data not used in the test. We further show in Appendix B that under faithful observability, a conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ is identified under test-wise deletion if and only if $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$. Based on these results, we next formulate a necessary and sufficient condition for the validity of the PC-algorithm using as input correct information about the (in)dependencies in the distributions under test-wise deletion ('oracle test-wise-deletion PC'). We use $\text{adj}(V, \mathcal{D})$ to denote the set of nodes adjacent to node V in DAG $\mathcal{D}$.

Definition 1 (Admissible separator condition)
Let $\mathcal{D}$ be a missingness DAG with node set $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$. We say that the admissible separator condition *holds if for all pairs (X, Y) of non-adjacent nodes in $\mathbf{V}$, there exists a (possibly empty) set $\mathbf{Z} \subset \mathbf{V}$ such that (i) $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, (ii) $\mathbf{Z} \subseteq \text{adj}(X, \mathcal{D})$ or $\mathbf{Z} \subseteq \text{adj}(Y, \mathcal{D})$ and (iii) $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$.*

Proposition 2
Let $\mathcal{D}$ be a missingness DAG with node set $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$, such that the distribution of $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$ is faithful to $\mathcal{D}$, and assume that faithful observability holds. Then oracle test-wise-deletion PC recovers the true CPDAG over $\mathbf{V}$ if and only if the admissible separator condition holds.

A proof of Proposition 2 is given in Appendix B. If faithful observability holds but the admissible separator condition does not hold, then the discovered CPDAG has additional edges compared to the true CPDAG, and may contain erroneous edge orientations. The admissible separator condition is not empirically verifiable and arguably difficult to assess in practice, where the true graph is not known. Consider the four missingness DAGs in Figure 3 for illustration. The missingness structure is the same in all graphs (i.e. Y is missing depending on the values of X and Y itself), but whether the CPDAG is correctly discovered by oracle test-wise-deletion PC under the assumptions of Proposition 2, depends on the presence or absence of the edge $X - Y$. Note that the correct CPDAG is recovered for the DAGs 1) and 2) in Figure 3 even though C is MNAR (as the missingness of Y depends on the values of Y itself). Consider also the missingness DAG in Figure 4. Here the missingness depends on fully observed variables only, which implies that the MAR assumption holds. The conditional independence $X \perp\!\!\!\perp Y \mid Z$ is not identified under test-wise deletion, however, as neither $R_Z \perp\!\!\!\perp X \mid (Y, Z)$ nor $R_Z \perp\!\!\!\perp Y \mid (X, Z)$.

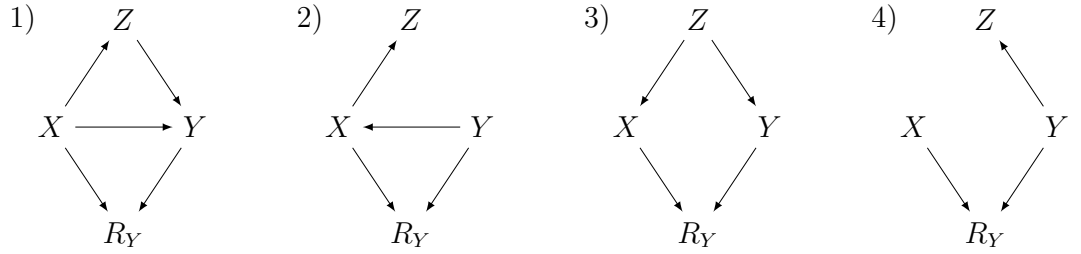


Figure 3: Four missingness DAGs with identical missingness structures. In all DAGs, $R_Y = R^{XYZ}$, as Y is the only variable containing missing values. DAGs 1 and 2: The true DAGs are such that oracle test-wise-deletion PC recovers the true CPDAG. DAGs 3 and 4: The conditional independence $X \perp\!\!\!\perp Y \mid Z$ is not identified under test-wise deletion, hence the CPDAG discovered by oracle test-wise-deletion PC using correct (in)dependence information will contain an edge between X and Y.

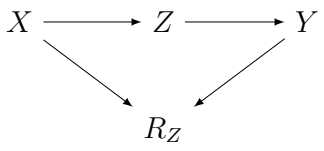


Figure 4: Example missingness DAG in which missingness depends on fully observed variables only, yet the conditional independence $X \perp\!\!\!\perp Y \mid Z$ is not identified under test-wise deletion. Oracle test-wise-deletion PC returns a fully connected graph.

Tu et al. (2019) (see also Tu et al., 2020) proposed two modifications of test-wise-deletion PC that can recover the correct CPDAG even if the admissible separator condition does not hold. Both aim at reconstructing relevant aspects of the full-data distribution. The first variant simulates values of all variables involved in the test based on models fitted to the observed data, the second variant re-weights the observed data. Both variants assume that no variable is a direct cause of its own missingness indicator, i.e. edges of the type $V_i \to R_i$ are not allowed, and that there are no edges between the missingness indicators.

4.1.1 Test-wise deletion vs. list-wise deletion

Proposition 2 holds for list-wise deletion if $R^{XY\mathbf{Z}}$ is replaced by $R^{\mathbf{V}} = \mathbf{R}(\mathbf{V})$ in the admissible separator condition. The condition then requires for a pair (X, Y) that $R^{\mathbf{V}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{\mathbf{V}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$, which is a stronger assumption than just requiring $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$. It follows that under the assumptions of Proposition 2, if oracle list-wise deletion PC recovers the true CPDAG, then oracle test-wise deletion does as well, but not the other way around. Consider now applying both variants to a given finite dataset. Then list-wise deletion PC uses only the completely observed data rows, while test-wise deletion also uses the incompletely observed rows for some of the conditional independence tests it performs. The graph discovered by test-wise-deletion PC is thus expected to be denser than the graph recovered by list-wise deletion PC, due to the larger power of some tests.

4.1.2 Parametric assumptions

So far, we have only considered non-parametric identification of conditional dependencies and independencies. In practice, parametric tests such as Fisher's z-test, which assumes that the variables follow a multivariate normal distribution, may be used. In that case, a complication arises for both list-wise and test-wise deletion, as the parametric assumptions need to hold conditionally on the response indicator being 1. As an example, suppose we have three variables *income*, *media* (measuring media consumption) and *sedentary* (measuring sedentary behaviour), and assume that the missingness mechanism is such that people with a high media consumption are less likely to answer the media question, implying $media \to R^{media}$, while the other two variables are completely observed. When applying test-wise-deletion PC using Fisher's z-test to the incomplete data, we make the following assumptions: For testing $income \perp\!\!\!\perp sedentary$, we assume that the full-data distribution of (*income*, *sedentary*) is normal, while for testing $income \perp\!\!\!\perp media$, we assume that the conditional distribution of (*income*, *media*) given $R^{media} = 1$ is normal. Under the assumed missingness mechanism, these assumptions are incompatible: If the full-data distribution of *media* is normal, then the observed distribution has a flattened right tail, since we assume that higher values are more likely to be missing. Hence at least one of these assumptions must be wrong, which potentially invalidates the type I error rate under the null hypothesis or decreases the power under the alternative.

4.2 Multiple imputation

Multiple imputation is a popular method for handling missing data especially in the context of regression analysis. It involves generating m predictions for each missing value using one of two strategies: For *joint model imputation*, a joint distribution over all

variables of interest is specified. Alternatively, separate models are specified for each incompletely observed variable given all other variables. This is called *fully conditional specification* or *multiple imputation by chained equations* (*MICE*) and is more flexible than joint model imputation when it comes to different measurement scales. In either case, Bayesian regression models are fitted and predictions are drawn from the posterior predictive distribution(s) of the missing data given the observed data. The resulting m datasets are separately analysed using the same method that would have been used in the absence of missing values. Finally, the m results are pooled according to Rubin's rules (Rubin, 1987) or other rules depending on the parameter of interest.

Standard implementations of multiple imputation rely on the MAR assumption, although known MNAR mechanisms can be accommodated as well. In addition, it is required that the modelling assumptions made in the imputation phase do not contradict the assumptions made during the analysis. This is further discussed below.

Two approaches are conceivable for combining constraint-based causal discovery with multiple imputation. One would be to estimate and pool m graphs. However, it is not clear what a good pooling method would be. The other one is to pool at the test level, as proposed by Foraita et al. (2020): First, m imputed datasets are generated using standard multiple imputation techniques. Then causal discovery is applied with the following modification: For each test, the test statistic is calculated using each of the m datasets in turn, and the m test statistics are pooled using appropriate rules. The test decision is based on the pooled statistic before going to the next test. This way, a single estimated graph is obtained.

Rubin's rules are valid for pooling Wald-type test statistics such as the z-statistic of Fisher's z-test (Rubin, 1987). For likelihood ratio statistics such as those of the G^2-test and the CG-test, appropriate rules have been proposed by Meng and Rubin (1992). See Appendix A for details on both sets of rules. Thus, for these and similar tests no new methodology is required for the pooling step. The rules guarantee that under the null hypothesis of conditional independence, the rejection rate is below the nominal α level. However, this assumes that an appropriate imputation model has been used. As discussed next, choosing the imputation models is more problematic in the context of causal discovery than in the regression context.

4.2.1 Choosing the imputation model

Rubin's rules (Rubin, 1987), as well as the rules by Meng and Rubin (1992), were derived within the joint model framework and assuming that the imputation model and the analysis model are *compatible*, meaning the models do not contradict each other (Meng, 1994; Bartlett et al., 2015). When using MICE, where imputation is based on a set of separate imputation models, a common joint distribution underlying all these models can exist only in special cases, e.g. when all imputation models are linear regression models or saturated logistic regression models (Hughes et al., 2014). In all other cases, the theoretical guarantees of the pooling rules do not apply, even though MICE has been found to be robust in many settings even in the absence of an underlying joint model (see Hughes et al., 2014, and the references therein).

In the context of causal discovery, two complications arise. One is that each conditional independence test assumes its own analysis model, and the different analysis models may

contradict each other. This is not the case for causal discovery using Fisher's z-test only: here we can impute using either a multivariate normal joint model, or MICE with linear main effects regression in order to ensure compatibility (Hughes et al., 2014). We do not recommend using predictive mean matching (Morris et al., 2014), as we found this method to lead to an increased type I error rate in several scenarios (results not shown). Similarly, for causal discovery using the G^2-test only, we can either use a multinomial joint model, or MICE with saturated (i.e. including all possible interactions) logistic regression (Hughes et al., 2014). In contrast, consider using the CG-test. The CG-distribution is not collapsible, i.e. if we assume a CG-distribution for all variables jointly, this does not imply that a given subset of the variables also follows a CG-distribution (Lauritzen and Wermuth, 1989; Lauritzen, 1990). The different analysis models thus contradict each other in general, and a compatible imputation model does not exist. Similarly, the analysis models will often contradict each other if a mix of different tests is used.

The second complication is that the number of variables in causal discovery analyses is often large, but imputation processes becomes instable when too many variables or model terms are involved (van Buuren, 2018; Hardt et al., 2012). Consider MICE using saturated logistic models: With 10 variables, the number of terms in each imputation model is $2^{10} = 1\,024$; with 100 variables, it equals $2^{100} > 10^{30}$. Some amount of model selection is necessary, but it is not clear what a good approach would be. For joint model imputation based on the multivariate normal distribution, it has been suggested to apply a ridge penalty (e.g. Schafer, 1997; Carpenter and Kenward, 2013), but this has not been generalised to other variable types. For discrete variables in particular, one could consider restricting the order of the interaction terms. Another idea is to use flexible imputation models, e.g. based on random forests (Doove et al., 2014; Shah et al., 2014).

4.2.2 Hybrid procedure

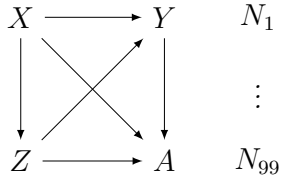
As an alternative to the above strategies for selecting the imputation models, we propose the following hybrid procedure. First, a preliminary graph skeleton is estimated using test-wise deletion, with a nominal α larger than the one to be used in the actual analysis. Tu et al. (2019) showed that under the assumptions of Proposition 2, the estimated graph will be a supergraph of the true skeleton (see Lemma 5). In a second step, MICE is performed such that the imputation model for variable V contains only V's neighbours and the neighbours of the neighbours. The rationale is that ideally, the imputation model for variable V would include all variables in the *Markov blanket* of the node V, which is defined as the set of V's parents, children and 'spouses', i.e. nodes with which V shares a common child. As the edges in the estimated preliminary skeleton are undirected, every neighbour of V is a potential child and every neighbour of a neighbour a potential 'spouse'.

4.3 Auxiliary information and noise – when multiple imputation is expected to outperform test-wise deletion

Both test-wise deletion and testing under multiple imputation yield type I error rates respecting the nominal significance level, under their respective assumptions. In addition, testing under multiple imputation has the potential to detect (conditional) associations with a higher power than test-wise deletion, as (i) no observations are deleted, and (ii) multiple imputation exploits information in the observed values about the missing values.

However, there are also situations in which multiple imputation is not expected to outperform test-wise deletion, e.g. when the incomplete variable is in the conditioning set of the conditional independence test, as illustrated in Scenario B of the following simulation experiment. Moreover, when the number of variables in the imputation model(s) is large, the imputation process could be dominated by noise and become unstable, as illustrated in Scenario D.

Illustration 3
Consider the following causal graph and covariance matrix, implying $X \not\perp\!\!\!\perp Y \mid Z$*:*



$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.2 & 0.2 & 0.5 & 0 & \dots & 0 \\ 0.2 & 1 & 0.2 & 0.2 & 0 & \dots & 0 \\ 0.2 & 0.2 & 1 & 0.2 & 0 & \dots & 0 \\ 0.5 & 0.2 & 0.2 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

We generated $n = 50$ *or* $n = 500$ *observations of* $(X, Y, Z, A, N_1, \dots, N_{99}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *and then deleted and imputed values as follows. Scenario A: 10, 30, 50 or 70 % of the values of* Z *were made MCAR. Imputation was based on a linear regression of* Z *on* (X, Y)*. Scenarios B,C,D: 10, 30, 50 or 70 % of the values of* X *were made MCAR. Scenario B: Imputation was based on a linear model of* X *on* (Y, Z)*. Scenario C: Imputation was based on a linear model of* X *on* (Y, Z, A)*. Scenario D: Imputation was based on a linear model of* X *on* $(Y, Z, A, N_1, \dots, N_{99})$*. The number of imputations was* $m = 100$*. In all scenarios,* $X \perp\!\!\!\perp Y \mid Z$ *was tested (i) using Fisher's z-test (*$\alpha = 0.05$*) with test-wise deletion and (ii) using Fisher's z-test (*$\alpha = 0.05$*) on the multiply imputed data. Figure 5 shows the rejection rate (power) over 10 000 replications.*

Figure 5 shows that in Scenario A, multiple imputation successfully exploited information in X and Y to partially recover the missing information about Z, resulting in a higher power for detecting $X \not\perp\!\!\!\perp Y \mid Z$. In contrast, multiple imputation did not result in a higher power in Scenario B, where missing values occurred in X. This is a phenomenon well-known in the context of regression analysis: When the analysis model is a model for $\mathrm{E}(X \mid Y, Z)$ and missingness occurs in X, then multiple imputation using the imputation model $\mathrm{E}(X \mid Y, Z)$ only adds noise to the analysis, hence restricting the analysis to the complete cases is preferred (Little and Rubin, 2002, page 237; Hughes et al., 2019). Testing $X \perp\!\!\!\perp Y \mid Z$ using Fisher's z-test is conceptually equivalent to modelling $\mathrm{E}(X \mid Y, Z)$ and testing for the coefficient of Y being zero. A different situation occurs when the imputation model includes additional variables not in the analysis model, such as the variable A in the simulation. In the context of regression analysis, such variables are called *auxiliary* to the variables in the analysis model. In Scenario C, where the imputation model for X included Y, Z and A, the multiple imputation procedure successfully exploited information in A, hence the power was (slightly) higher. In Scenario D, however, where the imputation model for X additionally included 99 noise variables, the noise outweighed the auxiliary information, hence the power was even lower than under test-wise deletion.

The above has consequences for causal discovery, where each variable can have different

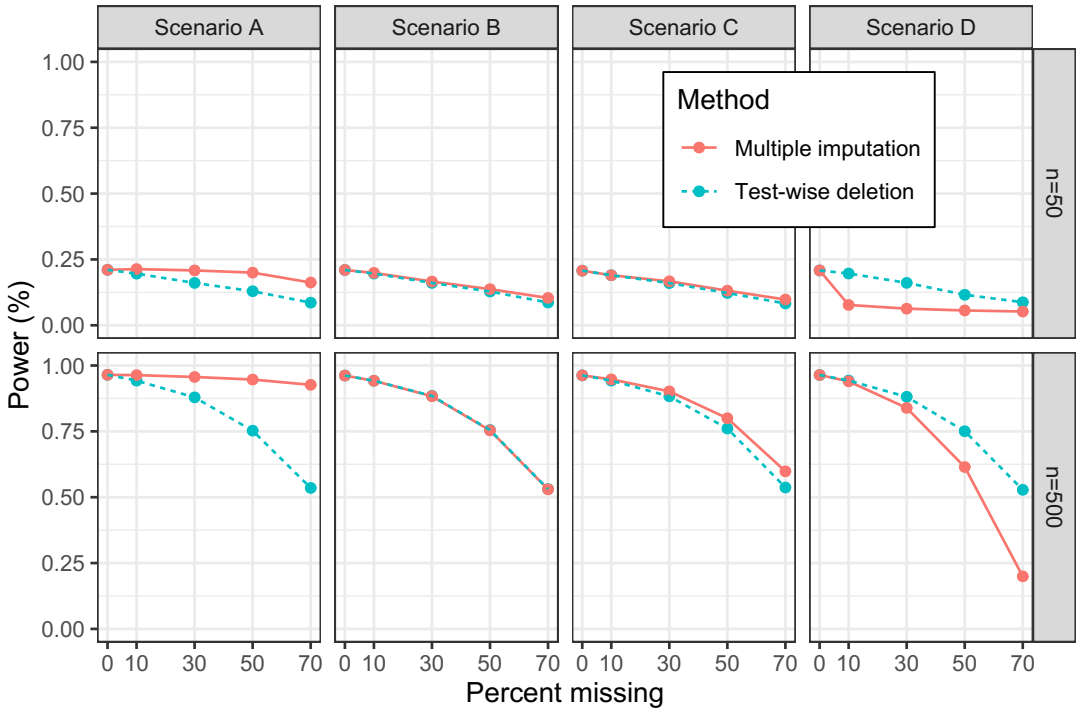


Figure 5: Power of Fisher's z-test combined with test-wise deletion versus multiple imputation. The null hypothesis is $X \perp\!\!\!\perp Y \mid Z$. **Scenario A**: Values in Z are MCAR. **Scenario B**: Values in X are MCAR. **Scenario C**: Values in X are MCAR, the imputation model includes an auxiliary variable. **Scenario D**: Values in X are MCAR, the imputation model includes an auxiliary variable and 99 noise variables.

roles (variable of interest, conditioning variable, auxiliary variable, noise variable) relative to the different tests that are performed. First, multiple imputation is expected to benefit from graphs with strong associations between the variables, as the observed values then contain more information about the missing ones. Second, multiple imputation is expected to benefit from dense graphs. This is because during the PC-algorithm, the conditional independence test $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ is only performed if X or Y still have more than $|\mathbf{Z}|$ neighbours. Hence, fewer conditional tests will be performed if the graph is sparse, but as argued above, multiple imputation is especially effective when missing values occur in the conditioning variable(s). These two trends become visible in Illustration 4 below. Third, as discussed in the previous section, variable selection on the imputation models is needed when the number of variables or terms in the imputation models is large, as otherwise the models are dominated by noise.

