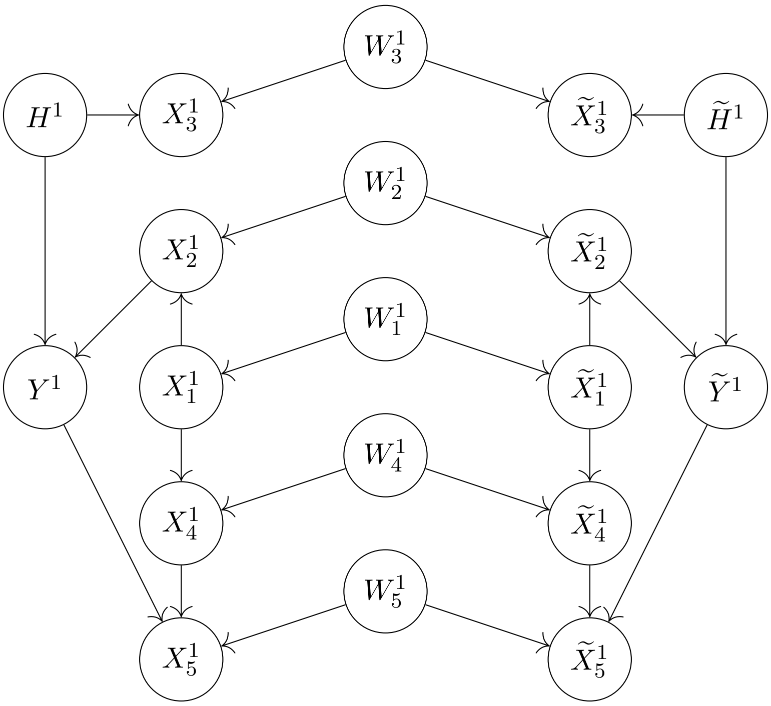
Causal Discovery from Interventional Data

A Bachelor of Science thesis by
Adam Gorm Hoffmann



Supervised by Prof. Jonas Peters
Co-supervised by Assist. Prof. Niklas Pfister
Department of Mathematical Sciences
University of Copenhagen, Denmark
August 2021

UNIVERSITY OF
COPENHAGEN

Abstract

We consider the task of learning the causes of a response in three closely related problems, all related to the scenario of two separate sets of interventional experiments with hidden variables and unknown intervention targets. In the first problem (which we haven't seen studied before), the covariates are observed in the first set of experiments and the response is observed in the second set of experiments, yielding unpaired data. We present a novel method, POLS, for its solution. In the second problem both the response and the covariates are observed in both sets of experiments; in the third problem we only conduct one set of experiments and then, in order to be able to use the methods from the first two problems (which require a second data set), we permute the rows to emulate data from a second set of experiments. We present another novel method, DPOLS, for the last two problems, and give a proof that it will select the correct parents asymptotically in a specific case. We give a strengthened version of Reichenbach's Common Cause Principle as motivation for the methods and investigate their performance through large scale simulation experiments. Our results show that both POLS and DPOLS beat the baseline methods. The results also indicate that DPOLS finds the correct parents asymptotically in many cases, including on permuted data, and that it is even useful for finding extra ancestors after having selected all parents.

Contributions from others

Jonas Peters and Niklas Pfister, my supervisors from the University of Copenhagen, originated the idea of two separate data sets $(\mathbf{X}, \mathbf{Y})$ and $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ linked only through the shift interventions W, the idea that this situation can be emulated by permuting the rows of a single dataset, as well as the specific methods POLS and DPOLS, except for the names which are my creation. The argument showing that $\beta^{\mathrm{DPOLS}} = \beta$, when $\operatorname{cov}(W)$ is invertible, is from an unpublished set of notes by Niklas Pfister.

Code availability

Scripts for running and analyzing simulations are available on GitHub at https://github.com/adamgorm/bsc-simulations.

Presented for the degree of *Bachelor of Science in Mathematics*.

Throughout German literature of the last ten years we find "to condition" almost everywhere used in place of "to cause" or "to effect." Since it is more abstract and indefinite it says less than it implies, and consequently leaves a little back door open to please those whose secret consciousness of their own incapacity inspires them with a continual fear of all definite expressions.

ARTHUR SCHOPENHAUER

Contents

1 Introduction

Discovering causal relationships is a fundamental goal of science. While randomized controlled experiments have long been used to distinguish causation from correlation, researchers have since proposed new methods for causal learning in non-randomized settings. This includes independence-based methods (*e.g.*, the PC algorithm; Spirtes et al., 2000) that, under the assumption of faithfulness, use the conditional independence statements of the observational distribution to find the Markov equivalence class of the directed acyclic graph entailed by the underlying structural causal model. If one is primarily interested in the causes of a specific variable, then the Markov equivalence class may disappoint, since many orientations of the edges can be possible. ICP, a more recent method proposed by Peters et al. (2016), will find a subset of the direct causes of a response given data from different environments, *e.g.,* different interventional settings.

Sometimes it may not be possible to observe the covariates and the response in the same experiment; perhaps the very act of measuring the covariates physically destroys the response. The researcher may then choose to conduct all experiments twice, first observing the covariates, then the response. If this is done with a series of experiments, then the only connection between the response and covariates from the separate experiments is the experimental setup itself, since any specific pairing of the observations will be arbitrary. Is any information about cause and effect left in this unpaired data set? When each experimental setup consists of intervening on a single known covariate, it is clear that a distribution shift of the response in a given experiment means that the intervention target is a cause of the response. It is, however, less clear whether enough information is left if we don't know the intervention targets. To the best of our knowledge, no one has studied this setting yet.

In this thesis, we present the method POLS for learning the causes of a response from unpaired data with unknown intervention targets in the presence of hidden variables. We present a strong version of Reichenbach's Common Cause Principle as a heuristic argument that the broken link between response and covariate can be a strength, rather than a weakness, since it removes hidden confounding. This motivates the question of whether POLS is useful, even when paired data is available? It is, at the very least, usable, since permuting the response vector within experimental setups will emulate the situation with unpaired data; by using the same permutation on each column of the matrix of covariate observations, we can emulate an entire data set from a separate set of experiments, leaving us with two complete data sets. We investigate whether POLS is useful on permuted data, and present another method, DPOLS, specifically for this setting. Niklas Pfister has proved that, given complete knowledge of the observational distributions, DPOLS can correctly identify the direct causes of a response, when applied on data from two separate experiments.

One of our main contributions is a series of large scale simulation experiments comparing POLS, DPOLS, and various baseline methods, both in the setting with data from two separate sets of experiments, and in the setting where a single data set is permuted.

2 Causal models

2.1 Why do we need more than statistical models?

Observational distributions give a wealth of useful knowledge. If we know the joint distribution of V (infected with corona virus), C (coughing), and F (went to a party last week), then we can find the probability $P(V \mid C = 1, F = 1)$ of a person being infected, given that they cough and went to a party. If, however, we start wondering "would fewer of us get infected if we eat more cough drops?" or "given that John tested positive on a PCR test today, is coughing, and went to a party last week, what would have happened if he had stayed at home and solved problems in Rudin instead?" then the observational distribution is no longer enough, since there is a large difference between observing $C = 0$ for a member of the population, and setting (or *doing*) $C := 0$, which in effect changes the population and hence the distribution. Unfortunately, classical statistical methodology is mostly concerned with inferring parameters of a single observational distribution or using it for predictions. According to Fisher (1922, Section 2), "briefly, and in its most concrete form, the object of statistical methods is the reduction of data [...] by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample". Different authors have given a wide range of definitions of the goals of statistics (see, *e.g.*, Barnett, 1999, Section 1.1). We would argue that a central theme is the idea of a statistical model.

> A *frequentist statistical model* is a family $\mathcal{P}$ of probability measures on a measurable space $(\Omega, \mathcal{K})$ (see, *e.g.*, Lehmann and Casella, 1998; Shao, 2003). A *Bayesian statistical model* furthermore assumes that $\mathcal{P}$ is a parameterized family with densities, and includes a prior distribution on the parameter space (see, *e.g.*, Gelman et al., 2014; Lauritzen, 2021).

Idealists then assume that there is a true $P \in \mathcal{P}$, while others hope for a $P \in \mathcal{P}$ that gives reasonable predictions, but, arguably, most would agree that "[t]he task of the statistician is to say something sensible about P, based on the observation x" (Lauritzen, 2021, Section 1.1). However, saying something sensible about the observational distribution is no longer enough when we start asking causal questions. To answer those, we can use many different, yet connected, distributions. We can use interventional distributions to reason about what would happen if everyone stops coughing and counterfactual distributions to answer what would likely have happened if John had stayed at home two weeks ago instead of going to that party. All of this is included with a *structural causal model* (*e.g.,* Pearl, 2009; Peters et al., 2017), which will be introduced in the following sections.

2.2 Directed graphs

Before defining structural causal models we introduce graph terminology. The meanings of most terms are obvious, but consult the formal definitions below when in doubt. The definitions can be found, in a slightly different form, in, *e.g.*, Lauritzen (1996, Section 2.1.1).

Definition 1 (Graph terms)**.** A directed *graph* is a pair $\mathcal{G} = (V, E)$ where V is a set of *vertices* (or *nodes*) and $E \subseteq V^2$ is a set of *edges*. A *path* is a tuple $(\gamma_1, e_1, \ldots, e_{k-1}, \gamma_k)$ where $\gamma_i \in V$ for all i, and $e_j \in \{(\gamma_j, \gamma_{j+1}), (\gamma_{j+1}, \gamma_j)\} \cap E$ for all j, and $\gamma_i \neq \gamma_j$ when $i \neq j$. We often write

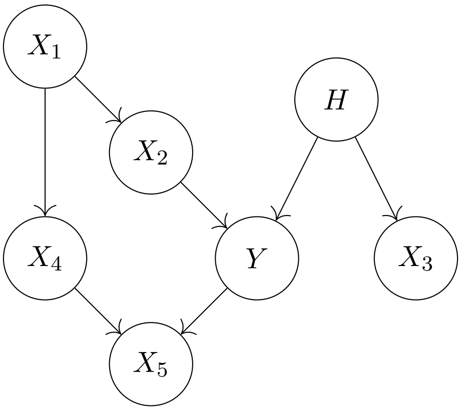


Figure 1: Example of a DAG with nodes $V = \{X_1, X_2, X_3, X_4, X_5, Y, H\}$ and edges given by the arrows in the picture. Here we have $\mathrm{pa}_Y = \{X_2, H\}$, $\mathrm{anc}_Y = \{X_1, X_2, H\}$, $\mathrm{ch}_Y = \{X_5\} = \mathrm{de}_Y$, and $\mathrm{nd}_Y = \{X_1, X_2, X_3, X_4, H\}$.

$\alpha \to \beta$ to indicate an edge $(\alpha, \beta) \in E$ and we write, *e.g.*, $\alpha \leftarrow \beta \to \gamma$ to indicate a path $(\alpha, (\beta, \alpha), \beta, (\beta, \gamma), \gamma)$. On a given path, the node β is said to be a collider if $\alpha \to \beta \leftarrow \gamma$ is in the path for some α and γ. A *directed path* from α to β is a path $(\alpha = \gamma_1, e_1, \ldots, e_{k-1}, \gamma_k = \beta)$ where $e_j = (\gamma_j, \gamma_{j+1}) \in E$ for all j. A *directed cycle* is a directed path, except that the first node and the last node are the same $(\alpha = \gamma_1, e_1, \ldots, e_{k-1}, \gamma_k = \alpha)$. A directed graph without directed cycles is called a *directed acyclic graph* or *DAG*. In DAGs a path is completely determined by the sequence of nodes, since there can only be one edge between any two nodes. The node α is a *parent* of β if $(\alpha, \beta) \in E$, an *ancestor* of β if there is a directed path from α to β, a *descendant* of β if β is an ancestor of α, and a non-descendant of β if it is not a descendant of β nor equal to β. A node without any parents is called a *source node*. The sets of all parents, ancestors, children, descendants, respectively non-descendants of β are denoted pa_β, anc_β, ch_β, de_β, respectively nd_β, or $\mathrm{pa}(\beta)$, …, $\mathrm{nd}(\beta)$; sometimes the superscript $\mathcal{G}$ is added to avoid confusion. When the nodes are random variables $X_1, \ldots, X_d$ we sometimes use uppercase PA_i to denote the parents of X_i, and lowercase pa_i to denote a specific outcome of PA_i (and similarly for descendants, ancestors, and non-descendants). See Fig. 1 for an example of a DAG.

2.3 Structural Causal Models

We are now ready to introduce structural causal models. All definitions and propositions in this section can be found, sometimes in a slightly different form, in Peters et al. (2017), Bongers et al. (2021), or Pearl (2009).

A structural causal model, or SCM, $\mathcal{C}$ over the d variables $X = (X_1, \ldots, X_d)$ consists of d structural assignments (also known as structural equations in, *e.g.*, Pearl, 2009)

$$\begin{aligned} X_1 &:= f_1(\mathrm{PA}_1, N_1) \\ &\vdots \\ X_d &:= f_d(\mathrm{PA}_d, N_d). \end{aligned}$$

and the distribution P_N of the jointly independent exogenous variables $N_1, \ldots, N_d$. The exogenous variables account for incomplete knowledge of the system or inherent randomness (what

Spirtes et al., 2000, call pseudo indeterminism respectively indeterminism). $\mathrm{PA}_i \subseteq X$ [1] is called the set of *causal parents* or *direct causes* of X_i, and play a central role in this thesis. There are two levels of arbitrariness concerning PA_i, the first of which we will simply have to accept, while the second will be addressed in Proposition 3 and Definition 4 below, which may be skimmed on a first reading.

First, the term *direct cause* insinuates that there is, in an absolute sense, a direct link between the cause and the effect with nothing in-between. However, the term should be understood relative to the given model, and, in particular, the set of observed variables, as discussed in Spirtes et al. (2000, Section 3.2). While L ("Likes movies?") is a direct cause of C ("Went to the cinema?") when we only observe (L, C), if we also observe G ("Got tickets for the cinema?") we would find that L causes G which causes C; whether L is a direct cause of C depends on the observed variables. Similarly, there is not a canonical choice of demarcation of what constitutes a variable. For instance, "Likes movies?" could be split into "Likes silent films?" and "Likes talkies?", or further into "Likes Charlie Chaplin movies?" and "Likes Buster Keaton movies?" and so on; perhaps, someday, even into specific states of the brain. Thus, when we say *direct cause* it is not to be understood as some absolute metaphysical concept, but rather as a property of a certain model of a restricted part of reality.