Illustration 4

Random DAGs with 8 nodes each were generated using the `randomDAG` *function from the* `R` *package* `pcalg` *(Kalisch et al., 2012). The edge density parameter (probability of connecting a newly added node to a node already in the graph) was set to 0.1 ('very sparse'), 0.25 ('sparse'), 0.4 ('medium'), 0.55 ('dense') or 0.7 ('very dense'). The graphs were parameterised as linear structural models, where the edge weights w were chosen such that the power of Fisher's z-test ($\alpha = 0.05$) for detecting the marginal dependence $X \not\perp\!\!\!\perp Y$*

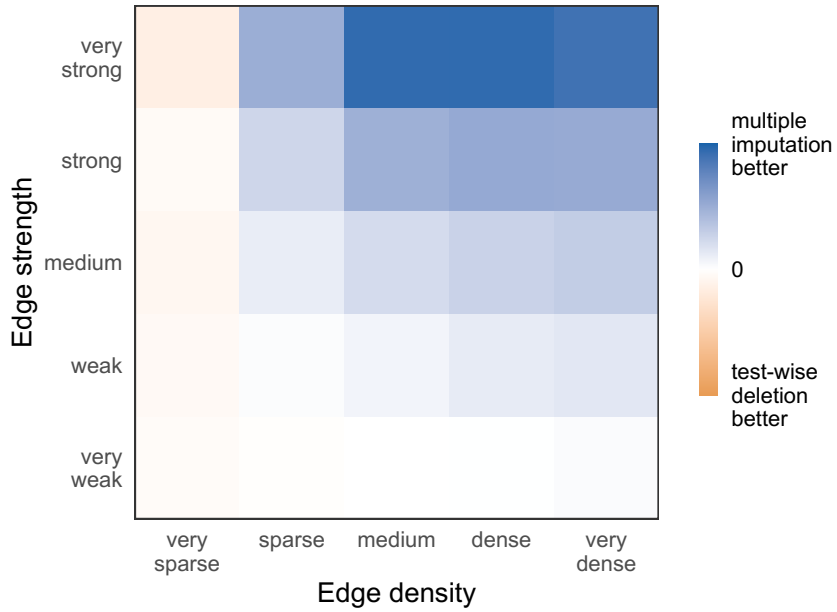


Figure 6: Relative performance of multiple imputation vs. test-wise deletion for discovering random graphs with 8 nodes and different edge densities and edge strengths. 10 % of data points were MCAR. The colour intensity is proportional to the relative Hamming distance (compared to using the full data); white means that the same relative Hamming distance was obtained for multiple imputation and test-wise deletion.

in the model $X \overset{w}{\to} Y$ *was 10 % (very weak), 30 % (weak), 50 % (medium), 70 % (strong) or 90 % (very strong). 500 observations were generated, and 10 % of the values were randomly deleted. Graphs were estimated by the PC-algorithm using Fisher's z-test* ($\alpha = 0.05$) *(i) with test-wise deletion, (ii) on multiply imputed data (based on linear regressions, 100 imputations) and (iii) on the full data (for comparison). Figure 6 shows the results averaged over 1 000 repetitions per scenario. The colour intensity is proportional to the difference in the relative Hamming distance, i.e.* $H_{MI}/H_{full} - H_{twd}/H_{full}$*, where* H_{MI}*,* H_{twd} *and* H_{full} *are the average Hamming distances (to the true graph) obtained using multiple imputation, test-wise deletion and the full data, respectively. The plot shows that under our simple data-generating model, multiple imputation yields better graph estimates in all but the extreme cases, and the advantage over test-wise deletion tends to be greatest in dense graphs with strong dependencies between variables.*

5 Detailed comparison on synthetic data

The aim of the simulation study was to compare the performance of test-wise deletion and multiple imputation with different imputation models, in order to help guide the choice between the different methods in practice. `R` code for replication can be found on GitHub (link will be provided after publication).

We considered 7 data-generating mechanisms (*ECOLI*, *MAGIC*, *ASIA*, *SACHS*, *HEALTHCARE*, *MEHRA*, *ECOLI_large*), 3 sample sizes (n=100, 1 000, 5 000) and 3 missingness mechanisms ('MCAR', 'MAR', 'MNAR'), yielding a total of 63 simulation scenarios.

Table 2: Data-generating mechanisms used in the simulation study. The footnotes indicate the variables of the selected subgraphs.

	#Variables			#Edges	#Categories
	total	Gaussian	discrete		
ECOLI[1]	12	12	–	17	
MAGIC[2]	7	7	–	7	
ASIA	8	–	8	8	2 each
SACHS	11	–	11	17	3 each
HEALTHCARE	7	4	3	9	2/3/3
MEHRA[3]	8	4	4	14	31/6/20/9
ECOLI_large	46	46	–	70	

[1] b1191, cchB, eutG, fixC, ibpB, sucA, tnaA, yceP, yfaD, ygbD, ygcE, yjbO

[2] MIL, G1217, G257, G2208, G1338, G524, G1945 of *MAGIC-NIAB*

[3] Zone, Type, Year, Region, co, pm10, pm2.5, so2

5.1 Synthetic incomplete datasets

Data were generated from benchmark causal graphs and their data-generating mechanisms according to the Bayesian Network Repository (`www.bnlearn.com/bnrepository`). Table 2 summarises their key features. The *ECOLI* graph is a subgraph of the *ECOLI_large* graph.

In the *ECOLI*, *MAGIC*, *ASIA*, *SACHS*, *HEALTHCARE* and *MEHRA* scenarios, missing values were generated as follows. For 'MCAR', 18 % of the values in the dataset were randomly chosen and deleted. Both multiple imputation and test-wise deletion are valid under this missingness mechanism. For 'MAR', one or two groups of three or four variables each were chosen at random. Using the `ampute` function from the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011), missing values were generated such that exactly one variable per group was missing in each data row, and the probability of missingness depended on the values of the other two or three variables in the group. Values in one other randomly chosen variable were randomly deleted until an overall missingness proportion of 18 % was reached. Under this 'MAR' mechanism, multiple imputation is valid, while test-wise deletion is not, as the admissible separator condition is not necessarily fulfilled for all pairs of variables. For 'MNAR', we chose one fixed 'key' variable and four to nine 'subordinate' variables per graph. In the data rows with the q % largest values of the 'key' variable, the values of the 'key' variable and all 'subordinate' variables were deleted, where q was chosen such that the overall missingness proportion was 18 %. Under this 'MNAR' mechanism, the admissible separator condition is satisfied for all pairs of variables, hence test-wise deletion is valid, while multiple imputation is not.

In the *ECOLI_large* scenarios, missing values were generated in the same variables and using the same mechanisms as in the *ECOLI* scenarios, leading to an overall missingness proportion of 4.7 % (instead of 18 %) in each scenario.

5.2 Missing data methods

The PC-stable algorithm as implemented in `pcalg` (Kalisch et al., 2012; Colombo and Maathuis, 2014) was applied, using the following methods for dealing with the missing values: 1) List-wise deletion, i.e. data rows with missing observations were deleted before applying PC-stable. 2) Test-wise deletion using `gaussCItwd`, `disCItwd` or `mixCItwd` from the `micd` package. 3-4) Test-wise deletion with the (3) density or (4) permutation correction method by Tu et al. (2019) as implemented in the `MVPC` repository (`www.github.com/TURuibo/MVPC`; only available for continuous or binary data). 5-9) Conditional independence testing under multiple imputation using `gaussMItest`, `disMItest` or `mixMItest` from the `micd` package, where the imputations were generated using the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011) with different imputation models, as follows: (5) each variable was imputed based on the variables in its Markov blanket (i.e. its parents, children and nodes with which it shares a common child) using linear or logistic regression imputation including all interaction terms ('oracle' multiple imputation; this is not possible to do in practice as the graph is not known, but is included here as a reference); (6) linear or logistic regression imputation including all interaction terms; (7) main effects linear or logistic regression imputation; (8) random forest imputation using the `rf` option (Doove et al., 2014); (9) random forest imputation using the `rfcont` or `rfcat` option from `CALIBERrfimpute` (Shah et al., 2014). 10) Missing values were singly imputed with the column mean (continuous data) or mode (discrete data) before applying PC-stable. For multiple imputation, we choose $m = 10$ imputations. Although this number is smaller than what is recommended in the literature (van Buuren, 2018; Carpenter and Kenward, 2013), we found in preliminary simulations (not shown) that the test rejection rates do not change considerably when more imputations are added. We still recommend using $m = 100$ or higher in real applications. For methods (5) and (6), the highest order of interaction was set to 2 or 3 if required to reduce the runtime. For random forest imputation, we set the number of trees to 100, as we found this to improve the quality of the estimated graphs, compared to the default of 10 trees, in preliminary simulations (not shown).

In the *ECOLI_large* scenarios, we additionally included three versions (A, B, C) of the hybrid procedure proposed in Section 4.2.2, where in step 1, the preliminary graph skeleton was estimated using `alpha=0.2`. In versions B and C, the skeleton search was stopped after all marginal independence tests had been performed, as the higher-order tests are expected to be less reliable. Additionally, in version C, the neighbours of the neighbours were ignored, in order to obtain even sparser imputation models.

5.3 Evaluation criteria

The performance was evaluated using the following metrics: number of edges in the estimated graph; proportion of discovered edges among the edges in the true CPDAG, ignoring edge orientation (recall); proportion of correctly discovered edges among the discovered edges, ignoring edge orientation (precision); number of edge insertion or deletions in order to transform the estimated graph into the true CPDAG, ignoring edge orientation (Hamming distance); and number of edge insertions, deletions or reversals in order to transform the estimated graph into the true CPDAG (structural Hamming distance; Tsamardinos et al., 2006).

5.4 Results

The runtime was about three weeks on a 240-node high-performing computer cluster. Figure 7 provides a first overview of the results. It compares the performance of test-wise deletion without correction vs. multiple imputation based on linear models (Gaussian variables), logistic models including interaction terms (discrete variables) or the CALIBER random forest method (mixed variables). The horizontal position of the points in the figure is determined by the difference in the relative Hamming distance as defined in Illustration 4.

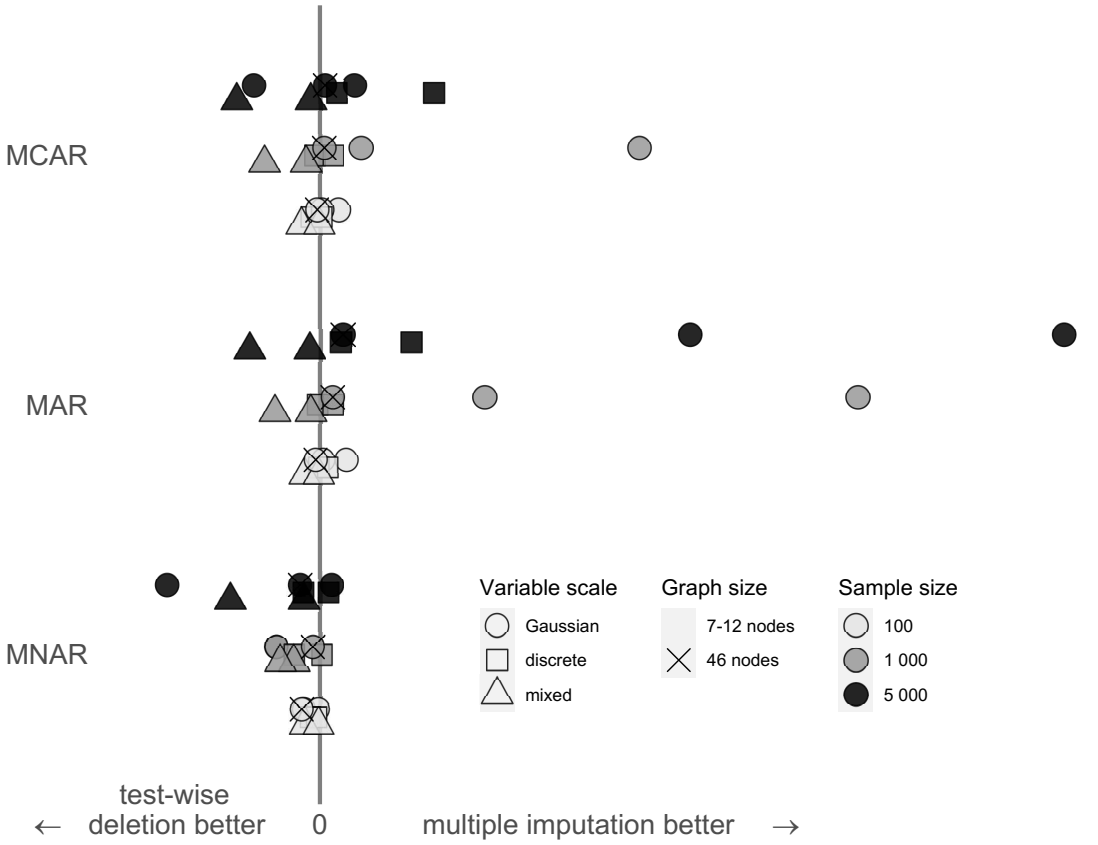


Figure 7: Overview over the simulation results. Shown on the x-axis is the difference in the relative Hamming distance (compared to using the full data).

The following trends are apparent: For $n = 100$, there were virtually no differences in the performance of test-wise deletion vs. multiple imputation, and the differences were most pronounced for $n = 5\,000$. For Gaussian variables, multiple imputation outperformed test-wise deletion in almost all 'MCAR' and 'MAR' scenarios, but not in the 'MNAR' scenarios. For discrete variables, the same trend can be observed, but the advantage of multiple imputation only occurred for $n = 5\,000$. For mixed variables, test-wise deletion outperformed multiple imputation in all simulation scenarios.

More detailed results are shown in Figures 8 and 9, and in the tables in the Online Supplement. The main observations are as follows: First, the edge recall was generally largest in the scenarios with Gaussian variables, smaller for discrete variables and very

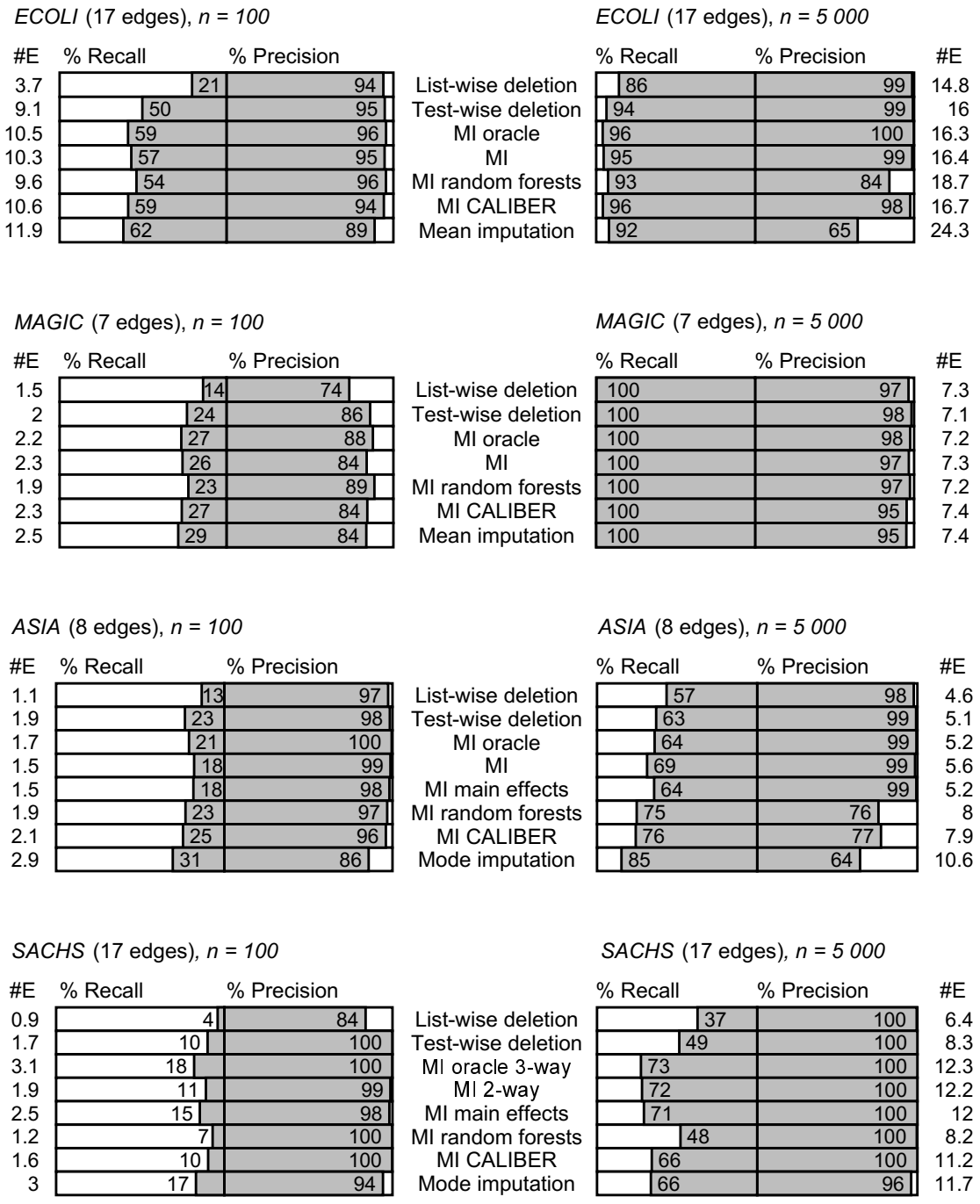


Figure 8: Simulation results, part I (*ECOLI*, *MAGIC*, *ASIA*, *SACHS*). Shown are the average edge recall (% Recall), the average edge precision (% Precision) and the average number of edges (#E) in 1 000 graphs estimated using the PC-algorithm combined with different methods for handling missing values. The sample size was either $n = 100$ (left) or $n = 5\,000$ (right) and missing values were generated using the 'MCAR' mechanism described in the text. MI=multiple imputation.