Second, consider the structural assignment $X_3 := 2X_1 + N_3$ where $\mathrm{PA}_3 = X_1$. The structural assignment $X_3 := 2X_1 + 0 \cdot X_2 + N_3$ leads to the exact same values of X_3 for given (X_1, X_2, N_3), but here $\mathrm{PA}_3 = (X_1, X_2)$. Our intended interpretation of PA_3 is that it is the set of observed variables which directly affect the value of X_3, so we clearly want to have $\mathrm{PA}_3 = X_1$ rather than $\mathrm{PA}_3 = (X_1, X_2)$. Luckily, we will see below that there is a unique representation of any SCM, where each function depends on all of its arguments. First we give a formal definition of SCMs.

Definition 2. A *structural causal model*, or *SCM*, over d variables is given by a triple

$$\mathcal{C} = ((\mathcal{N}, \mathcal{K}, P_N), (\mathcal{X}, \mathcal{F}), f),$$

where

- $(\mathcal{N}, \mathcal{K}, P_N) = \left(\bigtimes_{i=1}^d \mathcal{N}_i, \bigotimes_{i=1}^d \mathcal{K}_i, \bigotimes_{i=1}^d P_{N_i}\right)$ is a measure space representing the sample space and distribution of the exogenous variables.
- $(\mathcal{X}, \mathcal{F}) = \left(\bigtimes_{i=1}^d \mathcal{X}_i, \bigotimes_{i=1}^d \mathcal{F}_i\right)$ is a measurable space representing the sample space of the endogenous variables.
- $f = (f_1, \ldots, f_d)$ is a measurable function representing the causal mechanism, where

$$\text{for all } i \in \{1, \ldots, d\} \text{ there is a set } \mathcal{I}_i \subseteq \{1, \ldots, d\} \text{ such that } \quad f_i : \left(\bigtimes_{i \in \mathcal{I}_i} \mathcal{X}_i\right) \times \mathcal{N}_i \to \mathcal{X}_i.$$

In these terms, the choice of f is arbitrary, since $\mathcal{I}_i$ could always be the entire set $\{1, \ldots, d\}$. The proposition and definition below, which is a modified version of Peters et al. (2017, Remark 6.6), show that there is a natural choice, where each f_i depends on as few arguments as possible.

[1] We adhere to the common abuse of notation of using, *e.g.*, X to denote both a vector and a set of random variables as circumstances see fit.

This is done in some detail, since our thesis concerns inferring parents and ancestors, wherefore it is important for the terms to have a clear meaning.

Proposition 3. *Let $\mathcal{C} = ((\mathcal{N}, \mathcal{K}, P_N), (\mathcal{X}, \mathcal{F}), f)$ be an SCM over d variables. There exist unique sets $\mathrm{pa}_i^{\mathcal{C}} \subseteq \{1, \ldots, d\}$ for $i \in \{1, \ldots, d\}$ that satisfy the following:*

(i) For all $i \in \{1, \ldots, d\}$ the function f_i only depends on $\mathrm{pa}_i^{\mathcal{C}}$, i.e., there are measurable functions $g_i : \left(\bigtimes_{i \in \mathrm{pa}_i^{\mathcal{C}}} \mathcal{X}_i\right) \times \mathcal{N}_i \to \mathcal{X}_i$ such that

$$f_i(x_{\mathcal{I}_i}, n_i) = g_i(x_{\mathrm{pa}_i^{\mathcal{C}}}, n_i) \quad \textit{for all } x \in \mathcal{X} \textit{ and } P_N\textit{-almost all } n_i \in \mathcal{N}_i.$$

(ii) The functions (g_i) are unique in the sense that if functions (h_i) satisfy (i) then

$$h_i(x_{\mathrm{pa}_i^{\mathcal{C}}}, n_i) = g_i(x_{\mathrm{pa}_i^{\mathcal{C}}}, n_i) \quad \textit{for all } x \in \mathcal{X} \textit{ and } P_N\textit{-almost all } n_i \in \mathcal{N}_i.$$

(iii) The sets $\mathrm{pa}_i^{\mathcal{C}}$ are chosen as small as possible in the sense that there is no $i \in \{1, \ldots, d\}$, proper subset $\widetilde{\mathrm{pa}}_i^{\mathcal{C}} \subsetneq \mathrm{pa}_i^{\mathcal{C}}$, and measurable function $\widetilde{g}_i : \left(\bigtimes_{i \in \widetilde{\mathrm{pa}}_i^{\mathcal{C}}} \mathcal{X}_i\right) \times \mathcal{N}_i \to \mathcal{X}_i$ such that

$$g_i(x_{\mathrm{pa}_i^{\mathcal{C}}}, n_i) = \widetilde{g}_i(x_{\widetilde{\mathrm{pa}}_i^{\mathcal{C}}}, n_i) \quad \textit{for all } x \in \mathcal{X} \textit{ and } P_N\textit{-almost all } n_i \in \mathcal{N}_i.$$

Proof. Existence follows by repeatedly removing arguments that some f_i doesn't depend on, as follows.

1. First, let $\mathrm{pa}_i^{\mathcal{C}} := \mathcal{I}_i$ and $g_i := f_i$ for all $i \in \{1, \ldots, d\}$.

2. If, for some $i \in \{1, \ldots, d\}$ and $p \in \mathrm{pa}_i^{\mathcal{C}}$,

$$g_i(x_{\mathrm{pa}_i^{\mathcal{C}} \setminus \{p\}}, x_p, n_i) = g_i(x_{\mathrm{pa}_i^{\mathcal{C}} \setminus \{p\}}, x'_p, n_i)$$

for all $x \in \mathcal{X}, x'_p \in \mathcal{X}_p$ and P_N-almost all $n_i \in \mathcal{N}_i$ then assign $\mathrm{pa}_i^{\mathcal{C}} \leftarrow \mathrm{pa}_i^{\mathcal{C}} \setminus \{p\}$ and let g_i be redefined as a function $g_i : \left(\bigtimes_{i \in \mathrm{pa}_i^{\mathcal{C}}} \mathcal{X}_i\right) \times \mathcal{N}_i \to \mathcal{X}_i$ given by

$$g_i(x_{\mathrm{pa}_i^{\mathcal{C}}}, n_i) := f_i(x_{\mathcal{I}_i}, n_i), \quad \text{for all } x \in \mathcal{X},$$

and repeat 2.

Otherwise go on to 3.

3. (i) is satisfied by construction. (iii) must be satisfied as well; otherwise 2 would not have terminated yet. (ii) is satisfied, because if (h_i) also satisfy (i) then

$$h_i(x_{\mathrm{pa}_i^{\mathcal{C}}}, n_i) = f_i(x_{\mathcal{I}_i}, n_i) = g_i(x_{\mathrm{pa}_i^{\mathcal{C}}}, n_i) \quad \text{for all } x \in \mathcal{X} \text{ and } P_N\text{-almost all } n_i \in \mathcal{N}_i.$$

This proves existence.

We now prove uniqueness. Let pa_i and $\widetilde{\mathrm{pa}}_i$ be two sets satisfying (i), (ii) and (iii) with corresponding functions (g_i) and $(\widetilde{g}_i)$. Assume that there is some $p \in \widetilde{\mathrm{pa}}_i \setminus \mathrm{pa}_i$. Since

$$\widetilde{g}_i(x_{\widetilde{\mathrm{pa}}_i}, n_i) = f_i(x_{\mathcal{I}_i}, n_i) = g_i(x_{\mathrm{pa}_i}, n_i) \quad \text{for all } x \in \mathcal{X} \text{ and } P_N\text{-almost all } n_i \in \mathcal{N}_i$$

we have that g_i doesn't depend on p. But then, using step 2 above, p could be removed from $\widetilde{\mathrm{pa}}_i$, so $\widetilde{\mathrm{pa}}_i$ and $(\widetilde{g}_i)$ cannot satisfy (iii). Hence there can't be a $p \in \widetilde{\mathrm{pa}}_i \setminus \mathrm{pa}_i$, so $\widetilde{\mathrm{pa}}_i \subseteq \mathrm{pa}_i$, and by symmetry $\mathrm{pa}_i \subseteq \widetilde{\mathrm{pa}}_i$. ∎

Definition 4. The set $\mathrm{pa}_i^{\mathcal{C}}$ is called the set of *causal parents* of i. The representation of $\mathcal{C}$ given by the functions (g_i) in Proposition 3 is called *structurally minimal.* From now on all SCMs discussed in this thesis are assumed to be structurally minimal.

Now that the term *causal parents* has a clear meaning, we introduce the idea of an *entailed graph*, which, as we will see in Section 2.5, is a useful tool for reasoning about conditional independence statements related to an SCM.

Definition 5. The *causal graph $\mathcal{G}$ entailed* by the SCM $\mathcal{C}$ is the directed graph over $\{1, \ldots, d\}$ with edges satisfying $\mathrm{pa}_i^{\mathcal{G}} = \mathrm{pa}_i^{\mathcal{C}}$. For $i \in \{1, \ldots, d\}$ the *causal ancestors, causal descendants*, respectively *causal non-descendants* of i refer to the ancestors, descendants, respectively non-descendants of i in the causal graph.

So far we have made no mention of the random variables (X, N) satisfying the structural assignments of an SCM. The existence and uniqueness of such a *solution* will be dealt with now.

Definition 6. Random variables $X = (X_1, \ldots, X_d)$ on $(\mathcal{X}, \mathcal{F})$ and $N = (N_1, \ldots, N_d)$ on $(\mathcal{N}, \mathcal{K})$ are called a *solution* of the SCM $\mathcal{C} = ((\mathcal{N}, \mathcal{K}, P_N), (\mathcal{X}, \mathcal{F}), f)$ if N has distribution P_N and

$$X_i = f_i(X_{\mathrm{pa}_i^{\mathcal{C}}}, N_i) \quad \text{for all } i \in \{1, \ldots, d\}, \quad \text{a.s.}$$

Given a solution, we often use the variables $X_1, \ldots, X_d$ as nodes in the causal graph instead of the indices $1, \ldots, d$, and let PA_i denote $X_{\mathrm{pa}_i^{\mathcal{C}}}$.

It is possible that there exists no solution, or that there exist solutions with any distribution on $\mathbb{R}$ (Bongers et al., 2021, Example 2.4). However, these problems disappear when the causal graph is a DAG due to the implied existence of a causal order.

Definition 7. A permutation $\pi : \{1, \ldots, d\} \to \{1, \ldots, d\}$ is a *causal order* if all $i, j \in \{1, \ldots, d\}$ with $i \in \mathrm{anc}_j$ satisfy $\pi(i) < \pi(j)$.

Proposition 8. *If the entailed graph $\mathcal{G}$ of $\mathcal{C}$ is a DAG, then there exists a causal order.*

Proof. Since there is a finite number of nodes, and no directed cycles, there must be a source node i_1. Let $\pi(i_1) := 1$ and let $\mathcal{G}_2$ be the graph where i_1 (and all edges involving i_1) is removed from $\mathcal{G}$, but everything else stays the same. This must again be a DAG, so there must again be a source node i_2. Let $\pi(i_2) := 2$. Remove i_2 and continue until there are no nodes left. ∎

Proposition 9. *If the causal graph entailed by the SCM $\mathcal{C}$ is a DAG, then there exists a solution (X, N) and any other solution (X', N') has the same distribution as (X, N).*

Proof. Let π be a causal order and let N have distribution P_N. By substituting into the structural assignments in causal order, we can uniquely determine X from N as follows. We must set $X_{\pi^{-1}(1)} := f_{\pi^{-1}(1)}(N_{\pi^{-1}(1)})$ since it is a source node. $X_{\pi^{-1}(2)}$ can now be determined from $N_{\pi^{-1}(2)}$ and possibly $X_{\pi^{-1}(1)}$. Then $X_{\pi^{-1}(3)}$ can be determined from $N_{\pi^{-1}(3)}$ and possibly $X_{\pi^{-1}(1)}$

and $X_{\pi^{-1}(2)}$. This can be continued until $X_{\pi^{-1}(d)}$ is determined. By construction (X, N) is a solution. X' must be obtained by the same transformation of N', so since $N' \overset{\mathcal{D}}{=} N$ it follows that $X' \overset{\mathcal{D}}{=} X$. ∎

All the SCMs that we consider in the following sections are assumed to be acyclic. For more information on the cyclic case, see Bongers et al. (2021). In this thesis we will include some unobserved endogenous variables in SCMs. This is often done implicitly, as described in the following definition.

Definition 10. A *partially observed SCM* is a pair $(\mathcal{C}, O)$ where $\mathcal{C}$ is an SCM and $O \subseteq \{1, \dots, d\}$ is the set of observed variables. If the letter "H" is used to denote a set of variables in an SCM over variables V, then it implicitly means that the SCM is a partially observed SCM with $O = V \setminus H$. We refer to H as the set of hidden variables.

Example 11. As an example, consider the SCM $\mathcal{C}$ over the 7 variables $X_1, X_2, X_3, X_4, X_5, Y, H$, given by the structural assignments

$$\begin{aligned}
X_1 &:= N_{X_1} \\
X_2 &:= X_1 + N_{X_2} \\
H &:= N_H \\
X_3 &:= H + N_{X_3} \\
Y &:= X_2 + H + N_Y \\
X_4 &:= X_1 + N_{X_4} \\
X_5 &:= X_4 + Y + N_{X_5}
\end{aligned}$$

where all noise variables are independent with distributions $N_{X_1}, N_{X_2}, N_{X_3}, N_{X_4}, N_{X_5} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $N_H \sim \mathcal{N}(0, \tau)$. This is a *linear Gaussian SCM* and the causal graph entailed by this SCM is exactly the DAG from Fig. 1. Since the letter "H" is used, we know that it is a partially observed SCM with $O = \{X_1, X_2, X_3, X_4, X_5, Y\}$.

2.4 Interventions

Interventional distributions describe a situation where many aspects of the causal structure (that is, the SCM) remain the same, while some change (Peters et al., 2017).

Definition 12. Let $\mathcal{C}$ be an SCM. An *intervention* corresponds to changing the structural assignments f or the noise distribution P_N yielding a new modified SCM $\hat{\mathcal{C}}$. The modified SCM $\hat{\mathcal{C}}$ must be acyclic as well. Interventions are often written in *do*-notation, denoting, *e.g.*, the intervention changing f_2 to $(x_5, n_2) \mapsto x_5^2 + 3n_2$ by $do(X_2 := X_5^2 + 3N_2)$, so the modified SCM is denoted by $\mathcal{C}; do(X_2 := X_5^2 + 3N_2)$, and the observational distribution entailed by the modified SCM by $P^{\mathcal{C}; do(X_2 := X_5^2 + 3N_2)}$.

Example 13. Consider again the SCM from Example 11. Let $W_1, W_2, W_3, W_4, W_5 \overset{\text{iid}}{\sim} \mathcal{N}(0, \rho)$ such that all W's and N's are jointly independent. We can do a *shift intervention* by, for all i, adding W_i to the structural assignment for X_i yielding the modified SCM $\hat{\mathcal{C}}$ with structural assignments

$$\begin{aligned}
X_1 &:= N_{X_1} + W_1 \\
X_2 &:= X_1 + N_{X_2} + W_2 \\
H &:= N_H \\
X_3 &:= H + N_{X_3} + W_3 \\
Y &:= X_2 + H + N_Y \\
X_4 &:= X_1 + N_{X_4} + W_4 \\
X_5 &:= X_4 + Y + N_{X_5} + W_5.
\end{aligned}$$

We will refer to the variables W as *mean shifts*.

2.5 Markov properties and d-separation

Entailed DAGs are useful, because they allow us to quickly read off conditional independence statements that hold under the entailed distribution, as we will see in this section. The notion of d-separation in DAGs is central to the connection between DAGs and conditional independence statements (Pearl, 2009).

Definition 14 (d-separation)**.** Let $\mathcal{G} = (V, E)$ be a DAG. A path $(\gamma_1, \gamma_2, \ldots, \gamma_k)$ is said to be *blocked* by a subset $U \subseteq V$ if at least one of the following holds:

(i) There exists $i \in \{2, \ldots, k-1\}$ such that $\gamma_i \in U$ and the path contains either

$$\gamma_{i-1} \rightarrow \gamma_i \rightarrow \gamma_{i+1}, \quad \text{or} \quad \gamma_{i-1} \leftarrow \gamma_i \leftarrow \gamma_{i+1}, \quad \text{or} \quad \gamma_{i-1} \leftarrow \gamma_i \rightarrow \gamma_{i+1}.$$

(ii) There exists $i \in \{2, \ldots, k-2\}$ such that $(\{\gamma_i\} \cup \mathrm{de}(\gamma_i)) \cap U = \emptyset$ and γ_i is a collider on the path, *i.e.* the path contains $\gamma_{i-1} \rightarrow \gamma_i \leftarrow \gamma_{i+1}$.

If two disjoint subsets $W, Z \subseteq V$ are pairwise disjoint from U, and U blocks all paths between W and Z, then W and Z are said to be *d-separated by U in $\mathcal{G}$* and we write $W \perp\!\!\!\perp_{\mathcal{G}} Z \mid U$. A path that is not blocked is called *open*, and sets that are not d-separated are called *d-connected*.

A set of equivalent properties, known as the Markov properties, provide a link between d-separation and conditional independence statements (Lauritzen, 1996).

Definition 15 (Markov properties)**.** Let $X = (X_1, \ldots, X_d)$ be random variables with distribution P_X, and let $\mathcal{G}$ be a DAG. The following three properties are called the recursive density factorization property, the (directed) global Markov property, respectively the (directed) local Markov property.

(F) P has density p w.r.t. a product measure, and it factorizes as $p(x) = \prod_{i=1}^{d} p_i(x_i \mid \mathrm{pa}_i)$, where (for fixed pa_i) the function $x_i \mapsto p_i(x_i \mid \mathrm{pa}_i)$ is the density of the conditional distribution of X_i given $\mathrm{PA}_i = \mathrm{pa}_i$.

(G) For all $Y, Z, W \subseteq X$ where $Y \perp\!\!\!\perp_{\mathcal{G}} Z \mid W$ we have $Y \perp\!\!\!\perp_{P_X} Z \mid W$.

(L) For all $Y \in X$ we have $Y \perp\!\!\!\perp_{P_X} \mathrm{ND}_Y \setminus \mathrm{PA}_Y \mid \mathrm{PA}_Y$.

If P_X satisfies any of the above three properties w.r.t. $\mathcal{G}$, then it is said to *be Markov w.r.t.* $\mathcal{G}$.

As seen in Lauritzen (1996, Theorem 3.27), all three conditions turn out to be equivalent.

Theorem 16. *If P_X has density w.r.t. a product measure, then* (F) $\Leftrightarrow$ (G) $\Leftrightarrow$ (L).

One has to assume the Markov properties if only given a DAG and interventional distributions (Pearl, 2009, Definition 1.3.1), but for SCMs the entailed distribution is always Markov w.r.t. the entailed DAG (Pearl, 2009, Theorem 1.4.1).

Proposition 17. *Let $\mathcal{C}$ be an SCM over X with entailed distribution P_X and entailed DAG $\mathcal{G}$. If P_X has density w.r.t. a product measure, then P_X is Markov w.r.t. $\mathcal{G}$.*

Proof. (L) is immediate from the structural assignments, and (G) and (F) follow from Theorem 16. ∎

The assumption of density suits the needs of this thesis, but is not necessary since Theorem 16 and Proposition 17 also hold with (F) replaced by (R): *the recursive kernel factorization property* stating that P is a recursive combination of Markov kernels adapted to $\mathcal{G}$. If interested, see Lauritzen (2019, Section 2.6.2).

3 Causal discovery from interventional data

We will now consider three different causal discovery problems and present methods for their solution. The problems are related to what we call W-bridged SCMs; a class of SCMs that describe two separate repetitions of the exact same set of interventional experiments.

Sometimes the cause-effect relationships for a phenomenon are unknown, and the researcher has to infer properties of the underlying SCM from observations. This problem — known as causal discovery, causal learning, or causal inference — is the causal equivalent of inferring the observational distribution from observations; a problem known as statistical learning, or statistical inference. Rather than inferring the entire causal structure, we focus on the problem of inferring the causes of a single variable of interest.

Our overall goal is to infer the causal parents or ancestors of Y^0 in the linear Gaussian SCM $\mathcal{C}^0$ over (H^0, X^0, Y^0), where Y^0 is a response of interest, $X^0 = (X_1^0, \dots, X_d^0)$ are observable covariates, and H^0 is a set of hidden variables, with the following set of structural assignments

$$\begin{aligned}
H^0 &:= N_{H^0} \\
X^0 &:= A \begin{pmatrix} H^0 \\ X^0 \\ Y^0 \end{pmatrix} + N_{X^0} \\
Y^0 &:= \beta^t X^0 + \gamma^t H^0 + N_{Y^0},
\end{aligned}$$

where the matrix A and the vectors β and γ are coefficients. We wish to infer $\mathrm{pa}(Y^0) \cap X^0$ or $\mathrm{anc}(Y^0) \cap X^0$: the causal parents or ancestors of Y^0 among X^0. We assume that H^0 contains no descendant of X^0 and Y^0 (that is, $H^0 \subseteq \bigcap_{i=1}^d \mathrm{ND}_{X_i^0} \cap \mathrm{ND}_{Y^0}$), partly to avoid the situation $X_i^0 \to H_j^0 \to Y^0$, where a perfect method would wrongly infer $X_i^0 \in \mathrm{PA}_{Y^0}$.

3.1 W-bridged SCMs

3.1.1 Population case

Consider K different interventional experiments where the k'th experiment corresponds to a mean shift of X^0 by a variable W^k. This leads to the modified SCMs $(\mathcal{C}^k)_{k\in\{1,\ldots,K\}}$ given by

$$\begin{aligned} H^k &:= N_{H^k} \\ X^k &:= A \begin{pmatrix} H^k \\ X^k \\ Y^k \end{pmatrix} + N_{X^k} + W^k \\ Y^k &:= \beta^t X^k + \gamma^t H^k + N_{Y^k} \end{aligned} \tag{1}$$

where W and N are jointly independent. A separate repetition of the exact same experiments can be described by an SCM $(\widetilde{\mathcal{C}}^k)_{k\in\{1,\ldots,K\}}$ given by

$$\begin{aligned} \widetilde{H}^k &:= \widetilde{N}_{\widetilde{H}^k} \\ \widetilde{X}^k &:= A \begin{pmatrix} \widetilde{H}^k \\ \widetilde{X}^k \\ \widetilde{Y}^k \end{pmatrix} + \widetilde{N}_{\widetilde{X}^k} + W^k \\ \widetilde{Y}^k &:= \beta^t \widetilde{X}^k + \gamma^t \widetilde{H}^k + \widetilde{N}_{\widetilde{Y}^k}. \end{aligned} \tag{2}$$

for $k \in \{1,\ldots,K\}$. This has new noise variables $\widetilde{N}$ (representing that it is an independent repetition of the experiments), while the mean shifts $(W^k)_{k\in\{1,\ldots,K\}}$ and the causal mechanism (that is, the coefficients A, β and γ) are assumed to be the same for the two sets of experiments (representing that the experimental setting is exactly the same). We assume that W, N, and $\widetilde{N}$ are jointly independent. We similarly introduce an independent SCM $\widetilde{\mathcal{C}}^0$ for a separate repetition of the control experiment.

The pair of SCMs $(\mathcal{C}^0,\widetilde{\mathcal{C}}^0)$ can be seen as one SCM consisting of two separate components, since all noise variables are jointly independent. Intervening with the mean shifts W^k, which yields the pair $(\mathcal{C}^k,\widetilde{\mathcal{C}}^k)$, connects the two components via a bridge consisting of W^k.

Definition 18. The *W-bridged SCM* $(\mathcal{C}^k,\widetilde{\mathcal{C}}^k)$ is the SCM over $(X^k,Y^k,H^k,\widetilde{X}^k,\widetilde{Y}^k,\widetilde{H}^k,W^k)$ with structural assignments given by Eqs. (1) and (2) and

$$W^k := N_{W^k}.$$

To be clear: The noise variables are $N_{X^k},N_{Y^k},N_{H^k},\widetilde{N}_{\widetilde{X}^k},\widetilde{N}_{\widetilde{Y}^k},\widetilde{N}_{\widetilde{H}^k},N_{W^k}$ and are, as required for an SCM, assumed to be jointly independent. We will sometimes leave the structural assignment $W^k := N_{W^k}$ implicit, simply stating the structural assignments from Eqs. (1) and (2).

The entailed graph of the W-bridged SCM is obtained by drawing the graphs entailed by $\mathcal{C}^k$ and $\widetilde{\mathcal{C}}^k$ as two separate components, adding nodes W_1^k, …, W_d^k and edges $X_1^k \leftarrow W_1^k \rightarrow \widetilde{X}_1^k$, …, $X_d^k \leftarrow W_d^k \rightarrow \widetilde{X}_d^k$. We will sometimes refer to this graph as the *W-bridged DAG*. W^k forms a bridge between the, otherwise separate, $\mathcal{C}^k$ and $\widetilde{\mathcal{C}}^k$ components; the $\mathcal{C}^k$ component is equal to the graph entailed by $\mathcal{C}^0$ with, for all i, the node X_i^0 replaced by X_i^k.

Example 19. Consider the SCM with shift interventions from Example 13. We can construct two corresponding W-bridged SCMs $(\mathcal{C}^1, \widetilde{\mathcal{C}}^1)$ and $(\mathcal{C}^2, \widetilde{\mathcal{C}}^2)$ relating to the situation where we do two separate repetitions of one interventional experiment (so $K = 2$). The W-bridged SCM $(\mathcal{C}^1, \widetilde{\mathcal{C}}^1)$ is given by structural assignments $W^1 := N_{W^1}$ and

$$\begin{alignedat}{2}
X_1^1 &:= N_{X_1^1} + W_1^1 \qquad\qquad & \widetilde{X}_1^1 &:= \widetilde{N}_{\widetilde{X}_1^1} + W_1^1 \\
X_2^1 &:= X_1^1 + N_{X_2^1} + W_2^1 & \widetilde{X}_2^1 &:= \widetilde{X}_1^1 + \widetilde{N}_{\widetilde{X}_2^1} + W_2^1 \\
H^1 &:= N_{H^1} & \widetilde{H}^1 &:= \widetilde{N}_{\widetilde{H}^1} \\
X_3^1 &:= H^1 + N_{X_3^1} + W_3^1 & \widetilde{X}_3^1 &:= \widetilde{H}^1 + \widetilde{N}_{\widetilde{X}_3^1} + W_3^1 \\
Y^1 &:= X_2^1 + H^1 + N_{Y^1} & \widetilde{Y}^1 &:= \widetilde{X}_2^1 + \widetilde{H}^1 + \widetilde{N}_{\widetilde{Y}^1} \\
X_4^1 &:= X_1^1 + N_{X_4^1} + W_4^1 & \widetilde{X}_4^1 &:= \widetilde{X}_1^1 + \widetilde{N}_{\widetilde{X}_4^1} + W_4^1 \\
X_5^1 &:= X_4^1 + Y^1 + N_{X_5^1} + W_5^1 & \widetilde{X}_5^1 &:= \widetilde{X}_4^1 + \widetilde{Y}^1 + \widetilde{N}_{\widetilde{X}_5^1} + W_5^1,
\end{alignedat}$$

and $(\mathcal{C}^2, \widetilde{\mathcal{C}}^2)$ is given by $W^2 := N_{W^2}$ and

$$\begin{alignedat}{2}
X_1^2 &:= N_{X_1^2} + W_1^2 \qquad\qquad & \widetilde{X}_1^2 &:= \widetilde{N}_{\widetilde{X}_1^2} + W_1^2 \\
X_2^2 &:= X_1^2 + N_{X_2^2} + W_2^2 & \widetilde{X}_2^2 &:= \widetilde{X}_1^2 + \widetilde{N}_{\widetilde{X}_2^2} + W_2^2 \\
H^2 &:= N_{H^2} & \widetilde{H}^2 &:= \widetilde{N}_{\widetilde{H}^2} \\
X_3^2 &:= H^2 + N_{X_3^2} + W_3^2 & \widetilde{X}_3^2 &:= \widetilde{H}^2 + \widetilde{N}_{\widetilde{X}_3^2} + W_3^2 \\
Y^2 &:= X_2^2 + H^2 + N_{Y^2} & \widetilde{Y}^2 &:= \widetilde{X}_2^2 + \widetilde{H}^2 + \widetilde{N}_{\widetilde{Y}^2} \\
X_4^2 &:= X_1^2 + N_{X_4^2} + W_4^2 & \widetilde{X}_4^2 &:= \widetilde{X}_1^2 + \widetilde{N}_{\widetilde{X}_4^2} + W_4^2 \\
X_5^2 &:= X_4^2 + Y^2 + N_{X_5^2} + W_5^2 & \widetilde{X}_5^2 &:= \widetilde{X}_4^2 + \widetilde{Y}^2 + \widetilde{N}_{\widetilde{X}_5^2} + W_5^2.
\end{alignedat}$$