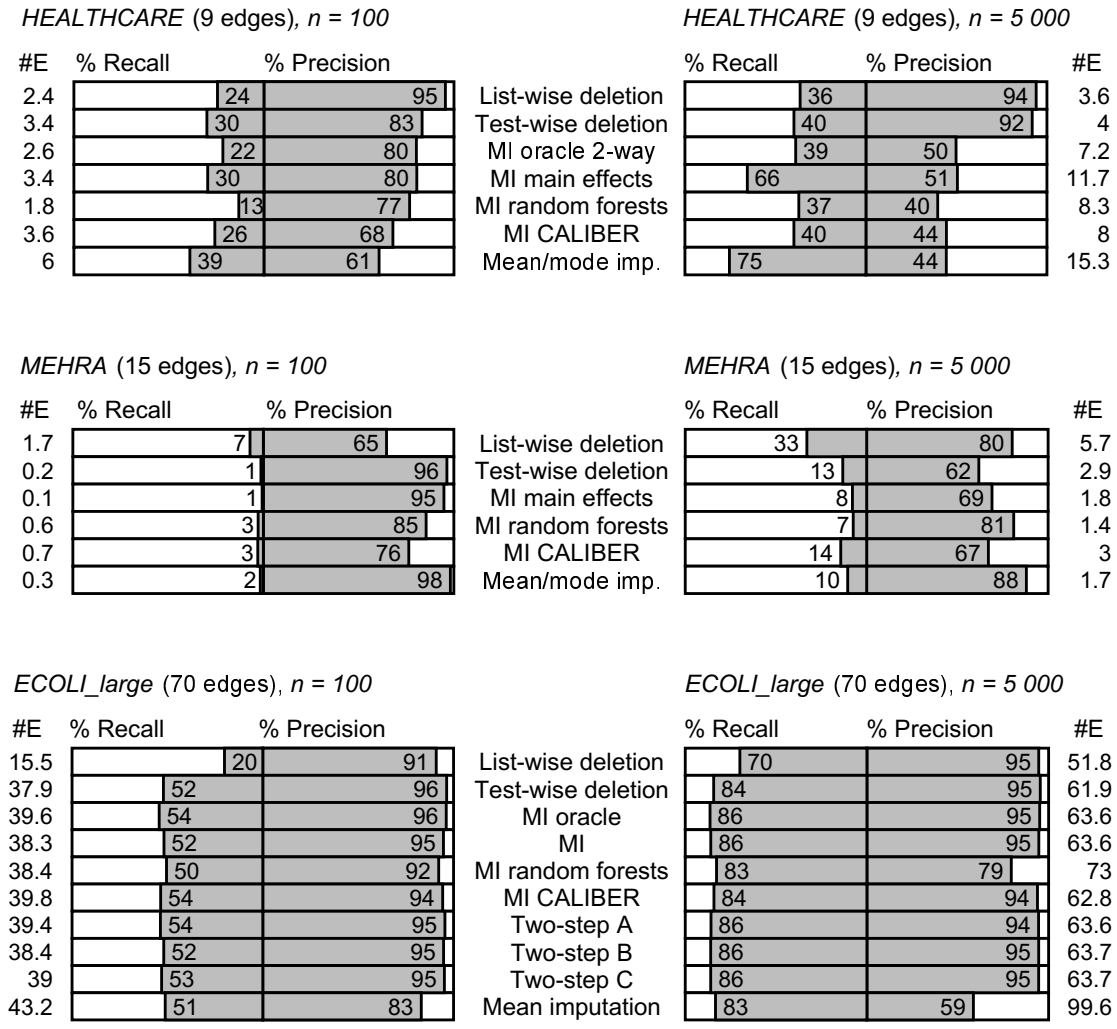


Figure 9: Simulation results, part II (*HEALTHCARE*, *MEHRA*, *ECOLI_large*). Shown are the average edge recall (% Recall), the average edge precision (% Precision) and the average number of edges (#E) in 1 000 graphs estimated using the PC-algorithm combined with different methods for handling missing values. The sample size was either $n = 100$ (left) or $n = 5\,000$ (right) and missing values were generated using the 'MCAR' mechanism described in the text. MI=multiple imputation.

small for mixed variables. Further, while an average precision of more than 95 % was attained by a subset of the missing data methods in all Gaussian and discrete scenarios, this was not the case for the mixed scenarios. This is in line with earlier results using data without missing values (Andrews et al., 2018) and indicates that causal discovery using mixed data is a particularly challenging task.

As expected, list-wise deletion resulted in sparse graphs with large (structural) Hamming distances, due to the low power. An exception occurred in the *MEHRA* 'MCAR' scenarios with $n = 100$ and $m = 5\,000$, where list-wise deletion resulted in denser graphs than test-wise deletion. This seemingly paradoxical behaviour can be explained as follows. The PC-algorithm starts with marginal tests and proceeds to conditional testing only if the nodes still have enough neighbours to be included in the conditioning set. Under list-wise deletion, only few edges remain after the marginal phase, hence the number of conditional tests performed is small. Under test-wise deletion, more edges survive the marginal phase, hence more conditional tests are performed, but this leads to the deletion of most remaining edges due to the very low power of the CG-test conditioning on the categorical *MEHRA* variables with 6–31 categories.

Single imputation by the column mean or mode led to graphs that were 'too large' (many edges but low precision). In order to understand why this happened, consider the structure $X \to Y \to Z$, implying $X \perp\!\!\!\perp Z \mid Y$ but $X \not\perp\!\!\!\perp Z$. If Y contains missing values and these are replaced by the column mean or mode, it is very likely that after conditioning on the imputed version of Y, there remains a residual association between X and Z. However, the sample size is as large as if the data had been complete to begin with, and the fact that values were imputed is not taken into account by the testing procedure. Hence, the null hypothesis is rejected with a probability larger than the nominal test level, so that the resulting graph is more likely to contain an edge between X and Z.

Test-wise deletion performed well overall. The results using the correction methods proposed by Tu et al. (2019) are not shown in Figures 8 and 9, as they were very similar to those using test-wise deletion without correction. This is not surprising, as the missingness mechanisms chosen for the simulation did not require these corrections. Tu et al. (2019) and Tu et al. (2020) showed that if they are required (which is usually not known in real data analyses), using the corrections improves the performance; we conclude that if they are not required, the performance is at least not substantially worsened.

Concerning multiple imputation, we observed that parametric imputation using interaction terms was computationally infeasible (producing errors) in many repetitions, especially for the datasets with mixed variables. See the Online Supplement for more information. We obtained inconclusive results for the usefulness of the two variants of random forest imputation. In the scenarios with only Gaussian or only categorical variables, the `rf` variant often produced graphs with a lower precision and worse Hamming distance than parametric imputation, and the CALIBER variant often ranged between parametric and `rf` imputation in terms of different evaluation metrics. In the scenarios with mixed variables, the performance of the two random forest methods was usually similar and also comparable to that of parametric imputation. The main difference between the two random forest options lies in how they guarantee that the multiply imputed values are sufficiently different from each other (in order to properly account for the uncertainty in the missing values). Using the `rf` option, a specified number of trees is fitted and one tree

is chosen at random. A prediction is made using this tree and the imputed value is randomly drawn from among the observed values that are in the same leaf as the prediction. The `CALIBERrfimpute` functions fit the random forest model on a bootstrap sample of the observed data. The imputed value is then either the best prediction plus a normal error (continuous case) or the predicted value from just one of the trees (discrete case). Based on our simulation results, we conclude that the CALIBER version is more appropriate for conditional independence testing, and we conjecture that the difference between the two versions is less pronounced when the goal is e.g. estimation of a regression coefficient.

The *ECOLI_large* results in Figure 9 demonstrate the potential of the hybrid method. For $n = 100$, test-wise deletion outperformed parametric multiple imputation ('MI' in the figure) in terms of the average precision and Hamming distance (34.9 for test-wise deletion vs. 35.8 for multiple imputation; see Online Supplement). Using the hybrid method A, the recall was as good as when using oracle multiple imputation, and the average Hamming distance was only 34.2. CALIBER random forest imputation also performed well and yielded an average Hamming distance of 34.8. We expect the differences to be larger in scenarios with even more variables.

6 Data application

To investigate the causal structure underlying the IDEFICS data, we used the `tpc` function from the `tpc` package (`www.github.com/bips-hb/tpc`), which is based on `pcalg` (Kalisch et al., 2012), but offers additional options for integrating background knowledge (see Andrews et al., 2021). We specified the following partial node ordering: (*income*, *isced*, *bage*, *migrant*, *sex*) < *smoke* < *week* < *bweight* < (*formula*, *hdiet*, *bf*) < (*age*, *school*) < (*bmi_m*, *fmeal*, *yhei*) < (*media*, *mvpa*, *sed*, *sleep*, *bmi*, *homa*, *wb*). In addition, we specified that *sex* and *age* are exogenous, i.e. do not have parent nodes. After obtaining rather sparse graphs in a test run, we set `alpha=0.1`. Missing data were dealt with using the following methods (in parentheses: name of the conditional independence test function used): list-wise deletion (`mixCItest`); test-wise deletion (`mixCItwd`); parametric multiple imputation based on main effects linear or logistic regression (`mixMItest`); random forest multiple imputation using `rf` in `mice` (`mixMItest`); random forest multiple imputation using `rf_cont` or `rf_cat` from the `CALIBERrfimpute` package (`mixMItest`); single imputation by the column mean or mode (`mixCItest`). For multiple imputation, we used 100 imputations and 100 trees where applicable. In order to get an impression of the variability of the estimated graphs, the whole analysis was repeated 50 times on bootstrap samples of the original data (Pigeot et al., 2015).

The graphs estimated in the main analysis are shown in Figure 10. All discovered graphs were sparser than what might be expected based on expert knowledge (Vandenbroeck et al., 2017). Possible reasons could be the small sample size, violations of the faithfulness assumption, or deviations from the CG assumption. Consequently, the absence of edges should be interpreted with care.

Table 3 compares the total number of edges and the number of edges adjoining nodes of 'critical' variables with more than 20 % missing values, i.e. *mvpa*, *sed*, *sleep* and *homa*, in the main analysis and the bootstrap analyses. The numbers reveal, first of all, that the variability among the bootstrap samples was rather large, which again might be explained

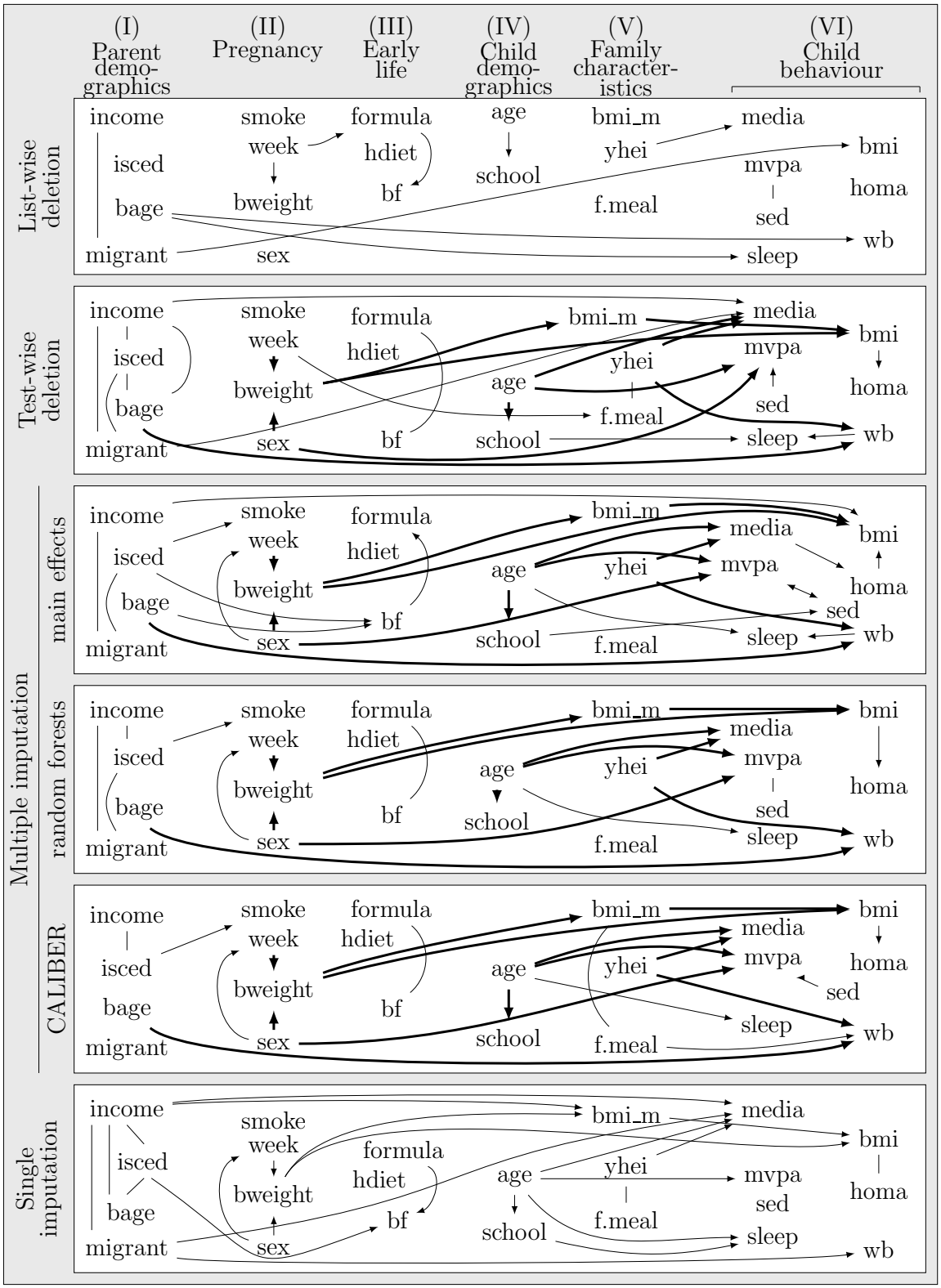


Figure 10: Estimated IDEFICS graphs. Bi-directed edges indicate that the direction could not be determined due to conflicting information in the data. Bold edges are present in all graphs estimated under test-wise deletion or multiple imputation.

Table 3: Number of edges in the graphs estimated from the IDEFICS data (in parentheses: minimum, average and maximum number of edges in the bootstrap analyses). The critical edges are those adjoining nodes *mvpa*, *sed*, *sleep* or *homa*, for which large proportions of values were missing.

	Total number of edges (min, average, max)	Number of critical edges (min, average, max)
List-wise deletion (lwd)	10 (6, 15.9, 25)	2 (1, 4.4, 8)
Test-wise deletion (twd)	26 (20, 27.3, 34)	6 (2, 4.9, 8)
Multiple imputation		
– main effects (MI)	26 (18, 25.3, 31)	7 (3, 4.9, 7)
– random forests (rfMI)	21 (22, 30.4, 37)	5 (4, 8.2, 13)
– CALIBER (cMI)	21 (20, 28.7, 37)	5 (4, 7.5, 11)
Single imputation (sing)	24 (25, 32.7, 41)	4 (4, 7.0, 13)

by the relatively small sample size. For random forest and single imputation, the number of edges obtained in the main analysis was smaller than the minimum number of edges obtained using the bootstrap samples. This is a known phenomenon and a correction has been proposed for score-based causal discovery (Steck and Jaakkola, 2003), but we are not aware of a correction method for the PC-algorithm. In line with the simulation results, list-wise deletion led to the sparsest graphs, while the densest graphs in the bootstrap analysis were discovered using single imputation. The multiple imputation methods tended to discover more edges adjoining 'critical' nodes than test-wide deletion. This might be because the sample size available for tests containing the 'critical' variables, where many values are missing, is rather small under test-wise deletion.

In the Online Supplement, we include diagnostic plots for the multiple imputation procedures in the main analysis. Based on visual inspection of the convergence plots, the algorithm converged in all three cases. The random forest (`rf`) method was most successful in generating imputed values with a distribution matching that of the observed values. Figure 11 illustrates this for the *wb* (well-being) variable. The distributions of the imputed values generated by the parametric and CALIBER methods are more symmetric. This may indicate that the (`rf`) method is better able to predict the missing values. However, as discussed previously (Shah et al., 2014), and as also witnessed in the simulation study, this does not necessarily mean that the graphs estimated using random forest (`rf`) imputation are closer to the truth.

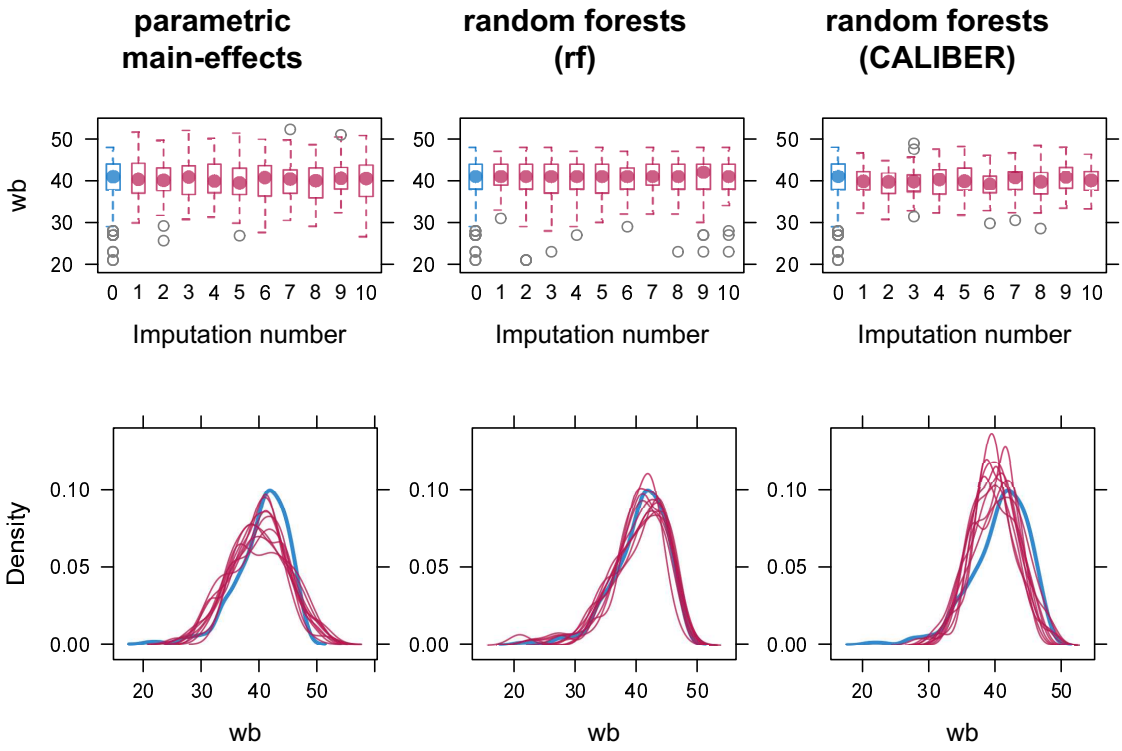


Figure 11: Diagnostic plots for variable *wb* (well-being). Shown are the distribution of the observed values (blue boxplot and curves) and the distribution of the values generated by 10 randomly chosen imputations (red boxplots and curves), for the three different imputation methods.

7 Conclusions

In this paper, we investigated test-wise deletion and multiple imputation for dealing with missing values in constraint-based causal discovery. Test-wise deletion relies on faithful observability and the admissible separator condition, whereas multiple imputation requires the missing values to be MAR. Both assumptions are implied by the stronger MCAR but are otherwise difficult to justify in practice.

In our empirical comparisons, we confirmed that test-wise deletion and multiple imputation are clearly superior to list-wise deletion and single imputation. We also demonstrated that while multiple imputation outperforms test-wise deletion in settings with small graphs and Gaussian variables, there is no overall best approach in realistically complex settings with a larger number of variables especially when these are a mix of continuous and discrete measurements. Random forest imputation or the hybrid method we proposed might be useful especially in settings with 50 or more variables, but comprehensive comparisons are difficult due to the long runtime of all three methods involved (causal discovery, multiple imputation and random forests).

An alternative missing value method for causal discovery not considered in this paper is in-

verse probability weighting (Gain and Shpitser, 2018). Likelihood-based approaches such as Expectation Maximisation can be used with score-based causal discovery (Friedman, 1997; Scutari, 2020) but are not straightforward to combine with constraint-based algorithms (see Sokolova et al., 2017, for a first idea assuming a joint nonparanormal distribution).

Future research should address model selection of the imputation models in MICE. This is relevant also outside the area of causal discovery, but the literature on this topic is surprisingly scarce (Noghrehchi et al., 2021). Finally, reliably learning (causal) graphs from data with mixed measurement scales remains a challenge especially with the additional complication of missing values.

Acknowledgements

We gratefully acknowledge financial support by the German Research Foundation (DFG—Project DI 2372/1-1).

A Conditional independence testing

In this appendix, we provide details about the three conditional independence tests we focussed on in this work. We first review how each test is implemented when complete data are available, and then describe how they can be applied to multiple imputed data.

Fisher's z-test

Consider a random vector $(X, Y, Z_1 \ldots, Z_s)^T \in \mathbb{R}^{s+2}$ with covariance matrix $\mathbf{\Sigma}$. The *partial correlation* between X and Y given $\mathbf{Z} = (Z_1 \ldots, Z_s)$ is defined as

$$\rho_{XY.\mathbf{Z}} = \frac{p_{12}}{\sqrt{p_{11}}\sqrt{p_{22}}},$$

where p_{ij} is the (i, j)-the element of the precision matrix $\mathbf{P} = \mathbf{\Sigma}^{-1}$. The corresponding *empirical partial correlation* can be estimated from n observations of $(X, Y, \mathbf{Z})^T$ as

$$\hat{\rho}_{XY.\mathbf{Z}} = \frac{\hat{\boldsymbol{\varepsilon}}_X^T \hat{\boldsymbol{\varepsilon}}_Y}{\sqrt{\hat{\boldsymbol{\varepsilon}}_X^T \hat{\boldsymbol{\varepsilon}}_X}\sqrt{\hat{\boldsymbol{\varepsilon}}_Y^T \hat{\boldsymbol{\varepsilon}}_Y}},$$

where $\hat{\boldsymbol{\varepsilon}}_X$ is the vector of residuals after regressing X on $\mathbf{Z}$, and $\hat{\boldsymbol{\varepsilon}}_Y$ is the vector of residuals after regressing Y on $\mathbf{Z}$.

For Fisher's z-test (Fisher, 1924), it is assumed that $(X, Y, \mathbf{Z})^T$ follows a multivariate normal distribution. Then $\rho_{XY.\mathbf{Z}} = 0$ if and only if $X \perp\!\!\!\perp Y \mid \mathbf{Z}$; this is the null hypothesis of Fisher's z-test. The test statistic is

$$z(\hat{\rho}_{XY.\mathbf{Z}}) = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_{XY.\mathbf{Z}}}{1 - \hat{\rho}_{XY.\mathbf{Z}}} \right). \tag{1}$$

Under the multivariate normal assumption, $z(\hat{\rho}_{XY.\mathbf{Z}})$ is asymptotically normal with variance $1/(n - s - 3)$, and mean zero under the null hypothesis.

Fisher's z-test under multiple imputation

Fisher's z-test can be applied to multiply imputed data using Rubin's rules, as follows (Schafer, 1997, page 109; Foraita et al., 2020).

Consider M completed datasets obtained by multiple imputation, and let $z^{(m)}(\hat{\rho}_{XY.\mathbf{Z}})$ be the z-statistic calculated according to Equation (1) from the m-th imputed dataset, $m = 1, \ldots, M$. The pooled test statistic is

$$\bar{z}(\hat{\rho}_{XY.\mathbf{Z}}) = \frac{1}{M} \sum_{m=1}^{M} z^{(m)}(\hat{\rho}_{XY.\mathbf{Z}}).$$

The variance of $\bar{z}(\hat{\rho}_{XY.\mathbf{Z}})$ is estimated as

$$T_{XY.\mathbf{Z}} = \overline{W}_{XY.\mathbf{Z}} + \left(1 + \frac{1}{M}\right) B_{XY.\mathbf{Z}}, \tag{2}$$

which has two components: $\overline{W}_{XY.\mathbf{Z}}$ is the average *within-imputation variance* and is calculated as

$$\overline{W}_{XY.\mathbf{Z}} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{n-s-3} = \frac{1}{n-s-3}.$$

The extra variance due to the missing values is captured in the *between-imputation variance*

$$B_{XY.\mathbf{Z}} = \frac{1}{M-1} \sum_{m=1}^{M} \left[z^{(m)}(\hat{\rho}_{XY.\mathbf{Z}}) - \bar{z}(\hat{\rho}_{XY.\mathbf{Z}})\right]^2.$$

The term $\left(1 + \frac{1}{M}\right)$ in Equation (2) adjusts for the fact that only a finite number M of imputations was drawn.

Under the null hypothesis $\rho_{XY.\mathbf{Z}} = 0$, $\bar{z}(\hat{\rho}_{XY.\mathbf{Z}})/\sqrt{T_{XY.\mathbf{Z}}}$ approximately follows a Student's t-distribution with degrees of freedom given by

$$\nu = (M-1)\left[1 + \frac{\overline{W}_{XY.\mathbf{Z}}}{(1+M^{-1})B_{XY.\mathbf{Z}}}\right]^2.$$

The G^2-test

Consider a vector $(X, Y, Z_1, \ldots, Z_s)^T$ of categorical random variables, and define $\mathbf{Z} = (Z_1, \ldots, Z_s)$. The sets of values that X, Y and $\mathbf{Z}$ can take are denoted by $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively. The vector $(X, Y, \mathbf{Z})^T$ thus defines a 3-way contingency table. Denote by $\theta_{xy\mathbf{z}}$ the probability of observing $(x, y, \mathbf{z})^T$, for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\mathbf{z} \in \mathcal{Z}$. This corresponds to one cell in the contingency table. Further, denote the marginal probabilities with respect to X and Y, respectively, as $\theta_{+y\mathbf{z}} = \sum_{x\in\mathcal{X}} \theta_{xy\mathbf{z}}$ for $y \in \mathcal{Y}$, $\mathbf{z} \in \mathcal{Z}$, and $\theta_{x+\mathbf{z}} = \sum_{y\in\mathcal{Y}} \theta_{xy\mathbf{z}}$ for $x \in \mathcal{X}$, $\mathbf{z} \in \mathcal{Z}$.

Without further assumptions, drawing n independent observations of $(X, Y, \mathbf{Z})^T$ can be viewed as sampling from a multinomial distribution with parameters n and

$$\boldsymbol{\theta} = \{\theta_{xy\mathbf{z}} : x \in \mathcal{X}, y \in \mathcal{Y}, \mathbf{z} \in \mathcal{Z}\}.$$

We refer to this as the saturated multinomial model. The number of elements of $\boldsymbol{\theta}$ is equal to $|\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}|$. As the elements must sum to 1, this corresponds to $d = |\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}| - 1$ degrees of freedom.

If $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, which is the null hypothesis of the G^2-test, then fewer parameters are required to describe the distribution of $(X, Y, \mathbf{Z})^T$. In particular, under $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ we have that for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $\mathbf{z} \in \mathcal{Z}$, $\theta_{xy\mathbf{z}} = \theta_{+y\mathbf{z}} \cdot \theta_{x+\mathbf{z}}$. Thus, under the null hypothesis the set of parameters can be reduced to

$$\boldsymbol{\theta}^0 = \{\theta_{+y\mathbf{z}} : y \in \mathcal{Y}, \mathbf{z} \in \mathcal{Z}\} \cup \{\theta_{x+\mathbf{z}} : x \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}\},$$

which has $|\mathcal{X}| \cdot |\mathcal{Z}| + |\mathcal{Y}| \cdot |\mathcal{Z}|$ elements. As $\sum_{y \in \mathcal{Y}} \sum_{\mathbf{z} \in \mathcal{Z}} \theta_{+y\mathbf{z}} = \sum_{x \in \mathcal{X}} \sum_{\mathbf{z} \in \mathcal{Z}} \theta_{x+\mathbf{z}} = 1$, this corresponds to $d^0 = (|\mathcal{X}| - 1) \cdot |\mathcal{Z}| + (|\mathcal{Y}| - 1) \cdot |\mathcal{Z}|$ degrees of freedom.

The G^2-test is a likelihood ratio test with test statistic

$$G^2 = -2 \left[l\left(\hat{\boldsymbol{\theta}}^0\right) - l\left(\hat{\boldsymbol{\theta}}\right) \right],$$

where $l(\cdot)$ denotes the log-likelihood and the parameter estimates in $\hat{\boldsymbol{\theta}}^0$ and $\hat{\boldsymbol{\theta}}$ are obtained from the sample by counting the number of observations in the corresponding cell or margin and dividing by n. Asymptotically and under the null hypothesis, G^2 follows a χ^2-distribution with $d - d^0 = (|\mathcal{X}| - 1) \cdot (|\mathcal{Y}| - 1) \cdot |\mathcal{Z}|$ degrees of freedom.

The CG-test

The CG-distribution is defined as follows: Consider a set of variables $\mathbf{V}$ partitioned into continuous variables $\mathbf{C}$ and discrete variables $\mathbf{B}$, where $\mathbf{B}$ can take values in $\mathcal{B}$. Then $\mathbf{V}$ is said to follow a CG-distribution if for every $\mathbf{b} \in \mathcal{B}$, the conditional distribution of $\mathbf{C}$ given $\mathbf{B} = \mathbf{b}$ is multivariate normal with mean vector $\boldsymbol{\mu}_\mathbf{b}$ and covariance matrix $\boldsymbol{\Sigma}_\mathbf{b}$ (Lauritzen and Wermuth, 1989; Lauritzen, 1990). Note that $\boldsymbol{\Sigma}_\mathbf{b}$ is allowed to depend on $\mathbf{b}$, which is in contrast to the *general location model* sometimes considered in the context of multiple imputation (Schafer, 1997, p. 335). We denote the set of parameters describing the distribution of $\mathbf{V}$ as $\boldsymbol{\psi}_\mathbf{V} = \{p_\mathbf{b}, \boldsymbol{\mu}_\mathbf{b}, \boldsymbol{\Sigma}_\mathbf{b} : \mathbf{b} \in \mathcal{B}\}$, where $p_\mathbf{b} = \mathrm{P}(\mathbf{B} = \mathbf{b})$. The family of CG-distributions is not closed under marginalisation, i.e. if $\mathbf{V}$ follows a CG-distribution, then a subset $\mathbf{V}' \subset \mathbf{V}$ does not in general follow a CG-distribution (Lauritzen, 1990, Section 6.1.1).

A likelihood ratio test for conditional independence between CG-distributed variables was proposed by Andrews et al. (2018). We call this the CG-test. Consider a random vector $(X, Y, Z_1, \ldots, Z_s)^T$ following a CG-distribution with parameter vector $\boldsymbol{\psi}_{XY\mathbf{Z}}$, where $\mathbf{Z} = (Z_1, \ldots, Z_s)$. For the CG-test, it is assumed that the marginal distributions of $(X, \mathbf{Z})$, $(Y, \mathbf{Z})$ and $\mathbf{Z}$ are well approximated by CG-distributions with parameters $\boldsymbol{\psi}_{X\mathbf{Z}}$, $\boldsymbol{\psi}_{Y\mathbf{Z}}$ and $\boldsymbol{\psi}_\mathbf{Z}$, respectively. As noted above, this does not in general follow from the assumption that $(X, Y, Z_1, \ldots, Z_s)^T$ is CG.

Denote by $\hat{\boldsymbol{\psi}}_{XY\mathbf{Z}}$, $\hat{\boldsymbol{\psi}}_{X\mathbf{Z}}$, $\hat{\boldsymbol{\psi}}_{Y\mathbf{Z}}$ and $\hat{\boldsymbol{\psi}}_\mathbf{Z}$ the maximum likelihood estimates of $\boldsymbol{\psi}_{XY\mathbf{Z}}$, $\boldsymbol{\psi}_{X\mathbf{Z}}$, $\boldsymbol{\psi}_{Y\mathbf{Z}}$ and $\boldsymbol{\psi}_\mathbf{Z}$, respectively, obtained from data, with corresponding log likelihoods $l(\hat{\boldsymbol{\psi}}_{XY\mathbf{Z}})$, $l(\hat{\boldsymbol{\psi}}_{X\mathbf{Z}})$, $l(\hat{\boldsymbol{\psi}}_{Y\mathbf{Z}})$ and $l(\hat{\boldsymbol{\psi}}_\mathbf{Z})$. The CG-test compares the log likelihood $L = l(\boldsymbol{\psi}_{XY\mathbf{Z}})/l(\boldsymbol{\psi}_{Y\mathbf{Z}})$ for modelling X given Y and $\mathbf{Z}$ with the log likelihood $L^0 =$

$l(\boldsymbol{\psi}_{X\mathbf{Z}})/l(\boldsymbol{\psi}_{\mathbf{Z}})$ for modelling X given Y only, which corresponds to the null hypothesis that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, or equivalently, $f(x \mid y, \mathbf{z}) = f(x \mid \mathbf{z})$.

The test statistic of the CG-test is

$$\chi^2 = -2(L^0 - L).$$

Under the null hypothesis, χ^2 approximately follows a χ^2-distribution. The degrees of freedom vary depending on which variables are continuous and which are discrete; for details see Andrews et al. (2018)[2].

G^2-test and CG-test under multiple imputation

Rules for combining likelihood ratio statistics have been suggested by Meng and Rubin (1992). Consider M completed datasets obtained by multiple imputation. Let $\boldsymbol{\phi}$ and $\boldsymbol{\phi}^0$ be sets of parameters characterising the full and reduced model of interest. For the G^2-test, $\boldsymbol{\phi} = \boldsymbol{\theta}$ and $\boldsymbol{\phi}^0 = \boldsymbol{\theta}^0$; for the CG-test, $\boldsymbol{\phi} = \boldsymbol{\psi}$ and $\boldsymbol{\phi}^0 = \boldsymbol{\psi}^0$. As before, we use the superscript $^{(m)}$ to indicate estimators obtained from the m-th completed dataset. We denote by $l_m(\cdot)$ the log likelihood function given the m-th completed dataset.

First, the average likelihood ratio statistic is calculated as

$$\bar{LR} = \frac{1}{M}\sum_{m=1}^{M} -2[l_m(\hat{\boldsymbol{\phi}}^{0(m)}) - l_m(\hat{\boldsymbol{\phi}}^{(m)})],$$

and the average parameter estimates as

$$\bar{\boldsymbol{\phi}}^0 = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\phi}^{0(m)}$$

and

$$\bar{\boldsymbol{\phi}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\phi}^{(m)}.$$

The log likelihoods are then re-evaluated given each of the M completed datasets, with the parameters fixed to the average parameter estimates, and the corresponding likelihood ratio statistics are averaged:

$$\tilde{LR} = \frac{1}{M}\sum_{m=1}^{M} -2[l_m(\bar{\boldsymbol{\phi}}^0) - l_m(\bar{\boldsymbol{\phi}})].$$

The pooled test statistic is

$$D_3 = \frac{\tilde{LR}}{k(1 + r_3)}$$

with $r_3 = (M + 1)(\bar{LR} - \tilde{LR})/[k(M - 1)]$, where k equals the degrees of freedom that would have been used had complete data been available. The test statistic D_3 can be

[2]Note that equation (11) of Andrews et al. (2018) should read $df_p(\hat{\theta}_p) = d(d+1)/2 + 1 + \mathbf{d}$, in order to account for the estimated vector of means (Bryan Andrews, personal communication).

approximated by an F-distribution with k and $4+[k(M-1)-4][1+(1-2k^{-1}(M-1)^{-1})/r_3]^2$ degrees of freedom. The name 'D_3' has no specific meaning; it is used in several popular books to distinguish it from the so-called D_1 statistic for multi-parameter Wald tests and the so-called D_2 statistic for general χ^2-tests (Schafer, 1997; Enders, 2010; van Buuren, 2018).

B Identifiability of conditional (in)dependencies under test-wise deletion

The following lemma on the identifiability of conditional dependencies is a rephrased version of Proposition 1 in Tu et al. (2019):

Lemma 5
Let $\mathcal{D}$ be a missingness DAG with node set $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$, such that the distribution of $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$ is faithful to $\mathcal{D}$, and assume that faithful observability holds. Let $X, Y \in \mathbf{V}$ with $X \neq Y$, and let $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$ such that $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$. Then $X \not\perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1)$.

In other words, under faithfulness and faithful observability, conditional dependencies are always identified under test-wise deletion. Tu et al. (2019) also show that conditional independencies are not always identified under test-wise deletion. Our next proposition provides a necessary and sufficient criterion for this type of identification. The proof builds on Theorem 6 of Didelez et al. (2010) and is based on the following properties of distributions faithful to DAGs (Pearl, 1988, Theorem 11):

Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ be disjoint subsets of a set of random variables $\mathbf{V}$ faithful to a DAG with node set $\mathbf{V}$.

Contraction: If $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ and $\mathbf{A} \perp\!\!\!\perp \mathbf{D} \mid (\mathbf{B}, \mathbf{C})$, then $\mathbf{A} \perp\!\!\!\perp (\mathbf{B}, \mathbf{D}) \mid \mathbf{C}$.

Weak union: If $\mathbf{A} \perp\!\!\!\perp (\mathbf{B}, \mathbf{D}) \mid \mathbf{C}$, then $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid (\mathbf{C}, \mathbf{D})$.

Weak transitivity: If $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid (\mathbf{C}, D)$ and $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$, then either $D \perp\!\!\!\perp \mathbf{A} \mid \mathbf{C}$ or $D \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$. Here D is required to be a singleton.

Lemma 6
Let $\mathbf{V}$ be a set of random variables with a joint distribution satisfying the properties of contraction, weak union and weak transitivity, and assume that faithful observability holds. Let $X, Y \in \mathbf{V}$ with $X \neq Y$, and let $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$. Then $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1)$ if and only if $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$.

Proof. By faithful observability, $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1) \Leftrightarrow X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}})$. We show that (i) $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}})$ if and only if (ii) $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$.

Suppose first that (i) holds. Then by weak transitivity, we have that either $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid \mathbf{Z}$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid \mathbf{Z}$. If $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid \mathbf{Z}$, then by contraction, $(R^{XY\mathbf{Z}}, Y) \perp\!\!\!\perp X \mid \mathbf{Z}$, and by weak union, $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$. Analogously, if $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid \mathbf{Z}$, then $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$. Hence, (ii) holds.

Suppose now that $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ holds. Since $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, by contraction, $X \perp\!\!\!\perp (Y, R^{XY\mathbf{Z}}) \mid \mathbf{Z}$. By weak union, $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}})$. By symmetry, if we instead suppose that $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$, then $Y \perp\!\!\!\perp X \mid (\mathbf{Z}, R^{XY\mathbf{Z}}) \Leftrightarrow X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}})$, which completes the proof. □

We are now ready to prove Proposition 2 from Section 4.1. For simplicity, it is assumed in the proof that oracle test-wise-deletion PC is based on the original version of oracle PC. However, the proposition also holds if the stable version proposed by Colombo and Maathuis (2014), in which additional conditional independencies are considered, is used. See Colombo and Maathuis (2014) for pseudo-code for both variants.

Proposition 2
Let $\mathcal{D}$ be a missingness DAG with node set $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$, such that the distribution of $\mathbf{V} \cup \mathbf{R}(\mathbf{V})$ is faithful to $\mathcal{D}$, and assume that faithful observability holds. Then oracle test-wise-deletion PC recovers the true CPDAG over $\mathbf{V}$ if and only if the admissible separator condition holds.

Proof. We first show that the skeleton part of oracle test-wise-deletion PC recovers the true skeleton if and only if the admissible separator condition holds.