Assume that the noise variables are all jointly independent with marginal distributions

$$N_{X_i^j}, \widetilde{N}_{\widetilde{X}_i^j}, N_{Y^j}, \widetilde{N}_{\widetilde{Y}^j} \sim \mathcal{N}(0,1), \quad \widetilde{N}_{H^j}, \widetilde{N}_{\widetilde{H}^j} \sim \mathcal{N}(0,\tau), \quad \text{and } N_{W_i^j} \sim \mathcal{N}(0,\rho) \quad \text{for all } i, j.$$

Due to independence of the noise variables, we see that

$$(X^1, Y^1, H^1, \widetilde{X}^1, \widetilde{Y}^1, \widetilde{H}^1, W^1) \perp\!\!\!\perp (X^2, Y^2, H^2, \widetilde{X}^2, \widetilde{Y}^2, \widetilde{H}^2, W^2).$$

On the other hand, within $(\mathcal{C}^1, \widetilde{\mathcal{C}}^1)$ we do not have

$$(X^1, Y^1, H^1) \perp\!\!\!\perp (\widetilde{X}^1, \widetilde{Y}^1, \widetilde{H}^1)$$

since W^1 acts as a bridge between the two. Indeed, we, *e.g.*, see that

$$\mathrm{cov}(X_4^1, \widetilde{Y}^1) = \mathrm{cov}(N_{X_1^1} + W_1^1 + N_{X_4^1} + W_4^1, \widetilde{N}_{\widetilde{X}_1^1} + W_1^1 + \widetilde{N}_{\widetilde{X}_2^1} + W_2^1 + \widetilde{N}_{\widetilde{H}^1} + \widetilde{N}_{\widetilde{Y}^1}) = VW_1^1 = \rho$$

while

$$\mathrm{cov}(X_3^1, \widetilde{Y}^1) = \mathrm{cov}(N_{H^1} + N_{X_3^1} + W_3^1, \widetilde{N}_{\widetilde{X}_1^1} + W_1^1 + \widetilde{N}_{\widetilde{X}_2^1} + W_2^1 + \widetilde{N}_{\widetilde{H}^1} + \widetilde{N}_{\widetilde{Y}^1}) = 0$$

even though

$$\mathrm{cov}(X_3^1, Y^1) = \mathrm{cov}(N_{H^1} + N_{X_3^1} + W_3^1, N_{X_1^1} + W_1^1 + N_{X_2^1} + W_2^1 + N_{H^1} + N_{Y^1})) = VN_{H^1} = \tau.$$

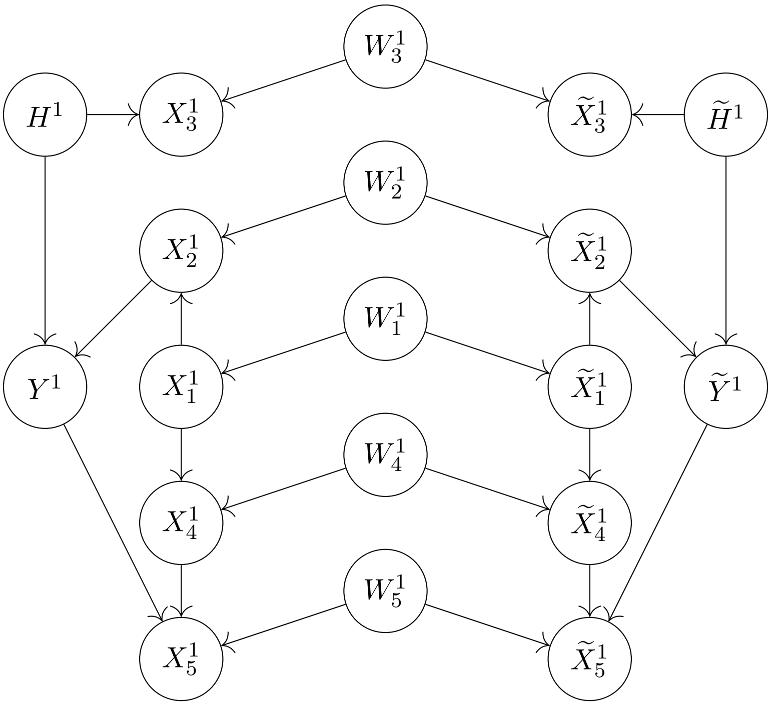


Figure 2: Entailed graph of the W-bridged SCM $(\mathcal{C}^1, \widetilde{\mathcal{C}}^1)$.

We can also see that $\operatorname{cov}(X_3^1, \widetilde{Y}^1) = 0$ from the entailed DAG of $(\mathcal{C}^1, \widetilde{\mathcal{C}}^1)$, depicted in Fig. 2, where all paths between X_3^1 and $\widetilde{Y}^1$ contain colliders, so X_3^1 and $\widetilde{Y}^1$ are d-separated, and hence, by the global Markov property, independent. On the other hand we see that X_4^1 and $\widetilde{Y}^1$ are not d-separated, since the path $X_4^1 \leftarrow X_1^1 \leftarrow W_1^1 \rightarrow \widetilde{X}_1^1 \rightarrow \widetilde{X}_2^1 \rightarrow \widetilde{Y}^1$ is open. The essential difference is that X_1^1 is a non-hidden common ancestor (confounder) of X_4^1 and Y^1, so we can go from X_4^1 to X_1^1, then cross the W-bridge from X_1^1 to $\widetilde{X}_1^1$ and proceed to $\widetilde{Y}^1$, while the only common ancestor of X_3^1 and Y^1 is H^1, which is not directly connected to the W-bridge, so there is no place to cross the W-bridge without running into a collider along the way. The following proposition formalizes this intuition.

Reichenbach's Common Cause Principle (Reichenbach, 1956), as stated in Peters et al. (2017, Principle 1.1), says that if $X \not\perp\!\!\!\perp Y$, then either X causes Y, Y causes X, or there is a third variable Z that causes both X and Y. This always holds in SCMs when "causes" is defined to mean "is an ancestor of" (Peters et al., 2017, Proposition 6.28). We now present a strengthened version for W-bridged SCMs.

Proposition 20 (A strong version of Reichenbach's Common Cause Principle)**.** *Let $k \in \{1, \ldots, K\}$, assume $i \neq j$, and consider the W-bridged SCM $(\mathcal{C}^k, \widetilde{\mathcal{C}}^k)$.*

If $X_i^k \not\perp\!\!\!\perp \widetilde{X}_j^k$ then at least one of the following holds:

(a) There is some non-hidden confounder $X_\ell^0 \in (\operatorname{anc}(X_i^0) \cap \operatorname{anc}(X_j^0)) \setminus H^0$, i.e., $\operatorname{anc}(X_i^0) \cap \operatorname{anc}(X_j^0) \not\subseteq H^0$.

(b) $X_i^0 \in \operatorname{anc}(X_j^0)$

(c) $X_j^0 \in \operatorname{anc}(X_i^0)$

Similarly, if $X_i^k \not\perp\!\!\!\perp \widetilde{Y}^k$ then at least one of the following holds:

(a) There is some non-hidden confounder $X_\ell^0 \in (\mathrm{anc}(X_i^0) \cap \mathrm{anc}(Y^0)) \setminus H^0$*, i.e.,* $\mathrm{anc}(X_i^0) \cap \mathrm{anc}(Y^0) \not\subseteq H^0$.

(b) $X_i^0 \in \mathrm{anc}(Y^0)$

This proposition holds without the assumption of Gaussianity or linearity.

Proof. Assume that $X_i^k \not\perp\!\!\!\perp \widetilde{X}_j^k$. By the global Markov property, this means that there must be a path between X_i^k and $\widetilde{X}_j^k$ without any colliders in the W-bridged DAG. Any path between them must pass at least one node in W^k to get from the $\mathcal{C}^k$-component to the $\widetilde{\mathcal{C}}^k$-component. If the path contains more than one node from W^k, then there must be W_t^k and W_s^k, where $t \neq s$, such that $W_t^k \to \cdots \leftarrow W_s^k$ is in the path, so the path contains a collider. Hence the open path must contain exactly one variable in W^k, call it W_ℓ^k. The path must contain X_ℓ^k and $\widetilde{X}_\ell^k$ as well, since they are the only nodes connected to W_ℓ^k. Assume first that $\ell \notin \{i, j\}$. The rest of the path between X_ℓ^k and X_i^k must be directed $X_\ell^k \to \cdots \to X_i^k$, since otherwise there would be a collider due to the first directed edge $W_\ell^k \to X_\ell^k$. This means that $X_\ell^k \in \mathrm{anc}(X_i^k)$, and since the $\mathcal{C}^k$-component of the W-bridged DAG is equal to the graph entailed by $\mathcal{C}^0$ (with the superscript of the nodes changed) it follows that $X_\ell^0 \in \mathrm{anc}(X_i^0)$. The exact same argument shows that $\widetilde{X}_\ell^k \in \mathrm{anc}(\widetilde{X}_j^k)$, so since the $\widetilde{\mathcal{C}}^k$-component of the W-bridged DAG is also equal to the graph entailed by $\mathcal{C}^0$ (with the nodes changed), we get $X_\ell^0 \in \mathrm{anc}(X_j^0)$. Hence $X_\ell^0 \in (\mathrm{anc}(X_i^0) \cap \mathrm{anc}(X_j^0)) \setminus H^0$; case (a). If $\ell = i$, then the argument that $X_\ell^0 \in \mathrm{anc}(X_j^0)$ still works, giving us that $X_i^0 \in \mathrm{anc}(X_j^0)$; case (b). Similarly, if $\ell = j$, then the argument that $X_\ell^0 \in \mathrm{anc}(X_i^0)$ still works, giving us that $X_j^0 \in \mathrm{anc}(X_i^0)$; case (c). If X_j was replaced by Y, the only change would be to skip the case $\ell = j$, meaning that either (a) or (b) holds, completing the proof. ∎

Loosely speaking, an observed variable in the $\mathcal{C}^k$-component and an observed variable in the $\widetilde{\mathcal{C}}^k$-component cannot be confounded by a hidden variable. The transposes of the statements in the proposition may make this more clear:

- If $X_i^0 \notin \mathrm{anc}(X_j^0)$, and $X_j^0 \notin \mathrm{anc}(X_i^0)$, and $\mathrm{anc}(X_i^0) \cap \mathrm{anc}(X_j^0) \subseteq H^0$, then $X_i^k \perp\!\!\!\perp \widetilde{X}_j^k$.
- If $X_i^0 \notin \mathrm{anc}(Y^0)$, and $\mathrm{anc}(X_i^0) \cap \mathrm{anc}(Y^0) \subseteq H^0$, then $X_i^k \perp\!\!\!\perp \widetilde{Y}^k$.

This motivates the methods proposed in Section 3.5.

3.1.2 Finite data case: truly separate data

In the real world, the exact observational distribution needed is often unknown, and one must instead rely on approximations using finite data. For all $k \in \{1, \ldots, K\}$, the W-bridged SCM $(\mathcal{C}^k, \widetilde{\mathcal{C}}^k)$ corresponds to a particular experimental setup. Assume that we conduct n_k repetitions in the k'th setup. The i'th repetition (where $i \in \{1, \ldots, n_k\}$) in the k'th setting corresponds to data from the W-bridged SCM $(\mathcal{C}^{k,i}, \widetilde{\mathcal{C}}^{k,i})$ given by

$$\begin{aligned} H^{k,i} &:= N_{H^{k,i}} & \widetilde{H}^{k,i} &:= \widetilde{N}_{\widetilde{H}^{k,i}} \\ X^{k,i} &:= A \begin{pmatrix} H^{k,i} \\ X^{k,i} \\ Y^{k,i} \end{pmatrix} + N_{X^{k,i}} + W^k & \widetilde{X}^{k,i} &:= A \begin{pmatrix} \widetilde{H}^{k,i} \\ \widetilde{X}^{k,i} \\ \widetilde{Y}^{k,i} \end{pmatrix} + \widetilde{N}_{\widetilde{X}^{k,i}} + W^k \\ Y^{k,i} &:= \beta^t X^{k,i} + \gamma^t H^{k,i} + N_{Y^{k,i}}. & \widetilde{Y}^{k,i} &:= \beta^t \widetilde{X}^{k,i} + \gamma^t \widetilde{H}^{k,i} + \widetilde{N}_{\widetilde{Y}^{k,i}}. \end{aligned}$$

We assume that W^k is the same across all repetitions $i \in \{1, \ldots, n_k\}$ within each experimental setting, but that the rest of the noise variables $(N_X, N_Y, N_H, \widetilde{N}_{\widetilde{X}}, \widetilde{N}_{\widetilde{Y}}, \widetilde{N}_{\widetilde{H}})$ are jointly independent across all repetitions. This means that, in the finite data setting, for all $k \in \{1, \dots, K\}$ we must use observations from the W-bridged SCMs $(\mathcal{C}^{k,i}, \widetilde{\mathcal{C}}^{k,i})$ to get the properties of the observational distribution entailed by $(\mathcal{C}^k, \widetilde{\mathcal{C}}^k)$ that we need.

To this end, we bundle the observations as follows. Let $\mathbf{X}^k$ denote the $n_k \times d$ matrix where the i'th row is an observation of $X^{k,i}$, let $\mathbf{Y}^k \in \mathbb{R}^{n_k}$ denote the vector where the i'th element is an observation of $Y^{k,i}$, and let

$$\mathbf{X} := \begin{pmatrix} \mathbf{X}^1 \\ \vdots \\ \mathbf{X}^K \end{pmatrix}, \quad \mathbf{Y} := \begin{pmatrix} \mathbf{Y}^1 \\ \vdots \\ \mathbf{Y}^K \end{pmatrix},$$

that is, $\mathbf{X}$ is a $\left(\sum_{k=1}^K n_k\right) \times d$ matrix and $Y \in \mathbb{R}^{\sum_{k=1}^K n_k}$.