Suppose first that the skeleton is correctly recovered by oracle test-wise-deletion PC. This implies that for all pairs (X, Y) of non-adjacent nodes in $\mathbf{V}$, there exists a (possibly empty) set $\mathbf{Z} \subseteq \text{adj}(X, \mathcal{D})$ or $\mathbf{Z} \subseteq \text{adj}(Y, \mathcal{D})$ such that $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1)$, as otherwise the edge between X and Y would not have been removed during the algorithm. By Lemma 5, $X \perp\!\!\!\perp Y \mid \mathbf{Z}$. By Lemma 6, $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1)$ and $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ together imply $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$, hence the admissible separator condition is satisfied.

Suppose now that the admissible separator condition is satisfied. Pick a pair (X, Y) of non-adjacent nodes in $\mathbf{V}$ and a set $\mathbf{Z}$ satisfying the admissible separator condition with respect to (X, Y), implying $\mathbf{Z} \subseteq \text{adj}(X, \mathcal{D})$ or $\mathbf{Z} \subseteq \text{adj}(Y, \mathcal{D})$ and $R^{XY\mathbf{Z}} \perp\!\!\!\perp X \mid (Y, \mathbf{Z})$ or $R^{XY\mathbf{Z}} \perp\!\!\!\perp Y \mid (X, \mathbf{Z})$. Then by Lemma 6, $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1)$. Lemma 5 implies that no edges are erroneously deleted by oracle test-wise-deletion PC, i.e. the nodes adjacent to X in any intermediate graph obtained while the algorithm runs is a superset of $\text{adj}(X, \mathcal{D})$, and analogous for Y. Hence, $X \perp\!\!\!\perp Y \mid (\mathbf{Z}, R^{XY\mathbf{Z}} = 1)$ is among the conditional independencies tested during the course of the algorithm. It follows that the skeleton is correctly recovered.

The adjacencies are not further modified after the skeleton phase is completed. Hence, the necessary condition for the recovery of the true CPDAG is that the admissible separator condition holds. This proofs the 'only if' direction of the statement in the proposition. For the other direction, suppose that the admissible separator condition holds for the remainder of the proof.

Consider the v-structure phase of oracle test-wise-deletion PC. This phase is based on checking, for triples (X, Y, Z) such that $X - Y - Z$ is in the estimated skeleton and $X - Z$ is not, whether Y is in the separating set $\mathbf{W}$ conditionally on which X and Z were found to be independent in the skeleton phase. If $Y \notin \mathbf{W}$, then $X - Y - Z$ is oriented as $X \rightarrow Y \leftarrow Z$. We have already established above that under the admissible separator

condition, $X \perp\!\!\!\perp Z \mid (\mathbf{W}, R^{XZ\mathbf{W}}) \Leftrightarrow X \perp\!\!\!\perp Z \mid \mathbf{W}$. Hence, as we assume faithfulness, $Y \notin \mathbf{W}$ if and only if the true structure is $X \to Y \leftarrow Z$. It follows that the v-structures are correctly recovered by oracle test-wise-deletion PC under the admissible separator condition.

Finally, the orientation of additional edges is based on logical rules and returns the correct CPDAG as long as the skeleton and the v-structures have correctly been recovered. □

References

Ahrens, W., Siani, A., Adan, R., De Henauw, S., Eiben, G., Gwozdz, W., Hebestreit, A., Hunsberger, M., Kaprio, J., Krogh, V., Lissner, L., Mólnar, D., Moreno, L. A., Page, A., Pico, C., Reisch, L., Smith, R. M., Tornaritis, M., Veidebaum, T., Williams, G., Pohlabeln, H., and Pigeot, I. on behalf of the I.Family consortium (2017). Cohort profile: The transition from childhood to adolescence in European children – how I.Family extends the IDEFICS cohort. *International Journal of Epidemiology*, 46(5):1394–1395j.

Alekseyenko, A. V., Lytkin, N. I., Ai, J., Ding, B., Padyukov, L., Aliferis, C. F., and Statnikov, A. (2011). Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biology Direct*, 6(1):25.

Andrews, B., Ramsey, J., and Cooper, G. F. (2018). Scoring Bayesian networks of mixed variables. *International Journal of Data Science and Analytics*, 6(1):3–18.

Andrews, R., Foraita, R., Didelez, V., and Witte, J. (2021). A practical guide to causal discovery with cohort data. *Working paper*.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and the Alzheimer's Disease Neuroimaging Initiative (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487.

Bessler, D. A. and Yang, J. (2003). The structure of interdependence in international stock markets. *Journal of International Money and Finance*, 22(2):261–287.

Börnhorst, C., Siani, A., Russo, P., Kourides, Y., Sion, I., Molnár, D., Moreno, L. A., Rodríguez, G., Ben-Shlomo, Y., Howe, L., Lissner, L., Mehlig, K., Regber, S., Bammann, K., Foraita, R., Ahrens, W., and Tilling, K. (2016). Early life factors and inter-country heterogeneity in BMI growth trajectories of European children: The IDEFICS study. *PLOS ONE*, 11(2):1–20.

Carpenter, J. R. and Kenward, M. G. (2013). *Multiple Imputation and Its Application*. John Wiley & Sons, Chichester, UK.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press, New Haven, Connecticut, USA.

Daniel, R. M., Kenward, M. G., Cousens, S. N., and De Stavola, B. L. (2012). Using causal

diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256.

Didelez, V. (2018). Causal concepts and graphical models. In Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M., editors, *Handbook of Graphical Models*, chapter 15, pages 355–382. CRC Press.

Didelez, V., Kreiner, S., and Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387.

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

Doretti, M., Geneletti, S., and Stanghellini, E. (2018). Missing data: a unified taxonomy guided by conditional independence. *International Statistical Review*, 86(2):189–204.

Ebert-Uphoff, I. and Deng, Y. (2012). Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665.

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guilford Press, New York, USA.

Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron*, 3:329–332.

Foraita, R., Friemel, J., Günther, K., Behrens, T., Bullerdiek, J., Nimzyk, R., Ahrens, W., and Didelez, V. (2020). Causal discovery of gene regulation with incomplete data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1747–1775.

Foraita, R., Witte, J., Börnhorst, C., De Henauw, S., Gwozdz, W., Krogh, V., Lissner, L., Lauria, F., Molnár, D., Moreno, L., Page, A., Reisch, L., Veidebaum, T., Tornaritis, M., Pigeot, I., and Didelez, V. (2021). A longitudinal causal graph analysis investigating modifiable risk factors and obesity in a European cohort of children and adolescents. *Working paper*.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In Fisher, D. H., editor, *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*, pages 125–133. Morgan Kaufmann Publishers.

Gain, A. and Shpitser, I. (2018). Structure learning under missing data. *Proceedings of Machine Learning Research*, 72:121–132.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Hardt, J., Herke, M., and Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Medical Research Methodology*, 12(1):184.

Hebestreit, A., Barba, G., De Henauw, S., Eiben, G., Hadjigeorgiou, C., Kovács, É., Krogh, V., Moreno, L. A., Pala, V., Veidebaum, T., Wolters, M., and Börnhorst, C. on

behalf of the IDEFICS Consortium (2016). Cross-sectional and longitudinal associations between energy intake and BMI z-score in European children. *International Journal of Behavioral Nutrition and Physical Activity*, 13(1):1–11.

Hughes, R. A., Heron, J., Sterne, J. A., and Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*, 48(4):1294–1304.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):1–10.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Lauritzen, S. L. (1990). *Graphical Models*. Oxford University Press, Oxford, UK.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17(1):31–57.

Lee, B. Y., Bartsch, S. M., Mui, Y., Haidari, L. A., Spiker, M. L., and Gittelsohn, J. (2017). A systems approach to obesity. *Nutrition Reviews*, 75(suppl_1):94–106.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley & Sons, Hoboken, New Jersey, USA, 2nd edition.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164.

Mealli, F. and Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102(4):995–1000.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In Besnard, P. and Hanks, S., editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410. Morgan Kaufmann Publishers.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111.

Moffa, G., Catone, G., Kuipers, J., Kuipers, E., Freeman, D., Marwaha, S., Lennox, B. R., Broome, M. R., and Bebbington, P. (2017). Using directed acyclic graphs in epidemiological research in psychosis: an analysis of the role of bullying in psychosis. *Schizophrenia Bulletin*, 43(6):1273–1279.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In Burges, C. J. C., Bottou, M. W., Ghahramani, Z., and Weinberger, K. Q.,

editors, *Advances in Neural Information Processing Systems 26 (NIPS-2013)*, pages 1277–1285.

Moreno-Betancur, M., Lee, K. J., Leacy, F. P., White, I. R., Simpson, J. A., and Carlin, J. B. (2018). Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. *American Journal of Epidemiology*, 187(12):2705–2715.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press, Cambridge, UK, 2nd edition.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1):1–13.

Noghrehchi, F., Stoklosa, J., Penev, S., and Warton, D. I. (2021). Selecting the model for multiple imputation of missing data: Just use an IC! *Statistics in Medicine*, 40(10):2467–2497.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Francisco, California.

Pigeot, I., Sobotka, F., Kreiner, S., and Foraita, R. (2015). The uncertainty of a selected graphical model. *Journal of Applied Statistics*, 42(11):2335–2352.

Pohlabeln, H., Rach, S., De Henauw, S., Eiben, G., Gwozdz, W., Hadjigeorgiou, C., Molnár, D., Moreno, L. A., Russo, P., Veidebaum, T., and Iris Pigeot on behalf of the IDEFICS consortium (2017). Further evidence for the role of pregnancy-induced hypertension and other early life influences in the development of ADHD: results from the IDEFICS study. *European Child & Adolescent Psychiatry*, 26(8):957–967.

Rau, M. A. and Scheines, R. (2012). Searching for variables and models to investigate mediators of learning from multiple representations. In Yacef, K., Zaïane, O., Hershkovitz, A., Yudelson, M., and Stamper, J., editors, *Proceedings of the 5th International Conference on Educational Data Mining*, pages 110–117.

Ray, S., Haney, M., Hanson, C., Biswal, B., and Hanson, S. J. (2015). Modeling causal relationship between brain regions within the drug-cue processing network in chronic cocaine smokers. *Neuropsychopharmacology*, 40(13):2960–2968.

Roberts, S. and Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLOS ONE*, 8(8):e70902.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, USA.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC, Boca Raton, Florida, USA.

Scutari, M. (2020). Bayesian network models for incomplete and dynamic data. *Statistica Neerlandica*, 74(3):397–419.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by 'missing at random'? *Statistical Science*, 28(2):257–268.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology*, 179(6):764–774.

Sokolova, E., von Rhein, D., Naaijen, J., Groot, P., Claassen, T., Buitelaar, J., and Heskes, T. (2017). Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD. *International Journal of Data Science and Analytics*, 3(2):105–119.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press, Cambridge, Massachusetts, 2nd edition.

Steck, H. and Jaakkola, T. (2003). Bias-corrected bootstrap and model uncertainty. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, pages 521–528.

Strobl, E. V., Visweswaran, S., and Spirtes, P. L. (2018). Fast causal inference with non-random missingness by test-wise deletion. *International Journal of Data Science and Analytics*, 6(1):47–62.

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., and Ellison, G. T. H. (2021). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*, 50(2):620–632.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78.

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., and Zhang, K. (2019). Causal discovery in the presence of missing data. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, pages 1762–1770. PMLR.

Tu, R., Zhang, K., Ackermann, P., Bertilson, B. C., Glymour, C., Kjellström, H., and Zhang, C. (2020). Causal discovery in the presence of missing data. *arXiv preprint arXiv:1807.04010*.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC, Boca Raton, Florida, USA.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Vandenbroeck, P., Goossens, J., and Clemens, M. (2017). Foresight, tackling obesities: future choices building the obesity system map. Report by the UK Government Office for Science. `www.gov.uk/government/publications/reducing-obesity-obesity-system-map`.

Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology*, 23(1):159–164.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16–17):1873–1896.

5.4 Outlook: Challenges in constraint-based causal discovery

Over the last years, constraint-based causal discovery has become more versatile and reliable, due to e.g. the development of the index-stable variants (Colombo and Maathuis, 2014) and new methods for dealing with missing values (see Sections 5.2 and 5.3). However, several challenges remain.

A fundamental assumption of constraint-based causal discovery is faithfulness. It is sometimes argued that this is a comparatively mild assumption, as the set of unfaithful distributions compatible with a given DAG has Lebesgue measure zero; this holds for Gaussian as well as multinomial distributions (Spirtes et al., 2000, p. 68; Meek, 1995b). However, when PC is applied to finite data, then even almost-violations of the faithfulness assumption can have a detrimental effect on the algorithm's performance. It has been shown that the proportion of near-unfaithful distributions among all distributions compatible with a given DAG can be large even for relatively small and sparse graphs (Uhler et al., 2013). As a consequence of (near-)unfaithfulness, the PC-algorithm will tend to delete certain edges even though they are present in the true underlying DAG. This is a fundamental problem of all constraint-based causal discovery algorithms and can only be mitigated by using a large sample size and a conditional independence test with a high power for detecting dependencies.

The choice of conditional independence test, however, poses a challenge of its own. Ideally, the test would not make distributional assumptions and at the same time have power against all possible alternatives, while respecting the nominal α level. The G^2-test applied in Paper 3 (Section 5.3) fulfils all of these criteria, but it can only be applied to discrete data. For continuous variables, constructing a test that meets all three criteria is provably impossible (Shah and Peters, 2020). There thus exists a trade-off between power and generality of assumptions. More traditional tests for continuous variables, such as Fisher's z-test (Fisher, 1924) or the rank correlation test by Liu et al. (2018) are computationally efficient and have high power under their respective parametric or monotonicity assumptions. More recently proposed tests make fewer assumptions and are often computationally demanding. Examples include tests based on reproducing kernel Hilbert spaces (Zhang et al., 2011), independence testing between residuals after non-parametric regression (Shah and Peters, 2020) and tests using knockoffs (Watson and Wright, 2021). Despite the large number of strategies that have been proposed, conditional independence testing between continuous variables is still considered an open problem

(Li and Fan, 2020).

Of particular importance for causal discovery from cohort data is that the conditional independence test can handle data measured on mixed scales. In Paper 3 (Section 5.3), we applied the Conditional Gaussian test by Andrews et al. (2018). The simulation results presented in Paper 3 demonstrated that despite the strong parametric assumptions, the power of the test can be low when the number of conditioning variables is large and/or the discrete variables have many categories. Another challenge is that the test reduces to a test of the partial correlation, similar to Fisher's z-test, when only continuous variables are involved, and to the G^2-test when all variables are discrete. Since the partial correlation test has a much higher power, on average, than the non-parametric G^2-test, the estimated graphs tend to have more edges between pairs of continuous variables than between pairs of discrete variables, which needs to be taken into account when interpreting the result. As an alternative to the Conditional Gaussian test, Andrews et al. (2019) proposed the Degenerate Gaussian test, where categorical variables are coded as dummy variables and then treated as if they were continuous. In their simulation study, the authors compared the Degenerate Gaussian method and the Conditional Gaussian method in the context of score-based learning and found that the Degenerate Gaussian method performed better in terms of edge recall and precision. It would be interesting to investigate whether this is also the case for conditional independence testing. An alternative parametric approach is to fit (main effects) linear or logistic regression models and then perform a t-test or likelihood ratio test, respectively, to test for a conditional independence (Raghu et al., 2018). This is problematic because the parametric assumptions of the different tests will in general contradict each other, as is the case for the Conditional Gaussian test. A non-parametric Bayesian test for mixed data was proposed by Boeken and Mooij (2021), but is only applicable when the conditioning set consists of a single continuous variable.

As the PC-algorithm and its variants have different stages building on each other's results, erroneous decisions due to e.g. misspecification of the conditional independence test or low power can propagate through the algorithm. Quantifying the uncertainty associated with the final graph estimate is challenging. In Paper 3 (Section 5.3), we bootstrapped the analysis in order to get an impression of the overall stability of the result. It is also common to combine the graphs estimated on the bootstrap samples into a summary graph (Pigeot et al., 2015), as done in Foraita, Witte et al. (2021). However, as also reported in Paper 3, bootstrapping a causal discovery analysis can have undesired effects such as bootstrap estimates

containing considerably more edges than the original estimate (Steck and Jaakkola, 2003). Another line of literature is concerned with estimating and controlling the false discovery rate of the PC-algorithm in order to quantify the uncertainty (Li and Wang, 2009; Armen and Tsamardinos, 2011; Strobl et al., 2019).

Another open question, especially when the aim is to apply the IDA algorithm (see Section 4.1.2) to the estimated graph, is how to obtain a valid CPDAG or MP-DAG. Ad-hoc solutions include a trial-and error strategy where the edges marked as ambiguous are oriented until the result is valid (implemented in `pcalg`, Kalisch et al., 2012), randomly orienting edges while ignoring the detected v-structures (implemented in `pcalg`) and restricting (optimal) IDA to potential causal nodes and potential parents thereof (implemented in `tpc`, Witte, 2021). However, a theoretically founded solution would be preferable.

Summarising, causal discovery involves many challenges, all of which have been tackled or are currently being tackled in the literature. At the moment, the methodological papers still outnumber the papers in which causal discovery is applied to real data, and many of the existing applications primarily serve as a proof of concept instead of generating genuinely new insights. The near future will show whether the obstacles can be overcome, and how useful causal discovery proves to be in real-world applications.

6 Discussion and conclusion

In this thesis, I investigated aspects of confounder selection and causal discovery. It became apparent that many data-driven confounder selection algorithms are based on conditional independence testing and are thus closely related to methods for constraint-based causal discovery. Stepwise regression selection of an outcome model essentially aims at discovering the Markov blanket of the outcome. Assuming that the procedure starts with a valid adjustment set and that all conditional independencies are correctly inferred, the selected set is valid and yields an efficient estimator, see Sections 3.4.5 and 4.5. In contrast, procedures for univariate regression selection only test for marginal independencies, which can lead to undesired results, as seen in Section 3.4.4. Some selection algorithms explicitly use causal discovery: An alternative implementation of the CovSel algorithm, which is suited for high-dimensional selection problems, uses causal discovery for iteratively searching for the Markov blanket of the treatment and the outcome variable (Häggström et al., 2015). Another example is the confounder selection algorithm by Entner et al. (2013), which is a restricted version of FCI. Conversely, constraint-based causal discovery can also be stated as a variable selection problem, and be solved using methods for regression selection (Wang and Michailidis, 2019).

It also became clear in Section 3.4 that confounder selection, just like causal discovery, relies on strong causal assumptions. In particular, almost all of the algorithms considered assume 'no unobserved confounding' (see Section 3.1) relative to the observed variables, meaning that the full set of observed covariates is a valid adjustment set. Further, the method by Entner et al. (2013) is the only one that is able to detect violations of this assumption. The 'no unobserved confounding' assumption is related to the causal sufficiency assumption, which essentially states that there is no unobserved confounding between any two variables in the analysis. Similar to many methods for confounder selection, constraint-based causal discovery algorithms relying on causal sufficiency are not able to detect violations thereof.

In light of the strong assumptions common confounder selection strategies rely

on, it is, once again, obvious that background knowledge is a central prerequisite for any causal analysis. The question is, however, whether this knowledge is best presented in form of a causal DAG, or if there are good and maybe simpler alternatives. It has been argued that attempting to draw a causal DAG for a given research question comes with more problems than benefits (Greenland, 2010), and it seems indeed unrealistic that the causal structure underlying any system of interest in epidemiology can be fully known (Pigeot and Foraita, 2011). However, in my opinion, the benefits of using causal DAGs for confounder selection clearly outweigh the challenges.

First, it must be emphasised that for choosing a valid adjustment set, not all details of the causal DAG need to be known. Consider the disjunctive cause criterion (VanderWeele and Shpitser, 2011) from Section 3.4.3: This only requires knowledge about which of the observed covariates are causes of the treatment or the outcome, plus a justification of the absence of any unobserved common causes. Hence, the exact causal relations among the observed covariates are not relevant. In contrast, it is of central importance to consider unobserved factors, which is arguably best done by sketching a causal graph showing both measured and unmeasured covariates.

Second, if there is uncertainty even about the basic causal structure, e.g. about whether a group of covariates should be considered as mediators or confounding factors, then identifying and acting upon this lack of knowledge is even more important. In some cases, this will mean choosing a new research question or study design. After all, causal assumptions that are not explicitly made are still made, even if they go unnoticed. This is in contrast to predictive and descriptive modelling, where a causal interpretation is optional.

Third, the conditional (in)dependencies implied by a causal DAG are empirically testable. They can thus be used to assess how well a postulated causal DAG fits the data (Ankan et al., 2021), or to repair a misspecified causal DAG (Oates et al., 2017). Oddly enough, these approaches for checking the plausibility of a specified causal DAG are often overlooked.

Forth, causal DAGs are useful not only for determining a valid adjustment set, but also for representing and assessing consequences of selection bias (Hernán et al., 2002), measurement error (Hernán and Cole, 2009) and missing data (see Section 5.3), among other things, under a common, intuitive framework.

Summarising, causal DAGs are a powerful tool for identifying sources of structural

bias in a given causal analysis. This does not imply, however, that drawing a useful causal DAG for a research question at hand is always an easy task. A particularly challenging aspect is causal feedback.

In principle, feedback can be represented in a DAG by adding a time axis. Consider the DAG in Figure 14, showing assumed causal relations between sleep quality and well-being on consecutive days. It seems reasonable to assume that well-being on any given day influences sleep quality in the following night, which in turn influences well-being on the next day. Additionally, it could be assumed e.g. that sleep quality in a given night also influences well-being two days later, which would be represented in the DAG in Figure 14 by a directed edge from $sleep_{i-1}$ to $well\text{-}being_{i+1}$.

In practice, however, data may not be available in such a high temporal resolution. Suppose that a dataset contains measures of average sleep quality and average well-being for week 1 and week 2. Then neither of the DAGs in Figure 15 is an accurate representation of the causal structure, assuming that Figure 14 is accurate. Similar issues can occur when the process is *subsampled*, i.e. measurements are taken e.g. only once a week (Runge, 2018). Another related situation emerges when interest lies in an event such as incident drug use, but the exact time point of the event is not known. For example, if it is only known that a drug was taken in the third quarter of the calendar year, then symptoms or diagnoses recorded in the same quarter could be either due to diseases that caused the patient to start taking the drug, or indicate side effects. When the aim is to estimate a causal effect, such situations can sometimes be avoided e.g. by re-defining the treatment or the outcome, or at least mitigated by performing sensitivity analyses. However, when the aim is causal discovery, there is no obvious solution.

Aalen et al. (2012) and Aalen et al. (2016) discussed another issue, going beyond aggregation and subsampling, which is causal effects that are transmitted continuously in time. As an example, consider *BMI* and *calory intake*. It is not clear how to draw a graph similar to the one in Figure 14 for these two variables. Aalen et al. (2012) and Aalen et al. (2016) argued that stochastic differential equations can be a suitable way of describing such time-continuous causal relations. These are tied to the concept of local independence, which describes independence relations between stochastic processes instead of random variables. Local independencies are represented in so-called local independence graphs (Didelez, 2008), which can be given a causal interpretation (Røysland, 2012; Didelez, 2015), and can be learned using specialised causal discovery algorithms (Meek, 2014; Mogensen et al., 2018).

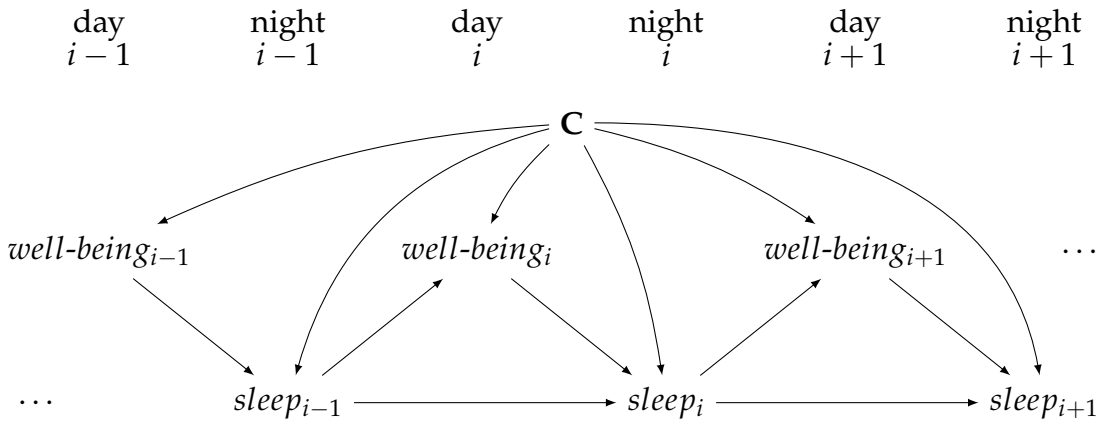


Figure 14: A DAG representing assumed causal relations between daily well-being and sleep quality. Measured and unmeasured confounding is represented by **C**.

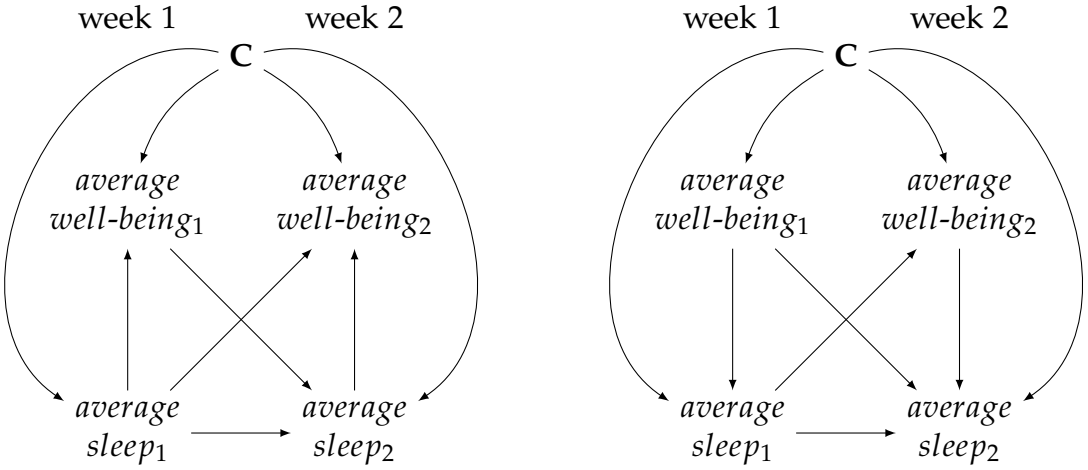


Figure 15: DAGs containing nodes for average well-being and sleep quality at two time points. Measured and unmeasured confounding is represented by **C**.

Cyclic graphs are another type of graph that may be useful for representing causal feedback. Mooij and Claassen (2020) showed that the output of the PC-algorithm or the FCI algorithm can alternatively be interpreted as representing an equivalence class of cyclic graphs. Yet another way of accounting for causal feedback is through causal chain graphs (Lauritzen and Richardson, 2002). These have directed and undirected edges, where an undirected edge represents an association between two variables in a state of equilibrium.

Causal feedback, subsampling, aggregation and continuous-time causation are all common issues in cohort data. Are DAGs the best option for representing causal mechanism, or should the time aspect and the possibility of causal feedback be taken into account by other types of graphs? I believe that this question deserves more attention. Additionally, for causal discovery in particular, it is also important to allow for unmeasured confounding, or lack of causal sufficiency, such as the FCI algorithm does, which outputs a PAG. My conclusion is that graphs are immensely useful for representing causal structures, but that more attention should be paid to graphs other than DAGs.

A Pseudocode

This appendix includes pseudocode for PC, LMPC-stable and tPC.

A.1 PC-algorithm

Algorithm 2 contains pseudocode for the PC-algorithm as described in Spirtes et al. (2000), with Meek's rules (Meek, 1995a) for edge orientation.

Algorithm 2 PC

INPUT: i.i.d. data on a set of variables $\mathbf{V}$; a procedure for testing conditional independencies

(I) skeleton phase

form the complete undirected graph $\mathcal{C}$ on node set $\mathbf{V}$
$\mathcal{C}' = \mathcal{C}$
$\ell = -1$
repeat
  $\ell = \ell + 1$
  repeat
    select new ordered pair of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ and $|\text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}| \geq \ell$
    repeat
      choose new $\mathbf{S} \subseteq \text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$
      if $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$ **then**
        delete edge $V_i - V_j$ from $\mathcal{C}'$
        record $\mathbf{S}$ as **Sepset**(V_i, V_j)
      end if
    until edge $V_i - V_j$ is deleted or all $\mathbf{S} \subseteq \text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$ have been chosen
  until all ordered pairs of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ have been chosen
until all ordered pairs of nodes (V_i, V_j) adjacent in $\mathcal{C}'$ satisfy $|\text{sib}_{\mathcal{C}'}(V_i) \setminus \{V_j\}| < \ell$

(II) detection of v-structures

17: $\mathcal{C}'' = \mathcal{C}'$
18: **repeat**
19: select new unordered triple of nodes (V_i, V_j, V_k) such that $\mathcal{C}''$ contains a path $V_i - V_j - V_k$, and V_i and V_k are non-adjacent
20: **if** $V_j \notin$ **Sepset**(V_i, V_k) **then**
21: replace $V_i - V_j - V_k$ by $V_i \to V_j \leftarrow V_k$ in $\mathcal{C}''$
22: **end if**
23: **until** all unordered triples of nodes (V_i, V_j, V_k) such that $\mathcal{C}''$ contains a path $V_i - V_j - V_k$, and V_i and V_k are non-adjacent have been selected

(III) application of Meek's rules

24: $\mathcal{C}''' = \mathcal{C}''$
25: apply the following four rules repeatedly to undirected edges in $\mathcal{C}''''$ until no further edges can be oriented:
Rule 1: Orient $V_i - V_j$ into $V_i \to V_j$ if $\mathcal{C}''''$ contains an edge $V_k \to V_i$ such that V_k and V_i are non-adjacent.
Rule 2: Orient $V_i - V_j$ into $V_i \to V_j$ if $\mathcal{C}''''$ contains a path $V_i \to V_k \to V_j$.
Rule 3: Orient $V_i - V_j$ into $V_i \to V_j$ if $\mathcal{C}''''$ contains two non-adjacent nodes V_k and V_m that are siblings of V_i and parents of V_j.

OUTPUT: estimated CPDAG $\mathcal{C}'''$

A.2 LMPC-stable

Algorithm 3 contains pseudocode for the LMPC-stable algorithm by Colombo and Maathuis (2014). The 'L' stands for 'lists', see lines 20, 50, 59, 68 and 77 of the algorithm. Here, information about edges to be deleted and arrow heads to be added is first collected until all variables have been considered, and only then is the graph transformed accordingly. The 'M' means 'majority rule'. This refers to how the v-structure orientation works if the different conditional independence tests performed in line 25 give conflicting results: If more than half of the test decisions indicate that a v-structure is present or not present, then the arrow heads are oriented accordingly or remain undirected, respectively. In the event of a tie, the edges remain undirected and the triple is added to the list of ambiguous triples.

Algorithm 3 LMPC-stable

INPUT: i.i.d. data on a set of variables $\mathbf{V}$; a procedure for testing conditional independencies

(I) skeleton phase

form the complete undirected graph $\mathcal{C}$ on node set $\mathbf{V}$
$\mathcal{C}' = \mathcal{C}$
$\ell = -1$
repeat
  $\ell = \ell + 1$
  for all nodes V_i in $\mathbf{V}$ **do**
    $\mathrm{a}(V_i) = \mathrm{sib}_{\mathcal{C}'}(V_i)$
  end for
  repeat
    select new ordered pair of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ and $|\mathrm{a}(V_i) \setminus \{V_j\}| \geq \ell$
    repeat
      choose new $\mathbf{S} \subseteq \mathrm{a}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$
      if $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$ **then**
        delete edge $V_i - V_j$ from $\mathcal{C}'$
      end if
    until edge $V_i - V_j$ is deleted or all $\mathbf{S} \subseteq \mathrm{a}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$ have been chosen
  until all ordered pairs of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ and $|\mathrm{a}(V_i) \setminus \{V_j\}| \geq \ell$ have been chosen
until all ordered pairs of nodes (V_i, V_j) adjacent in $\mathcal{C}'$ satisfy $|\mathrm{a}(V_i) \setminus \{V_j\}| < \ell$

(II) detection of v-structures

$\mathcal{C}'' = \mathcal{C}'$
create empty lists *orient_list* and *ambiguous_list*
repeat
  select new unordered triple of nodes (V_i, V_j, V_k) such that $\mathcal{C}''$ contains a path $V_i - V_j - V_k$, and V_i and V_k are non-adjacent
  $counter = 0$
  for all sets $\mathbf{S}$ such that $\mathbf{S}$ is a subset of $\mathrm{sib}_{\mathcal{C}''}(V_i)$ or $\mathrm{sib}_{\mathcal{C}''}(V_k)$ or both **do**
    if $V_i \perp\!\!\!\perp V_k \mid \mathbf{S}$ **then**
      if $V_j \in \mathbf{S}$ **then**
        $counter = counter - 1$
      else
        $counter = counter + 1$
      end if
    end if
  end for

33: **if** $counter = 0$ **then**
34: add (V_i, V_j) to *ambiguous_list*
35: add (V_j, V_k) to *ambiguous_list*
36: **else if** $counter > 0$ **then**
37: add (V_i, V_k) to *orient_list*
38: add (V_k, V_j) to *orient_list*
39: **end if**
40: **until** all unordered triples of nodes (V_i, V_j, V_k) such that $\mathcal{C}''$ contains a path $V_i - V_j - V_k$ and V_i and V_k are non-adjacent have been selected
41: **for** all pairs (V_m, V_n) in *orient_list* **do**
42: **if** $V_m - V_n$ is in $\mathcal{C}''$ **then**
43: replace by $V_m \rightarrow V_n$
44: **else if** $V_m \leftarrow V_n$ is in $\mathcal{C}''$ **then**
45: replace by $V_m \leftrightarrow V_n$
46: **end if**
47: **end for**

(III) application of Meek's rules

48: $\mathcal{C}''' = \mathcal{C}''$
49: **repeat**
50: create empty list *rule*1*_list*
51: apply the following rule to as many undirected edges in $\mathcal{C}'''$ not listed in *ambiguous_list* as possible:
Rule 1: Add (V_i, V_j) to *rule*1*_list* if $\mathcal{C}'''$ contains an edge $V_k \rightarrow V_i$ such that V_k and V_i are non-adjacent.
52: **for** all pairs (V_m, V_n) in *rule*1*_list* **do**
53: **if** $V_m - V_n$ is in $\mathcal{C}'''$ **then**
54: replace by $V_m \rightarrow V_n$
55: **else if** $V_m \leftarrow V_n$ is in $\mathcal{C}'''$ **then**
56: replace by $V_m \leftrightarrow V_n$
57: **end if**
58: **end for**
59: create empty list *rule*2*_list*
60: apply the following rule to as many undirected edges in $\mathcal{C}'''$ not listed in *ambiguous_list* as possible:
Rule 2: Add (V_i, V_j) to *rule*2*_list* if $\mathcal{C}'''$ contains a path $V_i \rightarrow V_k \rightarrow V_j$.
61: **for** all pairs (V_m, V_n) in *rule*2*_list* **do**
62: **if** $V_m - V_n$ is in $\mathcal{C}'''$ **then**
63: replace by $V_m \rightarrow V_n$
64: **else if** $V_m \leftarrow V_n$ is in $\mathcal{C}'''$ **then**
65: replace by $V_m \leftrightarrow V_n$
66: **end if**
67: **end for**

create empty list *rule*3_*list*
apply the following rule to as many undirected edges in $\mathcal{C}'''$ not listed in *ambiguous_list* as possible:
Rule 3: Add (V_i, V_j) to *rule*3_*list* if $\mathcal{C}'''$ contains two non-adjacent nodes V_k and V_m that are siblings of V_i and parents of V_j.
for all pairs (V_m, V_n) in *rule*3_*list* **do**
  if $V_m - V_n$ is in $\mathcal{C}'''$ **then**
    replace by $V_m \to V_n$
  else if $V_m \leftarrow V_n$ is in $\mathcal{C}'''$ **then**
    replace by $V_m \leftrightarrow V_n$
  end if
end for
create empty list *rule*4_*list*
apply the following rule to as many undirected edges in $\mathcal{C}''''$ not listed in *ambiguous_list* as possible:
Rule 4: Add V_i, V_j to *rule*4_*list* if $\mathcal{C}''''$ contains a path $V_k \to V_m \to V_j$ such that V_k and V_m are siblings of V_i, and V_j and V_k are non-adjacent.
for all pairs (V_m, V_n) in *rule*4_*list* **do**
  if $V_m - V_n$ is in $\mathcal{C}'''$ **then**
    replace by $V_m \to V_n$
  else if $V_m \leftarrow V_n$ is in $\mathcal{C}'''$ **then**
    replace by $V_m \leftrightarrow V_n$
  end if
end for
until no further edges can be oriented

OUTPUT: estimated CPDAG $\mathcal{C}'''$

A.3 tPC

Algorithm 4 contains pseudocode for the tPC-algorithm described in Section 5.1 and implemented in the `tpc` package (Witte, 2021). The package also contains functions for specifying context nodes that are forced to have edges into all other (non-context) nodes in the graph or all other (non-context) nodes in their respective tiers.