For the second set of experiments we introduce similar notation $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. Thus, $(\mathbf{X}, \mathbf{Y}, \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ denotes a complete data set from a set of W-bridged SCMs, corresponding to the real-world situation, where two separate sets of experiments are carried out. We call this type of data *truly separate data*, where the word "truly" is used to distinguish it from "permuted separate data", which is introduced below.

3.1.3 Permuted separate data

If we only conduct one set of experiments and obtain the data set $(\mathbf{X}, \mathbf{Y})$, then we can emulate a second set of experiments by permuting the rows of $\mathbf{X}$ and $\mathbf{Y}$ within each experimental setting, *i.e*, for all $k \in \{1, \dots, K\}$ we let P^k be an $n_k \times n_k$ permutation matrix, and permute $\mathbf{Y}^k$ and the rows of $\mathbf{X}^k$ using P^k to obtain

$$\breve{\mathbf{X}} := \begin{pmatrix} P^1\mathbf{X}^1 \\ \vdots \\ P^K\mathbf{X}^K \end{pmatrix}, \quad \breve{\mathbf{Y}} := \begin{pmatrix} P^1\mathbf{Y}^1 \\ \vdots \\ P^K\mathbf{Y}^K \end{pmatrix}.$$

The permuted data $(\breve{\mathbf{X}}, \breve{\mathbf{Y}})$ is interpreted as an emulation of data $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ from a separate set of experiments. We call this type of data *permuted separate data.*

$y^{k,i}$ from the same experiment as $x^{k,i}$ is still equal to $\breve{y}^{k,j}$ for some j, since we have merely permuted our observations; this is not the case for truly separate data. Another way to obtain a second data set circumvents this problem by, in effect, splitting a single data set into truly separate data. By splitting the data from each environment in two equally sized portions, instead of permuting, it does not merely *emulate* data from W-bridged SCMs; it *is* data from W-bridged SCMs. This means that any results about truly separate data holds for data split into two in this fashion, but for any concrete setting, it halves the number of observations per environment.

3.2 Problems we try to solve

Our thesis tackles three main problems that we state now. They are all formulated in finite data versions, but the first two problems are also addressed in the population scenarios with complete knowledge of relevant observational distributions, corresponding to the asymptotic case of practically infinite data. Our overall aim in all three problems is to infer the parents or ancestors of Y^0 among X^0; the difference between the problems lies in what we observe.

3.2.1 Problem A: causal discovery from unpaired data

Imagine a real-world scenario where X^k and Y^k cannot both be observed in the same experiment, for instance because observing X^k destroys the physical entity underlying Y^k and vice versa. Instead we conduct all experiments twice, observing X in the first set of experiments and Y in the second set. This means that we have incomplete truly separate data, since we only observe $(\mathbf{X}, \widetilde{\mathbf{Y}})$, and not $\mathbf{Y}$, nor $\widetilde{\mathbf{X}}$.

Problem A. Infer the parents or ancestors of Y^0 among X^0 from the observations $(\mathbf{X}, \widetilde{\mathbf{Y}})$.

In the k'th experiment our observations are

$$(\mathbf{X}^k, \widetilde{\mathbf{Y}}^k) = \begin{pmatrix} x^{k,1} & \widetilde{y}^{k,1} \\ \vdots & \vdots \\ x^{k,n_k} & \widetilde{y}^{k,n_k} \end{pmatrix}.$$

The pairing is artificial, since $x^{k,i}$ and $\widetilde{y}^{k,i}$ are no longer from the same experiment, but only from the arbitrarily numbered i'th repetition in their respective sets of experiments. Indeed, for all i, j, we have $(X^{k,i}, \widetilde{Y}^{k,i}) \overset{\mathcal{D}}{=} (X^{k,i}, \widetilde{Y}^{k,j})$, so $x^{k,i}$ might as well be paired with $\widetilde{y}^{k,j}$.

In the population case, this problem corresponds to, for all $k \in \{1, \ldots, K\}$, knowing the distribution of $(X^k, \widetilde{Y}^k)$, but not of Y^k, nor of $\widetilde{X}^k$.

3.2.2 Problem B: causal discovery from truly separate data

Of course, there may be useful information in $\mathbf{Y}$ or $\widetilde{\mathbf{X}}$. This leads us to the second problem.

Problem B. Infer the parents and ancestors of Y^0 among X^0 from truly separate data $(\mathbf{X}, \mathbf{Y}, \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$.

By the data splitting technique described in the bottom of Section 3.1.3, a single data set can be split into truly separate data, so Problem B does not only deal with the situation of two physically separate sets of experiments.

3.2.3 Problem C: causal discovery from permuted separate data

If we are able to observe $x^{k,i}$ and $y^{k,i}$ paired, and conduct a single set of experiments, then we obtain a single paired data set $(\mathbf{X}, \mathbf{Y})$. By permuting the rows within environments, as explained in Section 3.1.3, we obtain permuted separate data $(\mathbf{X}, \mathbf{Y}, \breve{\mathbf{X}}, \breve{\mathbf{Y}})$. Any method that is useful in Problem A and Problem B is also applicable in this case, since permuted separate data is an emulation of truly separate data. However, they may perform differently, due to the inherent differences between permuted- and truly separate data, which leads us to the third, and last, problem.

Problem C. Infer the parents and ancestors of Y^0 among X^0 from permuted separate data $(\mathbf{X}, \mathbf{Y}, \breve{\mathbf{X}}, \breve{\mathbf{Y}})$.

Remember that we only need to observe $(\mathbf{X}, \mathbf{Y})$ to obtain $(\mathbf{X}, \mathbf{Y}, \breve{\mathbf{X}}, \breve{\mathbf{Y}})$. This means that Problem C deals with the usual setting of one paired data set, but through the unusual means of permuting the rows to emulate a second data set.

When discussing the methods we will often write $\widetilde{\mathbf{X}}$ or $\widetilde{\mathbf{Y}}$, but one can plug in the permuted data sets $\breve{\mathbf{X}}$ or $\breve{\mathbf{Y}}$ instead when working with Problem C. However, in population arguments we strictly consider the case of truly separate data; heuristically, the population arguments for Problem B should carry over to Problem C to a certain degree, since the permuted data emulates data from an actual separate set of experiments.

3.3 Distributions of the mean shift W: alltargets and singletargets

We consider two specific distributions of $(W^1, \dots, W^K)$ in the rest of this thesis. In the first setting, which we call *alltargets*, there is a $\sigma^2 \in (0, \infty)$ such that $W^1, \dots, W^K \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$. In the second setting — named *singletargets* — W is distributed as follows: There is a $\nu^2 \in (0, \infty)$ such that for all $k \in \{1, \dots, K\}$ there exists $j \in \{1, \dots, d\}$ such that $W_j^k \sim \mathcal{N}(0, \nu^2)$ and, for all $i \neq j$, we have $W_i^k = 0$. In the singletargets setup we also include observations from a control experiment with no shift interventions[2]. Briefly, in alltargets we intervene on all X's in each setting, while in singletargets we only intervene on a single X at a time. We focus on alltargets, but include singletargets to have an example of a somewhat degenerate setting.

Except for the baseline method "mean-shift", our proposed methods do not assume knowledge of the distribution of W, nor, in particular, of the intervention targets in the different environments of the singletargets setting.

3.4 OLS

OLS is a well-known regression technique. We briefly discuss its shortcomings in our setting, before presenting two novel methods, both variations of OLS, designed to overcome its problems. We define the population parameter

$$\beta^{\text{OLS}} := \left(\sum_{k=1}^{K} \text{cov}(X^k) \right)^{-1} \sum_{k=1}^{K} \text{cov}(X^k, Y^k),$$

which is the solution to the following population version of the classic least squares regression problem[3]

$$\arg\min_{\beta} \sum_{k=1}^{K} E(Y^k - \beta^t X^k)^2. \tag{3}$$

The population parameter β^{OLS} can be estimated by the well-known estimator

$$\hat{\beta}^{\text{OLS}} := (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}.$$

[2] In the singletargets setup we include a baseline method ("mean-shift"), for which control observations would be needed in a real world case where $EY^0 = 0$ doesn't necessarily hold. See Section 3.6.

[3] See Appendix A for details.

In the alltargets setting $X^1, \dots, X^K$ all have the same distribution, since $W^1, \dots, W^K$ do, and $Y^1, \dots, Y^K$ all have the same distribution as well. This means that

$$\begin{aligned}\beta^{\mathrm{OLS}} &= \left(\sum_{k=1}^{K} \mathrm{cov}(X^k)\right)^{-1} \sum_{k=1}^{K} \mathrm{cov}(X^k, Y^k) \\ &= \left(K\mathrm{cov}(X^1)\right)^{-1} K\mathrm{cov}(X^1, Y^1) \\ &= \left(\mathrm{cov}(X^1)\right)^{-1} \mathrm{cov}(X^1, Y^1).\end{aligned}$$

Due to the calculation

$$\begin{aligned}&(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \\ &= \begin{pmatrix} \sum_{k=1}^K \sum_{i=1}^{n_k} x_1^{k,i} x_1^{k,i} & \cdots & \sum_{k=1}^K \sum_{i=1}^{n_k} x_1^{k,i} x_d^{k,i} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^K \sum_{i=1}^{n_k} x_d^{k,i} x_1^{k,i} & \cdots & \sum_{k=1}^K \sum_{i=1}^{n_k} x_d^{k,i} x_d^{k,i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{k=1}^K \sum_{i=1}^{n_k} x_1^{k,i} y^{k,i} \\ \vdots \\ \sum_{k=1}^K \sum_{i=1}^{n_k} x_d^{k,i} y^{k,i} \end{pmatrix} \\ &= \left[\frac{1}{K}\sum_{k=1}^{K} \frac{1}{n_k}\sum_{i=1}^{n_k} \begin{pmatrix} x_1^{k,i} x_1^{k,i} & \cdots & x_1^{k,i} x_d^{k,i} \\ \vdots & \ddots & \vdots \\ x_d^{k,i} x_1^{k,i} & \cdots & x_d^{k,i} x_d^{k,i} \end{pmatrix}\right]^{-1} \frac{1}{K}\sum_{k=1}^{K} \frac{1}{n_k}\sum_{i=1}^{n_k} \begin{pmatrix} x_1^{k,i} y^{k,i} \\ \vdots \\ x_d^{k,i} y^{k,i} \end{pmatrix}\end{aligned} \tag{4}$$

a modified version of the strong law of large numbers[4] gives strong consistency of $\hat{\beta}^{\mathrm{OLS}}$ in the alltargets setting with truly separate data: $\hat{\beta}^{\mathrm{OLS}} \overset{\text{a.s}}{\to} \beta^{\mathrm{OLS}}$ for $K \to \infty$. It is, however, not enough that $n_k \to \infty$ for fixed K, since W^k is fixed within environments and thus breaks independence. Intuitively, if we only have a single environment and observe a large value of W^1, then the matrices in Eq. (4) are going to be poor estimates of the covariance matrices, no matter how many observations we take in the one environment, since W^1 will not change.

The minimization problem (3) serves as motivation that β^{OLS} is similar to the ψ in $E(Y^0 \mid X^0 = x) = \psi^t x$. The nonzero entries of ψ ideally correspond to the *Markov blanket of* Y^0 *in* X^0 (Peters et al., 2017, Definition 6.26).

Definition 21 (Markov blanket)**.** Let $\mathcal{G}$ be a DAG over nodes V, let $O \subseteq V$, and let $Y \in V$. The *Markov blanket* of Y in O is the smallest subset $M \subseteq O$ such that

$$Y \perp\!\!\!\perp_{\mathcal{G}} O \setminus (\{Y\} \cup M) \mid M.$$

If V is a set of random variables and P_V is Markov w.r.t. $\mathcal{G}$, then

$$Y \perp\!\!\!\perp_{P_V} O \setminus (\{Y\} \cup M) \mid M.$$

A simple d-separation argument shows that the Markov blanket of Y in V is $M = \mathrm{PA}_Y \cup \mathrm{CH}_Y \cup \mathrm{PA}(\mathrm{CH}_Y)$. If $V = (X^0, Y^0, H^0)$, then the same argument shows that the Markov blanket of Y^0 in X^0 must contain $X \cap (\mathrm{PA}_Y \cup \mathrm{CH}_Y \cup \mathrm{PA}(\mathrm{CH}_Y))$, but it will also contain any X_i^0 where there exists a path where every second node is a hidden confounder: $Y^0 \leftarrow H_a^0 \rightarrow X_\alpha^0 \leftarrow \cdots \rightarrow X_\beta^0 \leftarrow H_b^0 \rightarrow X_i^0$. It will also contain any parent of such an X_i^0. This

[4] See Appendix B for details. A slight modification of the strong law of large numbers is needed, since the n_k may be different, leading to non-identical (but still independent) distributions.

indicates that the β^{OLS} coefficients for the parents of Y^0 will be nonzero, but that there will also be nonzero β^{OLS} coefficients for other variables that are neither parents nor ancestors of Y^0.

In spite of this, we will use OLS as a baseline method, where the variable with the largest absolute $\hat{\beta}^{\mathrm{OLS}}$ coefficient is taken to be the most likely parent or ancestor.

3.5 Novel methods

One of the problems of OLS is due to hidden confounding; a problem that doesn't affect $\mathrm{cov}(X_i^k, \widetilde{Y}^k)$ and $\mathrm{cov}(X_i^k, \widetilde{X}_j^k)$, as we saw in Proposition 20. This motivates the two novel methods, POLS and DPOLS, both of which consist of plugging $\widetilde{Y}$ or $\widetilde{X}$ into the formula for β^{OLS}. We also briefly present a third possible method, PICP, that we will not study in detail.

3.5.1 POLS (Permuted OLS)

We replace $\mathrm{cov}(X^k, Y^k)$ by $\mathrm{cov}(X^k, \widetilde{Y}^k)$ in β^{OLS} to obtain the population parameter

$$\beta^{\mathrm{POLS}} := \left(\sum_{k=1}^{K} \mathrm{cov}(X^k)\right)^{-1} \sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{Y}^k),$$

with the estimator

$$\hat{\beta}^{\mathrm{POLS}} := (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\widetilde{\mathbf{Y}}.$$

Again we find for the alltargets setting that

$$\beta^{\mathrm{POLS}} = \left(\mathrm{cov}(X^1)\right)^{-1} \mathrm{cov}(X^1, \widetilde{Y}^1),$$

and an argument similar to (4) gives that $\hat{\beta}^{\mathrm{POLS}} \overset{\text{a.s}}{\to} \beta^{\mathrm{POLS}}$ for $K \to \infty$ in the alltargets setting with truly separate data. Since POLS only uses $(\mathbf{X}, \widetilde{\mathbf{Y}})$, it is applicable in Problem A (causal discovery from unpaired data), unlike OLS.