Algorithm 4 tPC

INPUT: i.i.d. data on a set of variables $\mathbf{V}$; a procedure for testing conditional independencies; a partial topological ordering $\mathbf{V}^1 < \cdots < \mathbf{V}^T$

(I) skeleton phase

form the complete undirected graph $\mathcal{C}$ on node set $\mathbf{V}$
$\mathcal{C}' = \mathcal{C}$
$\ell = -1$
repeat
  $\ell = \ell + 1$
  for all nodes V_i in $\mathbf{V}$ **do**
    $\mathrm{a}(V_i) = \mathrm{sib}_{\mathcal{C}'}(V_i)$
  end for
  repeat
    select new ordered pair of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ and $|\mathrm{a}(V_i) \setminus \{V_j\}| \geq \ell$
    repeat
      choose new $\mathbf{S} \subseteq \mathrm{a}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$ and for all $S \in \mathbf{S}$, $t(S) \leq t(V_i)$
      if $V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$ **then**
        delete edge $V_i - V_j$ from $\mathcal{C}'$
      end if
    until edge $V_i - V_j$ is deleted or all $\mathbf{S} \subseteq \mathrm{a}(V_i) \setminus \{V_j\}$ such that $|\mathbf{S}| = \ell$ and for all $S \in \mathbf{S}$, $t(S) \leq t(V_i)$ have been chosen
  until all ordered pairs of nodes (V_i, V_j) such that V_i and V_j are adjacent in $\mathcal{C}'$ and $|\mathrm{a}(V_i) \setminus \{V_j\}| \geq \ell$ have been chosen
until all ordered pairs of nodes (V_i, V_j) adjacent in $\mathcal{C}'$ satisfy $|\mathrm{a}(V_i) \setminus \{V_j\}| < \ell$

(II) detection of v-structures

$\mathcal{C}'' = \mathcal{C}'$
create empty lists *orient_list* and *ambiguous_list*
repeat
  select new unordered triple of nodes (V_i, V_j, V_k) such that (i) $\mathcal{C}''$ contains a path $V_i - V_j - V_k$, (ii) V_i and V_k are non-adjacent in $\mathcal{C}''$, and (iii) $t(V_j) = \max(t(V_i), t(V_k))$
  $\mathrm{sib}^*_{\mathcal{C}''}(V_i) = \{V \mid V \in \mathrm{sib}_{\mathcal{C}''}(V_i) \text{ and } t(V) \leq t(V_i)\}$
  $\mathrm{sib}^*_{\mathcal{C}''}(V_k) = \{V \mid V \in \mathrm{sib}_{\mathcal{C}''}(V_k) \text{ and } t(V) \leq t(V_i)\}$
  $counter = 0$
  for all sets $\mathbf{S}$ such that $\mathbf{S}$ is a subset of $\mathrm{sib}^*_{\mathcal{C}''}(V_i)$ or $\mathrm{sib}^*_{\mathcal{C}''}(V_k)$ or both **do**
    if $V_i \perp\!\!\!\perp V_k \mid \mathbf{S}$ **then**
      if $V_j \in \mathbf{S}$ **then**
        $counter = counter - 1$
      else
        $counter = counter + 1$
      end if

end if
end for
if $counter = 0$ **then**
add (V_i, V_j) to *ambiguous_list*
add (V_j, V_k) to *ambiguous_list*
else if $counter > 0$ **then**
add (V_i, V_k) to *orient_list*
add (V_k, V_j) to *orient_list*
end if
until all unordered triples of nodes (V_i, V_j, V_k) such that $\mathcal{C}''$ contains a path $V_i - V_j - V_k$, V_i and V_k are non-adjacent in $\mathcal{C}''$ and $t(V_j) = \max(t(V_i), t(V_k))$ have been selected
for all pairs (V_m, V_n) in *orient_list* **do**
if $V_m - V_n$ is in $\mathcal{C}''$ **then**
replace by $V_m \rightarrow V_n$
else if $V_m \leftarrow V_n$ is in $\mathcal{C}''$ **then**
replace by $V_m \leftrightarrow V_n$
end if
end for

(III) orientation of between-block edges

$\mathcal{C}''' = \mathcal{C}''$
for all nodes V_i in **V** **do**
for all nodes V_j with $\mathrm{t}(V_j) > \mathrm{t}(V_i)$ **do**
if $V_i - V_j$ is in $\mathcal{C}'''$ **then**
replace by $V_i \rightarrow V_j$
end if
end for
end for

(IV) application of Meek's rules

$\mathcal{C}'''' = \mathcal{C}'''$
repeat
create empty list *rule*1*_list*
apply the following rule to as many undirected edges in $\mathcal{C}''''$ not listed in *ambiguous_list* as possible:
Rule 1: Add (V_i, V_j) to *rule*1*_list* if $\mathcal{C}''''$ contains an edge $V_k \rightarrow V_i$ such that V_k and V_i are non-adjacent.
for all pairs (V_m, V_n) in *rule*1*_list* **do**
if $V_m - V_n$ is in $\mathcal{C}''''$ **then**
replace by $V_m \rightarrow V_n$
else if $V_m \leftarrow V_n$ is in $\mathcal{C}''''$ **then**
replace by $V_m \leftrightarrow V_n$
end if
end for

create empty list *rule*2_*list*
apply the following rule to as many undirected edges in $\mathcal{C}''''$ not listed in *ambiguous_list* as possible:
Rule 2: Add (V_i, V_j) to *rule*2_*list* if $\mathcal{C}''''$ contains a path $V_i \to V_k \to V_j$.
for all pairs (V_m, V_n) in *rule*2_*list* **do**
  if $V_m - V_n$ is in $\mathcal{C}''''$ **then**
    replace by $V_m \to V_n$
  else if $V_m \leftarrow V_n$ is in $\mathcal{C}''''$ **then**
    replace by $V_m \leftrightarrow V_n$
  end if
end for
create empty list *rule*3_*list*
apply the following rule to as many undirected edges in $\mathcal{C}''''$ not listed in *ambiguous_list* as possible:
Rule 3: Add (V_i, V_j) to *rule*3_*list* if $\mathcal{C}''''$ contains two non-adjacent nodes V_k and V_m that are siblings of V_i and parents of V_j.
for all pairs (V_m, V_n) in *rule*3_*list* **do**
  if $V_m - V_n$ is in $\mathcal{C}''''$ **then**
    replace by $V_m \to V_n$
  else if $V_m \leftarrow V_n$ is in $\mathcal{C}''''$ **then**
    replace by $V_m \leftrightarrow V_n$
  end if
end for
create empty list *rule*4_*list*
apply the following rule to as many undirected edges in $\mathcal{C}''''$ not listed in *ambiguous_list* as possible:
Rule 4: Add (V_i, V_j) to *rule*4_*list* if $\mathcal{C}''''$ contains a path $V_k \to V_m \to V_j$ such that V_k and V_m are siblings of V_i, and V_j and V_k are non-adjacent.
for all pairs (V_m, V_n) in *rule*4_*list* **do**
  if $V_m - V_n$ is in $\mathcal{C}''''$ **then**
    replace by $V_m \to V_n$
  else if $V_m \leftarrow V_n$ is in $\mathcal{C}''''$ **then**
    replace by $V_m \leftrightarrow V_n$
  end if
end for
until no further edges can be oriented

OUTPUT: estimated maxPDAG $\mathcal{C}''''$

Bibliography

Aalen, O. O., Røysland, K., Gran, J. M., Kouyos, R., and Lange, T. (2016). Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. *Statistical Methods in Medical Research*, 25(5):2294–2314.

Aalen, O. O., Røysland, K., Gran, J. M., and Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(4):831–861.

Ahrens, W., Siani, A., Adan, R., De Henauw, S., Eiben, G., Gwozdz, W., Hebestreit, A., Hunsberger, M., Kaprio, J., Krogh, V., Lissner, L., Mólnar, D., Moreno, L. A., Page, A., Pico, C., Reisch, L., Smith, R. M., Tornaritis, M., Veidebaum, T., Williams, G., Pohlabeln, H., and Pigeot, I. on behalf of the I.Family consortium (2017). Cohort profile: The transition from childhood to adolescence in European children — how I.Family extends the IDEFICS cohort. *International Journal of Epidemiology*, 46(5):1394–1395j.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alekseyenko, A. V., Lytkin, N. I., Ai, J., Ding, B., Padyukov, L., Aliferis, C. F., and Statnikov, A. (2011). Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biology Direct*, 6(1):25.

Ali, M. S., Groenwold, R. H. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C. B., de Boer, A., and Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology*, 68(2):122–131.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, USA, 2nd edition.

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541.

Andrews, B., Ramsey, J., and Cooper, G. F. (2018). Scoring Bayesian networks of mixed variables. *International Journal of Data Science and Analytics*, 6(1):3–18.

Andrews, B., Ramsey, J., and Cooper, G. F. (2019). Learning high-dimensional directed acyclic graphs with mixed data-types. *Proceedings of Machine Learning Research*, 104:4–21.

Andrews, R., Foraita, R., Didelez, V., and Witte, J. (2021). A practical guide to causal discovery with cohort data. *arXiv preprint arXiv:2108.13395*.

Ankan, A., Wortel, I. M. N., and Textor, J. (2021). Testing graphical causal models using the R package 'dagitty'. *Current Protocols*, 1(2):e45.

Armen, A. P. and Tsamardinos, I. (2011). A unified approach to estimation and control of the false discovery rate in bayesian network skeleton identification. In *Proceedings of the Nineteenth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2011)*, Bruges, Belgium. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.226.7818&rep=rep1&type=pdf.

Asvatourian, V., Leray, P., Michiels, S., and Lanoy, E. (2020). Integrating expert's knowledge constraint of time dependent exposures in structure learning for Bayesian networks. *Artificial Intelligence in Medicine*, 107:101874.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*, 26(4):734–753.

Bhattacharya, J. and Vogt, W. B. (2007). Do instrumental variables belong in propensity scores? Technical report, NBER Technical Working Paper No. 343, National Bureau of Economic Research, Cambridge, MA, USA. Revised 2009. http://www.nber.org/papers/t0343.

Boeken, P. A. and Mooij, J. M. (2021). A Bayesian nonparametric conditional two-sample test with an application to local causal discovery. *arXiv preprint arXiv:2008.07382*. Accepted for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI-21).

Bollen, K. A. and Pearl, J. (2013). Eight myths about causality and structural equation models. In Morgan, S. L., editor, *Handbook of Causal Analysis for Social Research*, pages 301–328. Springer, Dordrecht, The Netherlands.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156.

Chatton, A., Le Borgne, F., Leyrat, C., Gillaizeau, F., Rousseau, C., Barbin, L., Laplaud, D., Léger, M., Giraudeau, B., and Foucher, Y. (2020). G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific Reports*, 10:9219.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press, New Haven, Connecticut, USA.

Dablander, F. (2020). An introduction to causal inference. *PsyArXiv preprint*, https://psyarxiv.com/b3fkw.

Daniel, R., Zhang, J., and Farewell, D. (2021). Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 63(3):528–557.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189. Corrigenda, ibid. (3):437.

Dawid, A. P. (2010). Beware of the DAG! In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, pages 59–86, Whistler, Canada. `https://proceedings.mlr.press/v6/dawid10a/dawid10a.pdf`.

Dawid, A. P. and Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statistics Surveys*, 4:184–231.

de Campos, L. M., Cano, A., Castellano, J. G., and Moral, S. (2019). Combining gene expression data and prior knowledge for inferring gene regulatory networks via Bayesian networks using structural restrictions. *Statistical Applications in Genetics and Molecular Biology*, 18(3):20180042.

de Campos, L. M. and Castellano, J. G. (2007). Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, 45(2):233–254.

de Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Denis, D. and Legerski, J. (2006). Causal modeling and the origins of path analysis. *Theory & Science*, 7(1):2–10.

Derryberry, D., Aho, K., Edwards, J., and Peterson, T. (2018). Model selection and regression t-statistics. *The American Statistician*, 72(4):379–381.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264.

Didelez, V. (2015). Causal reasoning for events in continuous time: a decision-theoretic approach. In Silva, R., Shpitser, I., Evans, R., Peters, J., and Claassen, T., editors, *Proceedings of the UAI 2015 Workshop on Advances in Causal Inference*, pages 40–45, Amsterdam, The Netherlands. `http://ceur-ws.org/Vol-1504/uai2015aci_paper3.pdf`.

Didelez, V. (2018). Causal concepts and graphical models. In Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M., editors, *Handbook of Graphical Models*, pages 355–382. CRC Press, Boca Raton, Florida, USA.

Didelez, V. and Pigeot, I. (1998). Maximum likelihood estimation in graphical models with missing values. *Biometrika*, 85(4):960–966.

Ding, P. and Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-Bias and Butterfly-Bias. *Journal of Causal Inference*, 3(1):41–57.

Ding, P., Vanderweele, T. J., and Robins, J. M. (2017). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104(2):291–302.

Diop, A., Lefebvre, G., Duchaine, C. S., Laurin, D., and Talbot, D. (2021). The impact of adjusting for pure predictors of exposure, mediator, and outcome on the variance of natural direct and indirect effect estimators. *Statistics in Medicine*, 40(10):2339–2354.

Eigenmann, M. F., Nandy, P., and Maathuis, M. H. (2017). Structure learning of linear gaussian structural equation models with weak edges. In Elidan, G., Kersting, K., and Ihler, A. T., editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI-17)*, page ID: 229, Sydney, Australia. AUAI Press. https://auai.org/uai2017/proceedings/papers/229.pdf.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.

Entner, D., Hoyer, P. O., and Spirtes, P. (2013). Data-driven covariate selection for nonparametric estimation of causal effects. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2016)*, pages 256–264, Scottsdale, Arizona, USA. PMLR. https://proceedings.mlr.press/v31/entner13a.pdf.

Fahrmeir, L., Künstler, R., Pigeot, I., and Tutz, G. (2003). *Statistik. Der Weg zur Datenanalyse. [Statistics. The Path to Data Analysis.]* Springer, Berlin, Germany, 4th edition. In German.

Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron*, 3:329–332.

Foraita, R., Friemel, J., Günther, K., Behrens, T., Bullerdiek, J., Nimzyk, R., Ahrens, W., and Didelez, V. (2020). Causal discovery of gene regulation with incomplete data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1747–1775.

Foraita, R. and Witte, J. (2020–2021). Multiple imputation for causal graph discovery (micd). https://github.com/bips-hb/micd.

Foraita, R., Witte, J., Börnhorst, C., De Henauw, S., Gwozdz, W., Krogh, V., Lissner, L., Lauria, F., Molnár, D., Moreno, L., Page, A., Reisch, L., Veidebaum, T., Tornaritis, M., Pigeot, I., and Didelez, V. (2021). A longitudinal causal graph analysis investigating modifiable risk factors and obesity in a European cohort of children and adolescents. *Working paper.*

Franzin, A., Sambo, F., and Di Camillo, B. (2017). bnstruct: an R package for Bayesian network structure learning in the presence of missing data. *Bioinformatics*, 33(8):1250–1252.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In Fisher, D. H., editor, *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*, pages 125–133, Nashville, Tennessee, USA. Morgan Kaufmann Publishers. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.8350&rep=rep1&type=pdf.

Gain, A. and Shpitser, I. (2018). Structure learning under missing data. In Studený, M. and Kratochvíl, V., editors, *Proceedings of the Ninth International Conference on Probabilistic Graphical Models (PGM 2018)*, pages 121–132, Prague, Czech Republic. PMLR. https:

//proceedings.mlr.press/v72/gain18a/gain18a.pdf.

Geiger, D. (1987). The non-axiomatizability of dependencies in directed acyclic graphs. Technical Report R-83, Department of Computer Science, University of California, Los Angeles, USA. https://ftp.cs.ucla.edu/tech-report/198_-reports/870048.pdf.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524.

Glymour, M. M., Weuve, J., and Chen, J. T. (2008). Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: measurement, selection, and bias. *Neuropsychology Review*, 18(3):194–213.

Goldberger, A. S. (1991). *A Course in Econometrics*. Harvard University Press, Cambridge, Massachusetts, USA.

Granger, E., Watkins, T., Sergeant, J. C., and Lunt, M. (2020). A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Medical Research Methodology*, 20:132.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.

Greenland, S. (2010). Overthrowing the tyranny of null hypotheses hidden in causal diagrams. In Dechter, R., Geffner, H., and Halpern, J. Y., editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 365–382. College Publications. http://bayes.cs.ucla.edu/TRIBUTE/festschrift-complete.pdf.

Greenland, S., Daniel, R., and Pearce, N. (2016). Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International Journal of Epidemiology*, 45(2):565–575.

Greenland, S. and Pearce, N. (2015). Statistical foundations for model-based adjustments. *Annual Review of Public Health*, 36:89–108.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Guo, F. R. and Perković, E. (2020). Efficient least squares for estimating total effects under linearity and causal sufficiency. *arXiv preprint arXiv:2008.03481v2*.

Häggström, J., Persson, E., Waernbaum, I., and de Luna, X. (2015). CovSel: an R package for covariate selection when estimating average causal effects. *Journal of Statistical Software*, 68(1):1–20.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *Review of Economics and Statistics*, 86(1):73–76.

Hardt, J., Brendler, C., Greiser, K. H., Timmer, A., Seidler, A., Weikert, C., and Latza,

U., editors (2011). Directed acyclic graphs (DAGs) – basic concepts and application of an approach for causal analyses in epidemiology (special issue). *Das Gesundheitswesen*, 73(12):877–926. In German.

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(1):2409–2464.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection — a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391.

Henckel, L., Perković, E., and Maathuis, M. H. (2019). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435v1*.

Hernán, M. A. and Cole, S. R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, 170(8):959–962.

Hernán, M. A., Hernández-Díaz, S., Werler, M. M., and Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology*, 155(2):176–184.

Hernán, M. A., Hsu, J., and Healy, B. (2019). A second chance to get causal inference right: a classification of data science tasks. *CHANCE*, 32(1):42–49.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Hill, J. and Stuart, E. A. (2015). Causal inference: overview. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences*, pages 255–260. Elsevier, Oxford, UK, 2nd edition.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J. D., Lacerda, G., and Shimizu, S. (2008). Causal discovery of linear acyclic models with arbitrary distributions. In McAllester, D. A. and Myllymäki, P., editors, *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI-08)*, pages 282–289, Helsinki, Finland. AUAI Press, `https://arxiv.org/abs/1206.3260`.

Hünermund, P. and Bareinboim, E. (2019). Causal inference and data-fusion in econometrics. *arXiv preprint arXiv:1912.09104*.

Hurink, T. (2020). Liefert das O–Set effiziente Schätzer für das marginale kausale Odds-Ratio? [Does the O-set yield efficient estimators for the marginal causal odds ratio?] Unpublished Master's thesis, University of Bremen, Bremen, Germany. In German.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–1179.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Kim, Y. and Steiner, P. M. (2021a). Causal graphical views of fixed effects and random effects models. *British Journal of Mathematical and Statistical Psychology*, 74(2):165–183.

Kim, Y. and Steiner, P. M. (2021b). Gain scores revisited: a graphical models perspective. *Sociological Methods & Research*, 50(3):1353–1375.

Lash, T. L., VanderWeele, T. J., Haneuse, S., and Rothman, K. J., editors (2020). *Modern Epidemiology*. Wolters Kluwer, Philadelphia, Pennsylvania, USA, 4th edition.

Lauritzen, S. and Sadeghi, K. (2018). Unifying Markov properties for graphical models. *The Annals of Statistics*, 46(5):2251–2278.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, New York, USA.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed markov fields. *Networks*, 20(5):491–505.

Lauritzen, S. L. and Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348.

Li, A. and Beek, P. (2018). Bayesian network structure learning with side constraints. In Studený, M. and Kratochvíl, V., editors, *Proceedings of the Ninth International Conference on Probabilistic Graphical Models (PGM 2018)*, pages 225–236, Prague, Czech Republic. PMLR, `https://proceedings.mlr.press/v72/li18a/li18a.pdf`.

Li, C. and Fan, X. (2020). On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1489.

Li, J. and Wang, Z. J. (2009). Controlling the false discovery rate of the association/causality structure learned with the PC algorithm. *Journal of Machine Learning Research*, 10(2):475–514.

Li, L., Dennis Cook, R., and Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):285–299.

Liu, Q., Li, C., Wanga, V., and Shepherd, B. E. (2018). Covariate-adjusted Spearman's rank correlation with probability-scale residuals. *Biometrics*, 74(2):595–605.

Liu, W., Brookhart, M. A., Schneeweiss, S., Mi, X., and Setoguchi, S. (2012). Implications of M bias in epidemiologic studies: a simulation study. *American Journal of Epidemiology*, 176(10):938–948.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M., editors (2018). *Handbook of Graphical Models*. CRC Press, Boca Raton, Florida, USA.

Maathuis, M. H. and Colombo, D. (2015). A generalised back-door criterion. *The Annals of Statistics*, 43(3):1060–1088.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164.

Malinsky, D. and Spirtes, P. (2017). Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*, 88:371–384.

Mansournia, M. A., Hernán, M. A., and Greenland, S. (2013). Matched designs and causal diagrams. *International Journal of Epidemiology*, 42(3):860–869.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London, UK.