We will not provide theoretical guarantees for POLS, but the following section presents a population argument for the similar method DPOLS. Heuristically, this argument also indicates that POLS is better than OLS, due to similarity between POLS and DPOLS.

When using POLS as a method to identify parents or ancestors below, we will concretely take the variable with the largest absolute $\hat{\beta}^{\mathrm{POLS}}$ coefficient to be the most likely parent or ancestor.

3.5.2 DPOLS (Double-Permuted OLS)

Assuming that $\sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{X}^k)$ is invertible, we can also replace $\mathrm{cov}(X^k)$ by $\mathrm{cov}(X^k, \widetilde{X}^k)$ to obtain the population parameter

$$\beta^{\mathrm{DPOLS}} := \left(\sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{X}^k)\right)^{-1} \sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{Y}^k)$$

with the estimator

$$\hat{\beta}^{\mathrm{DPOLS}} := (\mathbf{X}^t\widetilde{\mathbf{X}})^{-1}\mathbf{X}^t\widetilde{\mathbf{Y}},$$

assuming that $\mathbf{X}^t\widetilde{\mathbf{X}}$ is invertible. Again we find for the alltargets setting that

$$\beta^{\text{DPOLS}} = \left(\text{cov}(X^1, \widetilde{X}^1)\right)^{-1} \text{cov}(X^1, \widetilde{Y}^1),$$

and an argument similar to Eq. (4) gives that $\hat{\beta}^{\text{DPOLS}} \overset{\text{a.s}}{\to} \beta^{\text{DPOLS}}$ for $K \to \infty$ in the alltargets setting with truly separate data. Combined with the following proposition, due to Niklas Pfister, this makes a strong case for the usefulness of DPOLS. We use β in the meaning from Eqs. (1) and (2).

Proposition 22. *If $\sum_{k=1}^K \text{cov}(X^k, \widetilde{X}^k)$ is invertible, then $\beta^{\text{DPOLS}} = \beta$.*

Proof. Since

$$\text{cov}(X^k, \widetilde{Y}^k) = \text{cov}(X^k, \beta^t\widetilde{X}^k + \gamma^t\widetilde{H}^k + \widetilde{N}_{\widetilde{Y}^k}) = \text{cov}(X^k, \widetilde{X}^k)\beta$$

we have

$$\begin{aligned}\beta^{\text{DPOLS}} &= \left(\sum_{k=1}^K \text{cov}(X^k, \widetilde{X}^k)\right)^{-1} \sum_{k=1}^K \text{cov}(X^k, \widetilde{Y}^k)\\ &= \left(\sum_{k=1}^K \text{cov}(X^k, \widetilde{X}^k)\right)^{-1} \sum_{k=1}^K \text{cov}(X^k, \widetilde{X}^k)\beta\\ &= \beta.\end{aligned}$$

∎

We will now show that $\sum_{k=1}^K \text{cov}(X^k, \widetilde{X}^k)$ is invertible in the alltargets setting, and that it is invertible in the singletargets setting if and only if all X's are intervened on. To prove these statements, we first give a proposition, which is again based on an unpublished argument by Niklas Pfister.

Proposition 23. *The sum $\sum_{k=1}^K \text{cov}(X^k, \widetilde{X}^k)$ is invertible if and only if $\sum_{k=1}^K \text{cov}(W^k)$ is invertible.*

Proof. Let A be the coefficient matrix from Eqs. (1) and (2). Let Θ be the first d columns of A, let Λ be the next d columns of A, and let α be the last column of A. Then

$$\begin{aligned}X^k &= \Theta H^k + \Lambda X^k + \alpha Y^k + N_{X^k} + W^k\\ &= \Theta H^k + \Lambda X^k + \alpha(\beta^t X^k + \gamma^t H^k + N_{Y^k}) + N_{X^k} + W^k\\ &= (\Lambda + \alpha\beta^t)X^k + \Theta H^k + \alpha(\gamma^t H^k + N_{Y^k}) + N_{X^k} + W^k,\end{aligned}$$

and hence

$$X^k = (I - \Lambda - \alpha\beta^t)^{-1}(\Theta H^k + \alpha(\gamma^t H^k + N_{Y^k}) + N_{X^k} + W^k),$$

where invertibility of $I - \Lambda - \alpha\beta^t$ follows from the non-cyclic structure. Similarly, we have

$$\widetilde{X}^k = (I - \Lambda - \alpha\beta^t)^{-1}(\Theta\widetilde{H}^k + \alpha(\gamma^t\widetilde{H}^k + \widetilde{N}_{\widetilde{Y}^k}) + \widetilde{N}_{\widetilde{X}^k} + W^k),$$

so

$$\text{cov}(X^k, \widetilde{X}^k) = (I - \Lambda - \alpha\beta^t)^{-2}\text{cov}(W^k),$$

and thus

$$\sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{X}^k) = (I - \Lambda - \alpha\beta^t)^{-2} \sum_{k=1}^{K} \operatorname{cov}(W^k),$$

which is invertible if and only if $\sum_{k=1}^{K} \operatorname{cov}(W^k)$ is invertible. ∎

Using this proposition we can characterize when the invertibility assumption is satisfied in the alltargets and singletargets settings.

Proposition 24. $\sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{X}^k)$ *is invertible in the alltargets setting.*

Proof. In the alltargets setting $W^1, \ldots, W^K$ all have the same distribution so

$$\sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{X}^k) = (I - \Lambda - \alpha\beta^t)^{-2} \sum_{k=1}^{K} \operatorname{cov}(W^k) = (I - \Lambda - \alpha\beta^t)^{-2} K \operatorname{cov}(W^1),$$

showing that it is necessary and sufficient that $\operatorname{cov}(W^1)$ is invertible, which it is, since

$$\operatorname{cov}(W^1) \propto I_d$$

in the alltargets setting. ∎

Proposition 25. *In the singletargets setting,* $\sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{X}^k)$ *is invertible if and only if all X's are intervened on (i.e., if and only if, for all $i \in \{1, \ldots, d\}$, there is a $k \in \{1, \ldots, K\}$ such that we intervene on X_i^k in the k'th experimental setting).*

Proof. In the singletargets setting, $\operatorname{cov}(W^k)$ is a diagonal matrix with exactly one nonzero entry; indeed, if i is the index of the intervention target X_i^k in the k'th experiment, then the only nonzero entry of $\operatorname{cov}(W^k)$ is in position (i, i). This means that $\sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{X}^k)$ is invertible if and only if, for all $i \in \{1, \ldots, d\}$, there is at least one $k \in \{1, \ldots, K\}$ such that we intervene on X_i^k in the k'th experimental setting. ∎

In many of our singletargets simulations we will intervene on all X's, but we will also explore what happens when we only intervene on half of the X's, meaning that $\sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{X}^k)$ is non-invertible.

Combining the strong consistency with Propositions 22 and 24 we now have an asymptotic guarantee for DPOLS applied on truly separate data in the alltargets setting.

Proposition 26. *For truly separate data in the alltargets setting,* $\hat{\beta}_i^{\mathrm{DPOLS}} \overset{\mathrm{a.s}}{\to} 0$ *for* $K \to \infty$ *if and only if* X_i^0 *is not a parent of* Y^0.

As for POLS, when we below use DPOLS to select parents or ancestors, we concretely take the variable with the largest absolute $\hat{\beta}^{\mathrm{DPOLS}}$ coefficient to be the most likely parent or ancestor.

3.5.3 Example of OLS, POLS, and DPOLS

Example 27. Consider again the setup from Example 19, and note that it corresponds to an alltargets setting. By doing calculations similar to those in Example 19 we find that

$$\operatorname{cov}(X^1) = \begin{pmatrix} 1+\rho & 1+\rho & 0 & 1+\rho & 2+2\rho \\ 1+\rho & 2+2\rho & 0 & 1+\rho & 3+3\rho \\ 0 & 0 & \tau+1+\rho & 0 & \tau \\ 1+\rho & 1+\rho & 0 & 2+2\rho & 3+3\rho \\ 2+2\rho & 3+3\rho & \tau & 3\rho+3\rho & 8+7\rho+\tau \end{pmatrix},$$

$$\operatorname{cov}(X^1, \widetilde{X}^1) = \begin{pmatrix} \rho & \rho & 0 & \rho & 2\rho \\ \rho & 2\rho & 0 & \rho & 3\rho \\ 0 & 0 & \rho & 0 & 0 \\ \rho & \rho & 0 & 2\rho & 3\rho \\ 2\rho & 3\rho & 0 & 3\rho & 7\rho \end{pmatrix}, \quad \operatorname{cov}(X^1, Y^1) = \begin{pmatrix} 1+\rho \\ 2+2\rho \\ \tau \\ 1+\rho \\ 4+3\rho+\tau \end{pmatrix}, \quad \operatorname{cov}(X^1, \widetilde{Y}^1) = \begin{pmatrix} \rho \\ 2\rho \\ 0 \\ \rho \\ 3\rho \end{pmatrix}.$$

Assuming that $\rho \neq 0$, such that W is non-degenerate, both $\operatorname{cov}(X^1)$ and $\operatorname{cov}(X^1, \widetilde{X}^1)$ are invertible and we obtain that

$$\beta^{\mathrm{OLS}} = (\operatorname{cov}(X^1))^{-1}\operatorname{cov}(X^1, Y^1) = \frac{1}{\rho^2 + (2\tau+3)\rho + 3\tau + 2} \begin{pmatrix} 0 \\ (1+\rho)(\tau+1+\rho) \\ (1+\rho)\tau \\ -((\tau+1)\rho + 2\tau + 1) \\ (\tau+1)\rho + 2\tau + 1 \end{pmatrix},$$

$$\beta^{\mathrm{POLS}} = (\operatorname{cov}(X^1))^{-1}\operatorname{cov}(X^1, \widetilde{Y}^1) = \begin{pmatrix} 0 \\ \frac{\rho}{1+\rho} \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \beta^{\mathrm{DPOLS}} = (\operatorname{cov}(X^1, \widetilde{X}^1))^{-1}\operatorname{cov}(X^1, \widetilde{Y}^1) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

We see that $\beta^{\mathrm{DPOLS}} = \beta$, as expected from Proposition 22 since $\operatorname{cov}(X^1, \widetilde{X}^1)$ is invertible. Furthermore, we see that $\beta^{\mathrm{POLS}} = \frac{\rho}{1+\rho}\beta$. This means that $\beta_i^{\mathrm{POLS}} = 0$ if and only if X_i^0 is a parent of Y^0, so β^{POLS} identifies non-hidden parents of Y^0 in this example, just like β^{DPOLS} does generally. We also see that β^{OLS} finds the Markov blanket of Y^0; the only 0-entry in β^{OLS} is the first, even though X_1^0 is an ancestor of Y^0. When ρ is small and τ is large X_4 and X_5 will be selected first, while X_2 will be selected first when ρ is large and τ is small. So, with usual OLS, the intervention has to be strong enough to outweigh the confounding due to the hidden variable.

3.5.4 p-values for OLS, POLS, and DPOLS

We will briefly present possible ways to calculate p-values for OLS, POLS, and DPOLS, and argue why they may not lead to the desired Type I error rate. We will include the p-values as a ranking mechanism in our simulation studies, but our focus is elsewhere. For simplicity we regard $\mathbf{X}$ as fixed, and use Y to denote the random vector observed as $\mathbf{Y}$.

Under the linear regression assumption $Y \sim \mathcal{N}(\mathbf{X}\beta^{\text{OLS}}, \sigma^2_{\text{OLS}} I)$, with $n = \sum_{k=1}^{K} n_k$ observations and d-dimensional X^0, the usual formula for the p-value of the hypothesis $\mathcal{H}_i : \beta_i^{\text{OLS}} = 0$ is

$$p_i^{\text{OLS}} := 2P\left(T \geq \left|\frac{\hat{\beta}_i^{\text{OLS}}}{\sqrt{\hat{\sigma}^2_{\text{OLS}}((\mathbf{X}^t\mathbf{X})^{-1})_{i,i}}}\right|\right), \quad \text{where } T \sim t_{n-d}. \tag{5}$$

This is not directly applicable to our case, since, for all $k \in \{1, \ldots, K\}$, the variables $Y^{k,1}, \ldots, Y^{k,n_k}$ are not independent, because W^k is the same in all repetitions of the experiment, so the covariance assumption is violated.

Under the assumption that $\widetilde{Y} \sim \mathcal{N}(\mathbf{X}\beta^{\text{POLS}}, \sigma^2_{\text{POLS}} I)$, we can similarly get a p-value for the hypothesis $\mathcal{H}_i : \beta_i^{\text{POLS}} = 0$ as

$$p_i^{\text{POLS}} = 2P\left(T \geq \left|\frac{\hat{\beta}_i^{\text{POLS}}}{\sqrt{\hat{\sigma}^2_{\text{POLS}}((\mathbf{X}^t\mathbf{X})^{-1})_{i,i}}}\right|\right), \quad \text{where } T \sim t_{n-d}. \tag{6}$$

However, under these assumptions $E\widetilde{Y}$ is determined by $\mathbf{X}$, which is not true since they represent separate experiments. Also, the independence assumption is violated within environments, as we saw for OLS.

If we assume that $\widetilde{Y} \sim \mathcal{N}(\widetilde{\mathbf{X}}\beta^{\text{DPOLS}}, \sigma^2_{\text{DPOLS}} I)$ then

$$\hat{\beta}^{\text{DPOLS}} = (\mathbf{X}^t\widetilde{\mathbf{X}})^{-1}\mathbf{X}^t\widetilde{Y} \sim \mathcal{N}(\beta^{\text{DPOLS}}, \sigma^2_{\text{DPOLS}}(\mathbf{X}^t\widetilde{\mathbf{X}})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\widetilde{\mathbf{X}})^{-t})$$

so under the hypothesis $\mathcal{H}_i : \beta_i^{\text{DPOLS}} = 0$ we have

$$\frac{\hat{\beta}_i^{\text{DPOLS}}}{\sqrt{\hat{\sigma}^2_{\text{DPOLS}}\left((\mathbf{X}^t\widetilde{\mathbf{X}})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\widetilde{\mathbf{X}})^{-t}\right)_{i,i}}} \sim t_{n-d}.$$

Hence we can obtain a p-value as

$$p_i^{\text{DPOLS}} = 2P\left(T \geq \left|\frac{\hat{\beta}_i^{\text{DPOLS}}}{\sqrt{\hat{\sigma}^2_{\text{DPOLS}}\left((\mathbf{X}^t\widetilde{\mathbf{X}})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\widetilde{\mathbf{X}})^{-t}\right)_{i,i}}}\right|\right), \quad \text{where } T \sim t_{n-d}. \tag{7}$$

Again, the independence assumption is violated within environments.

We will, despite these shortcomings, be using the p-values above as one criterion to select parents or ancestors. Concretely we introduce methods OLS-pvals, POLS-pvals, and DPOLS-pvals, where we take the variable with the smallest p-value to be the most likely parent or ancestor.

3.5.5 PICP (Permuted ICP)

POLS introduced the idea of plugging $(\mathbf{X}, \widetilde{\mathbf{Y}})$ into usual OLS. Similarly, one can plug $(\mathbf{X}, \widetilde{\mathbf{Y}})$ into ICP. We call this method PICP and mention it as another candidate method that can be used in Problem A, but we won't discuss it any further, except for including it in a few simulations. Concretely, we take the variable with the smallest p-value to be the most likely parent or ancestor.

3.6 Baseline method for singletargets: mean-shift

Assume that we are in the singletargets setting and know the intervention target in each environment. If EY^k is different from EY^0, then the intervention target in the k'th environment must be an ancestor of Y^0; otherwise the expression for Y^k in terms of noise variables would be the same as for Y^0, so they would have the same mean. Thus, a reasonable method is to compare the average of $\mathbf{Y}^k$ (the Y-observations from the k'th environment) to the average of $\mathbf{Y}^0$ (the control Y-observations), and if the difference is sufficiently large (as measured, *e.g.*, by a t-test), then infer that the intervention target of the k'th environment is an ancestor of Y^0.

In our setup we know that $EY^0 = 0$, so instead of a t-test we do the following simplified version, which we call *mean-shift*. For all i, we give X_i^0 a weight as follows. If X_i^0 is not the intervention target in any environment, then give it weight 0. Else, let $A_i \subseteq \{1, \ldots, K\}$ be the set of all k where X_i^0 is the intervention target in the k'th environment, and, for all $k \in A_i$, calculate the absolute value of the average of $\mathbf{Y}^k$, that is, $a_k := \left| \frac{1}{n_k} \sum_{j=1}^{n_k} y^{k,j} \right|$. The average of these absolute averages (that is, $v_i := \frac{1}{\# A_i} \sum_{k \in A_i} a_k$) is the weight for X_i^0. We concretely take the variable with the largest weight to be the most likely parent or ancestor.

Note that we can use $\widetilde{\mathbf{Y}}$ instead of $\mathbf{Y}$ to calculate the weights without changing their distribution, since we don't use the specific index of a given observation, only its experiment number. Hence "mean-shift" also works in Problem A (but still only under the assumption of singletargets with known intervention targets).

4 Simulation experiments

We now present simulation experiments that investigate the performance of POLS and DPOLS, and compare them to OLS and other baseline methods. All code for simulating and analyzing data is available on GitHub at https://github.com/adamgorm/bsc-simulations.

4.1 Simulation of data

Since the methods may perform differently depending on the true underlying causal structure, we first simulate a large number of random DAGs where each DAG is simulated as follows. Say that we want to simulate a DAG with n_h hiddens, d observed covariates (X's), and a response Y, that is, a total of $1 + d + n_h$ variables. First a causal order $(\pi_1^{-1}, \ldots, \pi_{1+d+n_h}^{-1})$ of $\{1, \ldots, 1 + d + n_h\}$ is fixed. Then, for all $i \in \{1, \ldots, d + n_h\}$ and $j > i$, there is probability 0.4 of adding the edge $\pi_i^{-1} \to \pi_j^{-1}$. Each path coefficient is then sampled by choosing absolute size uniformly in $(0.1, 0.9)$ and sign by a fair coin flip. The first n_h nodes $(\pi_1^{-1}, \ldots, \pi_{n_h}^{-1})$ is H; Y is placed in the middle of the remaining nodes, that is, the node $\pi^{-1}_{\texttt{round}(1+(1+n_h+d+n_h)/2)}$; the remaining nodes become X.

The proof of Proposition 9 provides the idea (known as *ancestral sampling*; Peters et al., 2017) of how to sample from an SCM: first noise variables are simulated from the noise distribution, and then they are substituted into the structural assignments in causal order, to obtain a sample of the endogenous variables. Since[5]

$$(H^{k,i}, X^{k,i}, Y^{k,i}) = B(H^{k,i}, X^{k,i}, Y^{k,i}) + N^{k,i} + (\mathbf{0}, W^k, 0),$$

where $\mathbf{0}$ is a 0-vector of the same dimension as $H^{k,i}$, and $B = \begin{pmatrix} \mathbf{0} \cdots \mathbf{0} \\ A \\ (\gamma^t \; \beta^t \; 0) \end{pmatrix}$, we see that

$$(H^{k,i}, X^{k,i}, Y^{k,i}) = (I - B)^{-1}(N^{k,i} + (\mathbf{0}, W^k, 0)), \tag{8}$$

where invertibility of $I - B$ follows from the acyclic structure. So by simulating $N^{k,i}$ and W^k from their respective distributions and plugging into Eq. (8) we get a sample of $(H^{k,i}, X^{k,i}, Y^{k,i})$. For all environments $k \in \{1, \ldots, K\}$ we do this for all $i \in \{1, \ldots, n_k\}$ using the same W^k. For truly separate data, we repeat this process a second time (but reusing the already simulated values of W^k) to obtain a separate data set. For permuted separate data, we permute our observations from the first simulation within environments. See Algorithms 1 and 2 in Appendix C for a simplified version of our simulation process.

We have simulated DAGs with 30 X's and 30 H's (61 variables in total, including Y), and with 5 X's and 5 H's (11 variables in total, including Y). We always let N_X and N_Y be standard Gaussian. We let `sdh` denote the standard deviation of H, so $N_H \sim \mathcal{N}(0, \texttt{sdh}^2 \cdot I)$. In the alltargets setting, we do the same number of repetitions `no` in each environment, that is, we set $n_1 = \cdots = n_k = \texttt{no}$. We denote the total number of environments (which is above referred to by the symbol K) by `ne`, and we let $W_k \sim \mathcal{N}(0, \texttt{sdw}^2 \cdot I)$. In the singletargets setting we let `nxi` denote the total number of X's intervened on; the `nxi` targets for interventions are chosen randomly among all X. `nei` denotes the total number of different interventional settings for each of the intervention targets, and `no` denotes the number of observations to take from each of these interventional settings. `noc` denotes the number of control observations, that is, without any interventions. In the singletargets setting `sdw` denotes the standard deviation of the nonzero W_i^k. We simulated a total of 4.5 TB of data for the analyses performed in this thesis. See Tables 1 and 2 for a summary of the parameters, and Tables 3 and 4 for an overview of all simulations presented below.

4.2 Summary of methods

We explore the proposed methods (POLS, DPOLS, and PICP) and compare them to various baseline methods. OLS, POLS, and DPOLS calculate regression coefficients ($\hat{\beta}^{\text{OLS}}$, $\hat{\beta}^{\text{POLS}}$, and $\hat{\beta}^{\text{DPOLS}}$), and take the variable with the largest absolute estimated coefficient to be the most likely parent or ancestor. We also use a variant of these methods, called OLS-pvals, POLS-pvals, and DPOLS-pvals, that instead calculate p-values and take the variable with the smallest p-value to be the most likely parent or ancestor. For the ICP and PICP methods, the ranking is based on p-values. For the "mean-shift" method, the intervention target leading to the largest absolute mean shift of Y^0 is taken to be the most likely parent or ancestor. We also include

[5] When a and b are column vectors, we write (a, b) for the column vector obtained by stacking a on top of b, and $(a^t \; b^t)$ for its transpose. A, β and γ are the coefficients from Eqs. (1) and (2), and $\mathbf{0} \cdots \mathbf{0}$ corresponds to a matrix with number of rows equal to the dimension of $H^{k,i}$ and number of columns equal to the dimension of $(H^{k,i}, X^{k,i}, Y^{k,i})$.

Table 1: Parameters for alltargets setting

no	Number of observations per environment
ne	Number of environments
sdw	Standard deviation of the mean shifts W
sdh	Standard deviation of the hidden variables

Table 2: Parameters for singletargets setting

no	Number of observations per environment
nei	Number of environments per intervention target
nxi	Number of different intervention targets
noc	Number of control observations
sdw	Standard deviation of the mean shifts W
sdh	Standard deviation of the hidden variables

Fig.	# DAGs	$\#X$	$\#H$	no	ne	sdw	sdh
3, 4	1000	30	30	2	$\{50,\dots,15000\}$	7	5
3, 4	1000	30	30	10	$\{50,\dots,15000\}$	7	5
3, 4	min. 100	30	30	2	$\{25000,\dots,100000\}$	7	5
3, 4	min. 100	30	30	10	$\{25000,\dots,100000\}$	7	5
5	1000	30	30	$\{2,\dots,300\}$	500	7	5
5	min. 100	30	30	$\{650,1000\}$	500	7	5
6	1000	30	30	$\{2,\dots,5000\}$	5000/no	7	5
7	1000	30	30	2	2500	$\{1,\dots,100\}$	5
7	1000	30	30	10	500	$\{1,\dots,100\}$	5
8	1000	30	30	2	2500	7	$\{1,\dots,100\}$
8	1000	30	30	10	500	7	$\{1,\dots,100\}$
10	100	5	5	2	$\{50,\dots,20000\}$	7	5
10	100	5	5	10	$\{5,\dots,20000\}$	7	5

Table 3: alltargets simulations. For all choices of parameters, we simulate both truly separate data and permuted separate data.

Figure	# DAGs	# X	# H	no	nei	nxi	noc	sdw	sdh
9	min. 100	30	30	2	$\{50,\dots,2000\}$	15	$2\cdot$ nei	7	5
9	min. 100	30	30	10	$\{50,\dots,2000\}$	15	$10\cdot$ nei	7	5
9	min. 100	30	30	2	$\{50,\dots,2000\}$	30	$2\cdot$ nei	7	5
9	min. 100	30	30	10	$\{50,\dots,2000\}$	30	$10\cdot$ nei	7	5
11	100	5	5	2	$\{50,\dots,1000\}$	15	$2\cdot$ nei	7	5
11	100	5	5	10	$\{50,\dots,1000\}$	15	$10\cdot$ nei	7	5
11	100	5	5	2	$\{50,\dots,1000\}$	30	$2\cdot$ nei	7	5
11	100	5	5	10	$\{50,\dots,1000\}$	30	$10\cdot$ nei	7	5

Table 4: singletargets simulations. For all choices of parameters, we simulate both truly separate data and permuted separate data.

two baseline methods based on random guessing. The method “all-random” selects variables at random, and serves as a worst-case baseline method. Propositions 22 and 26 show that DPOLS is useful for finding parents. It is, however, still unclear what it does after having selected all parents; will it start selecting ancestors or guess at random? In order to assess this, we also include the baseline method “random-after-parents” that is directly given information about what the correct parents are, but guesses ancestors at random among the remaining variables. See Tables 5 and 6 for a summary of all methods used here.

Table 5: Novel methods

Name	Order of selection
POLS	The variable with the largest absolute $\hat{\beta}^{\text{POLS}}$-coefficient is the most likely parent or ancestor.
POLS-pvals	The variable with the smallest p-value from Eq. (6) is the most likely parent or ancestor.
DPOLS	The variable with the largest absolute $\hat{\beta}^{\text{DPOLS}}$-coefficient is the most likely parent or ancestor.
DPOLS-pvals	The variable with the smallest p-value from Eq. (7) is the most likely parent or ancestor.
PICP	The variable with the smallest p-value from PICP is the most likely parent or ancestor

Table 6: Baseline methods

Name	Order of selection
OLS	The variable with the largest absolute $\hat{\beta}^{\text{OLS}}$-coefficient is the most likely parent or ancestor.
OLS-pvals	The variable with the smallest p-value from Eq. (5) is the most likely parent or ancestor.
ICP	The variable with the smallest p-value from ICP is the most likely parent or ancestor
mean-shift	The intervention target leading to the largest absolute mean shift of Y^0 is the most likely parent or ancestor.
all-random	Selects at random.
random-after-parents	Selects the correct parents first, then selects the remaining variables in random order.

4.3 Evaluating the methods

For a given data set, the methods employed here allow us to rank all variables in terms of how likely they are to be parents or ancestors of Y^0, based on either the absolute magnitude of the regression coefficients (larger values are better) or the size of the p-values (smaller values are better). However, the methods do not suggest a specific set of such ancestors and parents. If we, for a given method, want to select the n most likely variables, we therefore take the n variables with the n largest coefficients, or the n variables with the n smallest p-values. To compare the performance of the different methods we use the area under the ROC curve (AUC) measure. A ROC (Receiver Operating Characteristic) curve is a plot of the true positive rate against the false positive rate for a classification method evaluated at several different classification thresholds; in our case, the classification threshold is the number of variables selected. Thus, a ROC curve for a given method and a given data set is created by first selecting 0 variables as parent or ancestor, then the 1 most highly ranked variable, then the two highest ranked, and so on, until all variables have been selected, and at each step calculating the true positive rate (TPR) and false positive rate (FPR). AUC is then calculated as the area between the horizontal axis and the ROC curve. Specifically, the TPR and FPR relating to parents or ancestors are computed as shown below, where $\mathrm{select}(\mathrm{method}, n)$ is the set of the n highest ranked variables for a specified method:

$$\begin{aligned}
\mathrm{TPR}_{\mathrm{pa}}(\mathrm{method}, n) &= \frac{\#(\mathrm{select}(\mathrm{method}, n) \cap \mathrm{pa}(Y^0))}{\#(X^0 \cap \mathrm{pa}(Y^0))} \\
\mathrm{FPR}_{\mathrm{pa}}(\mathrm{method}, n) &= \frac{\#(\mathrm{select}(\mathrm{method}, n) \setminus \mathrm{pa}(Y^0))}{\#(X^0 \setminus \mathrm{pa}(Y^0))} \\
\mathrm{TPR}_{\mathrm{anc}}(\mathrm{method}, n) &= \frac{\#(\mathrm{select}(\mathrm{method}, n) \cap \mathrm{anc}(Y^0))}{\#(X^0 \cap \mathrm{anc}(Y^0))} \\
\mathrm{FPR}_{\mathrm{anc}}(\mathrm{method}, n) &= \frac{\#(\mathrm{select}(\mathrm{method}, n) \setminus \mathrm{anc}(Y^0))}{\#(X^0 \setminus \mathrm{anc}(Y^0))}.
\end{aligned}$$

For each data set we construct two ROC curves for each method considered; one comparing against parents, and one comparing against ancestors. The ROC curve for parents is thus a plot of the points

$$\{(\mathrm{FPR}_{\mathrm{pa}}(\mathrm{method}, i), \mathrm{TPR}_{\mathrm{pa}}(\mathrm{method}, i)) : i \in \{0, \dots, d\}\}$$

connected by line segments, while the ancestor ROC curve is made from the points

$$\{(\mathrm{FPR}_{\mathrm{anc}}(\mathrm{method}, i), \mathrm{TPR}_{\mathrm{anc}}(\mathrm{method}, i)) : i \in \{0, \dots, d\}\}.$$

We perform multiple simulations for each choice of parameters (between 100 and 1000; see Tables 3 and 4), compute AUC for each simulation, and finally report the average AUC (and possibly the quartiles) for that choice of parameters.