Meek, C. (1995a). Causal inference and causal explanation with background knowledge. In Besnard, P. and Hanks, S., editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410, Montreal, Quebec, Canada. Morgan Kaufmann Publishers, `https://arxiv.org/abs/1302.4972`.

Meek, C. (1995b). Strong completeness and faithfulness in bayesian networks. In Besnard, P. and Hanks, S., editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 411–419, Montreal, Quebec, Canada. Morgan Kaufmann Publishers, `https://arxiv.org/abs/1302.4973`.

Meek, C. (2014). Toward learning graphical and causal process models. In Mooij, J. M., Janzing, D., Peters, J., Claassen, T., and Hyttinen, A., editors, *Proceedings of the UAI 2014 Workshop on Causal Inference: Learning and Prediction*, pages 43–48, Quebec City, Quebec, Canada. `http://ceur-ws.org/Vol-1274/uai2014ci_paper8.pdf`.

Miettinen, O. S. and Cook, E. F. (1981). Confounding: essence and detection. *American Journal of Epidemiology*, 114(4):593–603.

Mogensen, S. W., Malinsky, D., and Hansen, N. R. (2018). Causal learning for partially observed stochastic dynamical systems. In Globerson, A. and Elidan, G., editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI-18)*, pages 350–360, Monterey, California, USA. AUAI Press, `https://auai.org/uai2018/proceedings/papers/142.pdf`.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.

Mooij, J. M. and Claassen, T. (2020). Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI-20)*, page ID: 481, Virtual Conference. `https://www.auai.org/uai2020/proceedings/481_main_paper.pdf`.

Mooij, J. M., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108.

Morabia, A. (2011). History of the modern epidemiological concept of confounding. *Journal of Epidemiology & Community Health*, 65(4):297–300.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press, Cambridge, UK, 2nd edition.

Murtaugh, P. A. (2014). In defense of p values. *Ecology*, 95(3):611–617.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11):1213–1222.

Nandy, P., Hauser, A., and Maathuis, M. H. (2018). High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183.

Nandy, P., Maathuis, M. H., and Richardson, T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674.

Neyman, J. (1923[1990]). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472. Translation by Dabrowska, D. M. and Speed, T. P.

Oates, C. J., Kasza, J., Simpson, J. A., and Forbes, A. B. (2017). Repair of partly misspecified causal diagrams. *Epidemiology*, 28(4):548–552.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Francisco, California, USA.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, 174(11):1223–1227.

Pearl, J. (2015). Comment on Ding and Miratrix: 'To adjust or not to adjust?'. *Journal of Causal Inference*, 3(1):59–60.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, USA.

Perković, E. (2020). Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI-20)*, pages 530–539, Virtual Conference. `https://www.auai.org/uai2020/proceedings/229_main_paper.pdf`.

Perković, E., Kalisch, M., and Maathuis, M. H. (2017). Interpreting and using CPDAGs with background knowledge. In Elidan, G., Kersting, K., and Ihler, A. T., editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI-17)*, page ID: 120, Sydney, Australia. AUAI Press. `https://auai.org/uai2017/proceedings/papers/120.pdf`.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference*. The MIT Press, Cambridge, Massachusetts.

Petersen, A. H., Osler, M., and Ekstrøm, C. T. (2021). Data-driven model building for life-course epidemiology. *American Journal of Epidemiology*, 190(9):1898–1907.

Pigeot, I. and Foraita, R. (2011). Invited commentary: Directed acyclic graphs – realization of a dream? *Das Gesundheitswesen*, 73(12):921–922. In German.

Pigeot, I., Sobotka, F., Kreiner, S., and Foraita, R. (2015). The uncertainty of a selected graphical model. *Journal of Applied Statistics*, 42(11):2335–2352.

Pressat-Laffouilhère, T., Jouffroy, R., Leguillou, A., Kerdelhue, G., Benichou, J., and Gilibert, A. (2021). Variable selection methods were poorly reported but rarely misused in major medical journals: literature review. *Journal of Clinical Epidemiology*, 139:12–19.

Quintana, R. (2020). The structure of academic achievement: searching for proximal mechanisms using causal discovery algorithms. *Sociological Methods & Research*, in press.

Raghu, V. K., Ramsey, J. D., Morris, A., Manatakis, D. V., Sprites, P., Chrysanthis, P. K., Glymour, C., and Benos, P. V. (2018). Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *International Journal of Data Science and Analytics*, 6(1):33–45.

Rao, P. (1971). Some notes on misspecification in multiple regressions. *The American Statistician*, 25(5):37–39.

Reisach, A. G., Seiler, C., and Weichwald, S. (2021). Beware of the simulated DAG! Varsortability in additive noise models. *arXiv preprint arXiv:2102.13647*.

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for the Statistics and the Social Sciences, University of Washington, Seattle, Washington, USA. `https://csss.uw.edu/files/working-papers/2013/wp128.pdf`.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period — application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12):1393–1512.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491–515.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1):27–42.

Rojas-Saunero, L. P., Hilal, S., Murray, E. J., Logan, R. W., Ikram, M. A., and Swanson, S. A. (2021). Hypothetical blood-pressure-lowering interventions and risk of stroke and dementia. *European Journal of Epidemiology*, 36(1):69–79.

Rosenbaum, P. R. (2002). *Observational Studies*. 2nd edition, Springer, New York, USA.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

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

Røysland, K. (2012). Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics*, 40(4):2162–2194.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (2009). Author's reply. *Statistics in Medicine*, 28(9):1420–1423.

Runge, J. (2018). Causal network reconstruction from time series: from theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310.

Runge, J. (2021). Necessary and sufficient conditions for optimal adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2102.10324v1*.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Rejoinder. *Journal of the American Statistical Association*, 94(448):1135–1146.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. (1998). The TETRAD

project: constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522.

Schomaker, M., Heumann, C., and Shalabh (2016). *Introduction to Statistics and Data Analysis*. Springer, Cham, Switzerland.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22.

Scutari, M. (2020). Bayesian network models for incomplete and dynamic data. *Statistica Neerlandica*, 74(3):397–419.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. (2014). Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39.

Shpitser, I. and Tchetgen Tchetgen, E. (2016). Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In Grünwald, P. and Spirtes, P., editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536, Catalina Island, California, USA. AUAI Press, `https://event.cwi.nl/uai2010/papers/UAI2010_0016.pdf`.

Shrier, I. (2008). Letter to the editor. *Statistics in Medicine*, 27(14):2740–2741.

Shrier, I. and Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, 8(1):70.

Sjölander, A. (2009). Letter to the editor. *Statistics in Medicine*, 28(9):1416–1420.

Smucler, E., Rotnitzky, A., and Robins, J. M. (2019). A unifying approach for doubly-robust l_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737v3*.

Smucler, E., Sapienza, F., and Rotnitzky, A. (2021). Efficient adjustment sets in causal graphical models with hidden variables. *Biometrics*, in press.

Sokolova, E., von Rhein, D., Naaijen, J., Groot, P., Claassen, T., Buitelaar, J., and Heskes, T. (2017). Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD. *International Journal of Data Science and Analytics*, 3(2):105–119.

Sonis, J. (1998). A closer look at confounding. *Family Medicine*, 30(8):584–588.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press, Cambridge, Massachusetts, USA, 2nd edition.

Staplin, N., Herrington, W. G., Judge, P. K., Reith, C. A., Haynes, R., Landray, M. J., Baigent, C., and Emberson, J. (2017). Use of causal diagrams to inform the design and interpretation of observational studies: An example from the study of heart and renal protection (SHARP). *Clinical Journal of the American Society of Nephrology*, 12(3):546–552.

Steck, H. and Jaakkola, T. (2003). Bias-corrected bootstrap and model uncertainty. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, pages 521–528, Vancouver, Canada. MIT Press, `https://proceedings.neurips.cc/paper/2003/file/2aaaddf27344ee54058548dc081c6541-Paper.pdf`.

Steck, H. and Tresp, V. (1999). Bayesian belief networks for data mining. In *Proceedings of the Workshop 'Data Mining und Data Warehousing als Grundlage moderner entscheidungsunterstützender Systeme'*, pages 145–154, Magdeburg, Germany. `https://www.dbs.ifi.lmu.de/~tresp/papers/SteckTresp.pdf`.

Strobl, E. V., Spirtes, P. L., and Visweswaran, S. (2019). Estimating and controlling the false discovery rate of the PC algorithm using edge-specific p-values. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–37.

Studenmund, A. H. (2014). *Using Econometrics. A Practical Guide*. Pearson Education, Harlow, UK, 6th edition.

Sun, G.-W., Shook, T. L., and Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*, 49(8):907–916.

Suttorp, M. M., Siegerink, B., Jager, K. J., Zoccali, C., and Dekker, F. W. (2015). Graphical presentation of confounding in directed acyclic graphs. *Nephrology Dialysis Transplantation*, 30(9):1418–1423.

Talbot, D. and Massamba, V. K. (2019). A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. *European Journal of Epidemiology*, 34(8):725–730.

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., and Ellison,

G. T. H. (2020). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*, 50(2):620–632.

Thoemmes, F. (2015). M-bias, butterfly bias, and butterfly bias with correlated causes — a comment on Ding and Miratrix (2015). *Journal of Causal Inference*, 3(2):253–258.

Tian, J., Paz, A., and Pearl, J. (1998). Finding minimal d-separators. Technical Report R-254, Department of Computer Science, University, University of California, Los Angeles. `https://ftp.cs.ucla.edu/pub/stat_ser/r254.pdf`.

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., and Zhang, K. (2019). Causal discovery in the presence of missing data. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, pages 1762–1770, Naha, Okinawa, Japan. PMLR, `https://proceedings.mlr.press/v89/tu19a/tu19a.pdf`.

Tu, R., Zhang, K., Ackermann, P., Bertilson, B. C., Glymour, C., Kjellström, H., and Zhang, C. (2020). Causal discovery in the presence of missing data. *arXiv preprint arXiv:1807.04010*.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC, Boca Raton, Florida, USA, 2nd edition.

van der Zander, B., Liskiewicz, M., and Textor, J. (2014). Constructing separators and adjustment sets in ancestral graphs. In Zhang, N. L. and Tian, J., editors, *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI-14)*, pages 907–916, Quebec City, Quebec, Canada. AUAI Press, `https://bioinformatics.bio.uu.nl/textor/uai14.pdf`.

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219.

VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413.

VanderWeele, T. J. and Shpitser, I. (2013). On the definition of a confounder. *Annals of Statistics*, 41(1):196–220.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1):7–30.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University, University of California, Los Angeles, California, USA. `https://ftp.cs.ucla.edu/pub/stat_ser/R150.pdf`.

Vowels, M. J., Camgoz, N. C., and Bowden, R. (2021). D'ya like DAGs? A survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582v2*.

Walter, S. and Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. *European Journal of Epidemiology*, 24(12):733–736.

Wang, P.-L. and Michailidis, G. (2019). Directed acyclic graph reconstruction leveraging prior partial ordering information. In Nicosia, G., Pardalos, P., Umeton, R., Giuffrida, G., and Sciacca, V., editors, *Proceedings of the International Conference on Machine Learning, Optimization, and Data Science (LOD 2019)*, pages 458–471, Siena, Italy. Springer, Cham, Switzerland.

Watson, D. S. and Wright, M. N. (2021). Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110(8):2107–2129.

Webster-Clark, M., Stürmer, T., Wang, T., Man, K., Marinac-Dabic, D., Rothman, K. J., Ellis, A. R., Gokhale, M., Lunt, M., Girman, C., and Glynn, R. J. (2021). Using propensity scores to estimate effects of treatment initiation decisions: state of the science. *Statistics in Medicine*, 40(7):1718–1735.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., and Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13(12):841–853.

Westreich, D. (2019). *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press, New York, USA.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Williamson, E. J., Aitken, Z., Lawrie, J., Dharmage, S. C., Burgess, J. A., and Forbes, A. B. (2014). Introduction to causal diagrams for confounder selection. *Respirology*, 19(3):303–311.

Witte, J. (2021). tpc: temporal PC-algorithm. `https://github.com/bips-hb/tpc`.

Witte, J. and Didelez, V. (2018). Exploring the causal structure in cohort data using generalised IDA. *Rostock Retreat on Causality, unpublished conference poster*. Available from the author upon request.

Witte, J. and Didelez, V. (2019). Covariate selection strategies for causal inference: classification and comparison. *Biometrical Journal*, 61(5):1270–1289.

Witte, J., Foraita, R., and Didelez, V. (2021). Multiple imputation and test-wise deletion for causal discovery with incomplete cohort data. *arXiv preprint arXiv:2108.13331*.

Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V. (2020). On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7):557–580.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215.

Wyss, R., Fireman, B., Rassen, J. A., and Schneeweiss, S. (2018). Erratum: High-

dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 29(6):e63–e64.

Zhang, J. (2008a). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16–17):1873–1896.

Zhang, J. and Spirtes, P. (2003). Strong faithfulness and uniform consistency in causal inference. In Kjærulff, U. and Meek, C., editors, *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 632–639, Acapulco, Mexico. Morgan Kaufmann Publishers, https://dl.acm.org/doi/pdf/10.5555/2100584.2100661.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In Cozman, F. and Pfeffer, A., editors, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 804–813, Barcelona, Spain. AUAI Press, https://arxiv.org/abs/1202.3775.

Zhang, Z. (2008b). Estimating a marginal causal odds ratio subject to confounding. *Communications in Statistics—Theory and Methods*, 38(3):309–321.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 9472–9482. Curran Associates, https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

List of Abbreviations

ACE‡	average treatment effect
ADMG	acyclic directed mixed graph, see Section 2.5
AIC	Akaike's information criterion
BIC	Bayesian information criterion
CG*	Conditional Gaussian (distribution or test)
CIE‡	change-in-estimate
CIE-MSE‡	the change-in-estimate variant by Greenland et al. (2016)
CovSel	the covariate procedure by de Luna et al. (2011), see Section 3.4.6
CPDAG	completed partially directed acyclic graph, see Section 2.4
DAG	directed acyclic graph, see Section 2.1
EHS‡	the algorithm by Entner, Hoyer and Spirtes (2013)
FCI	fast causal inference (causal discovery algorithm), see Section 2.6
FCS‡	the focussed confounder selection method by Vansteelandt et al. (2012)
GES	greedy equivalence search, see Section 2.6
HMP19#	Henckel et al. (2019)
IDA	Intervention Calculus When the DAG is Absent, see Section 4.1.2
LMPC-stable	lists-majority-rule PC-stable (causal discovery algorithm), see Section 2.6
maxPDAG#	see MPDAG
MAR*	missing at random
MCAR*	missing completely at random
MCOR‡	marginal causal odds ratio
MICE*	multiple imputation by chained equations
MNAR*	missing not at random
MPDAG	maximally oriented partially directed acyclic graph, see Section 2.4
MSE	mean squared error
O-set	optimal adjustment set
OLS	ordinary least squares

PAG	partial ancestral graph, see Section 2.6.2
PC	Peter&Clark (causal discovery algorithm), see Section 2.6
PS‡	propensity score
RS20#	Rotnitzky and Smucler (2020)
SWIG	single world intervention graph, see Section 2.3.2
tPC	temporal / time-ordered / tiers / (partial) topological ordering PC-algorithm (causal discovery algorithm), see Section 5.1

‡ These abbreviations are used in Paper 1 (Section 3.5) only.
# These abbreviations are used in Paper 2 (Section 4.5) only.
* These abbreviations are used in Paper 3 (Section 5.3) only.

List of Symbols

$\perp_{\mathcal{G}}$	d-separation or m-separation in graph $\mathcal{G}$, see Sections 2.1, 2.4 and 2.5
$\not\perp_{\mathcal{G}}$	absence of d-separation or m-separation in graph $\mathcal{G}$
$\perp\!\!\!\perp$	(conditional) independence in a probability distribution, see Section 2.2
$\not\perp\!\!\!\perp$	(conditional) dependence in a probability distribution
$\beta_{y.\mathbf{xz}}$	intercept in a regression of Y on $\mathbf{X}$ and $\mathbf{Z}$
$\beta_{y\mathbf{x}.\mathbf{z}}$	(vector of) coefficient(s) of $\mathbf{X}$ in a regression of Y on $\mathbf{X}$ and $\mathbf{Z}$
$\hat{\mu}_0(\mathbf{Z})$	the estimated mean outcome under $do(X = 0)$, a function of random variables $\mathbf{Z}$
$\hat{\mu}_1(\mathbf{Z})$	the estimated mean outcome under $do(X = 1)$, a function of random variables $\mathbf{Z}$
τ	average causal effect, see Section 3.1
$\text{an}(\mathbf{A})$	the set of ancestors of node set $\mathbf{A}$, see Section 2.1
$a.var(\cdot)$	asymptotic variance, see Section 4.3
$\text{ch}(\mathbf{A})$	the set of children of node set $\mathbf{A}$, see Section 2.1
$\text{cn}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$	the causal nodes with respect to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$, see Section 3.2
$\mathcal{D}$	a (causal) DAG
$\text{de}(\mathbf{A})$	the set of descendants of node set $\mathbf{A}$, see Section 2.1
$\hat{e}(\mathbf{Z})$	the estimated propensity score, a function of random variables $\mathbf{Z}$
$f(\cdot)$	probability density, see Section 2.2
$do(\cdot)$	Pearl's do-operator, see Section 2.3.1
$\text{forb}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$	the forbidden set with respect to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$, see Sections 3.2 and 4.1.1
$\mathcal{G}$	a (causal) graph
$\mathcal{G}^{\mathbf{X}Y}$	the forbidden projection of $\mathcal{G}$ with respect to $(\mathbf{X}, Y)$, see Sections 4.2 and 4.5
$\text{nde}(\mathbf{A})$	the set of non-descendants of node set $\mathbf{A}$, see Section 2.1
$\mathbf{O}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$	the optimal adjustment set ($\mathbf{O}$-set) with respect to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$, see Sections 4.2 and 4.5

$\text{pa}(\mathbf{A})$	the set of parents of node set $\mathbf{A}$, see Section 2.1
$\text{possan}(\mathbf{A})$	the set of possible ancestors of node set $\mathbf{A}$, see Section 2.1
$\text{possde}(\mathbf{A})$	the set of possible descendants of node set $\mathbf{A}$, see Section 2.1
$\text{posscn}(\mathbf{X}, \mathbf{Y}, \mathcal{G})$	the possibly causal nodes with respect to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$, see Section 4.1.1
R_V	the missingness indicator of variable V, see Paper 3 (Section 5.3)
$R^{\mathbf{A}}$	the missingness indicator for the set of variables $\mathbf{A}$, see Paper 3 (Section 5.3)
$\mathbf{R}(\mathbf{V})$	the set of missingness indicators of the variables in $\mathbf{V}$, see Paper 3 (Section 5.3)
$\text{sib}(\mathbf{A})$	the set of siblings of node set $\mathbf{A}$, see Section 2.1
$\mathbf{S}$	usually a conditioning set
T	the treatment in Paper 1 (Section 3.5)
$t(V)$	the tier of node V
$\mathbf{V}$	usually the set of nodes in a graph
$\mathbf{W}$	usually the set of observed covariates
$\mathbf{W}^*$	the selected adjustment set
X or $\mathbf{X}$	usually the treatment; in Paper 1 (Section 3.5), $\mathbf{X}$ denotes the covariates
$\mathcal{X}$	the state space of $\mathbf{X}$
Y or $\mathbf{Y}$	usually the outcome
$\mathbf{Z}$	usually a (valid) adjustment set