4.4 Results

To test the performance of our methods, we first simulated large data sets from DAGs with 30 X's and 30 H's (see Tables 3 and 4). We don't include ICP and PICP in these simulations, since they are slow on the large data sets, and because Bonferroni correction for the 30 X's resulted in most p-values being practically equal to 1 (in small pilot experiments). For both random baseline methods we plot the average AUC, and a ribbon indicating the first and third quartiles of the AUC.

We first investigate the performance for different numbers of environments (see Figs. 3 and 4). In all of these settings our proposed methods, POLS and DPOLS, beat the baseline methods OLS and “all-random”. In particular, we see that POLS is a viable method for Problem A (causal discovery from unpaired data) since it beats “all-random”. Our experiments further affirm Proposition 26, since the average AUC of DPOLS converges to 1 as $\texttt{ne} \to \infty$ on truly separate data. However, the average AUC of DPOLS doesn’t seem to converge to 1 on permuted separate data (*i.e.*, data where the separate data set is obtained by permuting the rows of the first data set), so the equivalent of Proposition 26 for permuted separate data seems untrue generally. On truly separate data, DPOLS is better at finding ancestors than the baseline method “random-after-parents”, so it is even useful for finding some of the remaining ancestors after all parents have been selected. Selecting variables based on coefficient size beats selection based on p-values, and the performance of OLS-pvals and POLS-pvals decreases for a sufficiently large number of environments. DPOLS-pvals performs slightly worse than DPOLS, and on permuted data with 10 observations per environment it gets worse for a sufficiently large number of environments. The quartiles show that the distributions of the observed AUC for POLS, DPOLS, and OLS are heavily skewed with many large observations (see Fig. 4). The first quartiles of the AUC for POLS, DPOLS, and OLS are above the third quartile of “all-random”.

In Fig. 5 we investigate the performance for different numbers of observations per environment. The average AUC of DPOLS converges to 1 for truly separate data as $\texttt{no} \to \infty$. It also converges to 1 (and does so faster) on permuted data for $\texttt{no} \to \infty$, unlike what we saw in Fig. 3 for $\texttt{ne} \to \infty$. Again we find that DPOLS is better at selecting ancestors than if it started guessing at random after having selected all parents.

With a total of 5000 observations (that is, $\texttt{no} \cdot \texttt{ne} = 5000$), is it best to have 1 environment with all 5000 observations, or 2500 environments with 2 observations each, or somewhere in between, for instance 50 environments with 100 observations each? And how much does the performance of POLS and DPOLS vary over the different possible allocations? To investigate this, we simulate from different allocations of 5000 observations (see Fig. 6). At all allocations of the 5000 observations, POLS and DPOLS beat the “all-random” baseline method. When there are few environments with a large number of observations each, POLS and DPOLS are comparable to OLS, but for many environments with few observations each, DPOLS beats OLS for both ancestors and parents, and POLS beats OLS for ancestors.

In Fig. 7 we investigate the performance of POLS and DPOLS for different standard deviations of the mean shifts W. The average parent AUC of POLS, DPOLS, and OLS converge to 1 for $\texttt{sdw} \to \infty$, meaning that they are able to correctly identify parents if the mean shifts are sufficiently strong. Also, they all beat “random-after-parents”, so they are able to identify some extra ancestors after having selected all parents. When ordering is based on p-values, the methods also get better as sdw increases, but at a slower pace.

In Fig. 8 we see that OLS, POLS, and DPOLS become worse for $\texttt{sdh} \to \infty$. The differences between their average AUC converge to 0, but even at $\texttt{sdh} = 100$ they still beat “all-random”.

Finally, we investigate the performance of POLS and DPOLS in the singletargets setting (see Fig. 9). Both POLS and DPOLS beat random guessing, and in many of the settings they beat OLS. They beat the “mean-shift” baseline method when $\texttt{no} = 2$ and when $\texttt{nxi} = 15$. DPOLS seems to converge to selecting the right parents for $\texttt{nei} \to \infty$ on truly separate data with $\texttt{no} = 10$, but it requires a higher total number of observations than in the alltargets settings in Figs. 3 and 5.

To test the performance of the methods when there are fewer variables, we simulate data sets from 100 DAGs with 5 X's and 5 H's (see Tables 3 and 4). In the alltargets setup, our proposed methods, POLS and DPOLS, beat both baseline methods (OLS and "all-random"), and PICP performs better than ICP. POLS and DPOLS have average AUC close to 1 in all settings (see Fig. 10). In the singletargets setup, POLS and DPOLS both beat the baseline methods OLS and "all-random", and their performance is comparable to that of "mean-shift" (see Fig. 11).

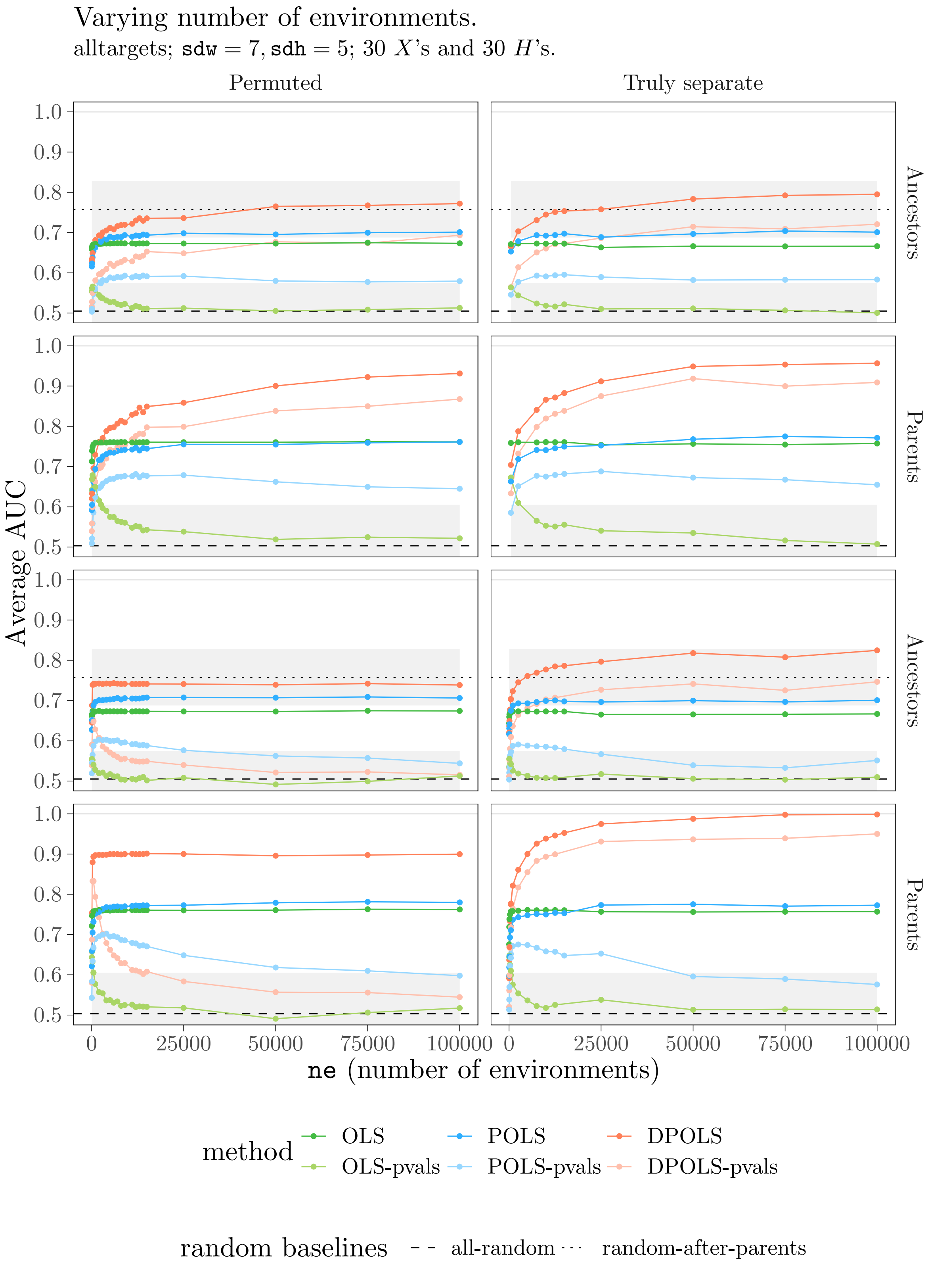


Figure 3: Our proposed methods, POLS and DPOLS, beat the baseline methods OLS and "all-random". Since POLS beats "all-random" it is a viable method for Problem A (causal discovery from unpaired data). DPOLS selects the correct parents asymptotically on truly separate data (the average AUC converges to 1), and is close to being as good on permuted separate data (with average AUC between 0.9 and 1). On truly separate data, we even see that DPOLS beats "random-after-parents", so it is able to select extra ancestors after having selected all parents.

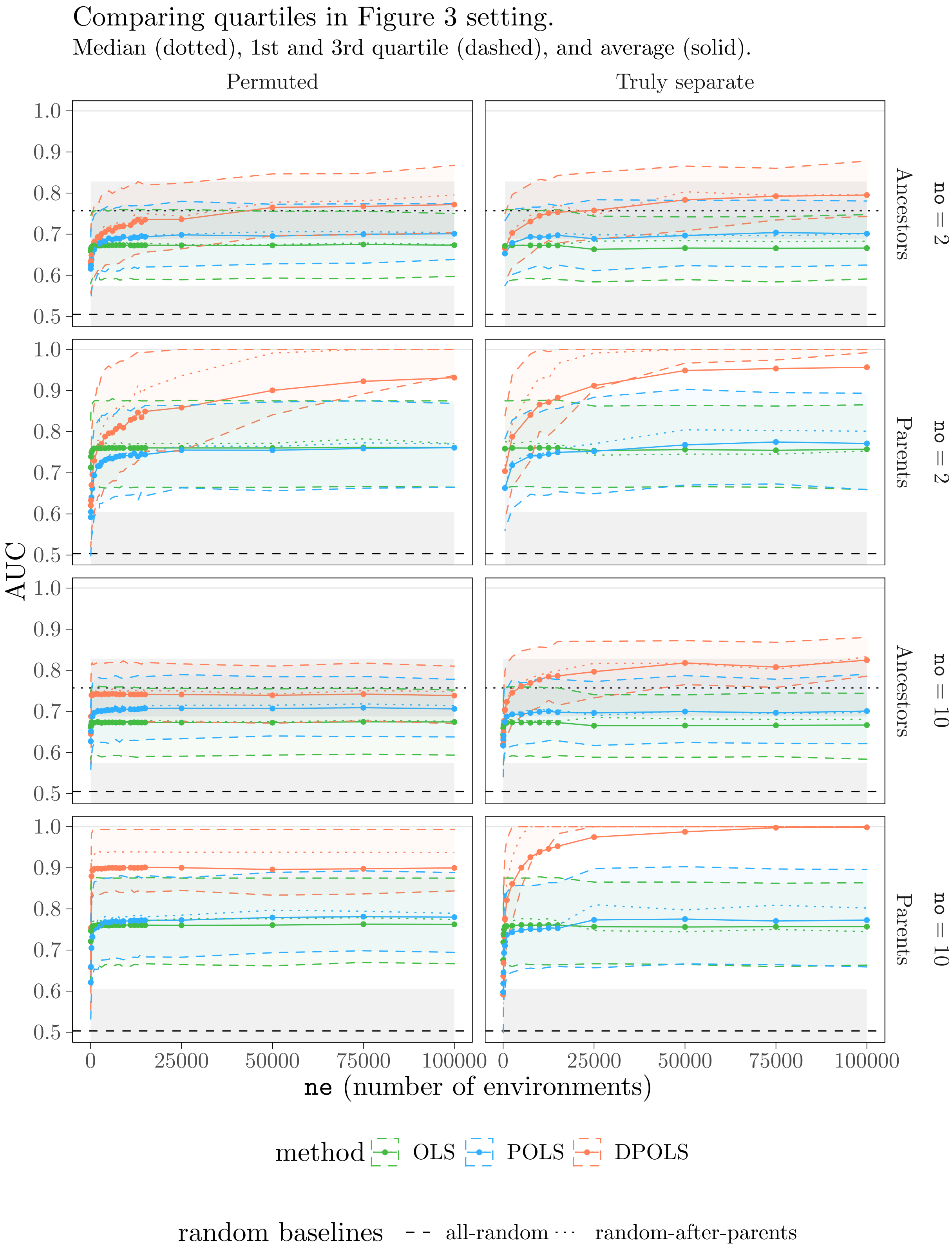


Figure 4: In this figure, we compare the distributions of observed AUC values by plotting their quartiles. The first quartile of AUC for our methods, POLS and DPOLS, are above the third quartile of AUC for random guessing. The quartiles of POLS beat the quartiles of OLS. The setting in this plot is the same as in Fig. 3.

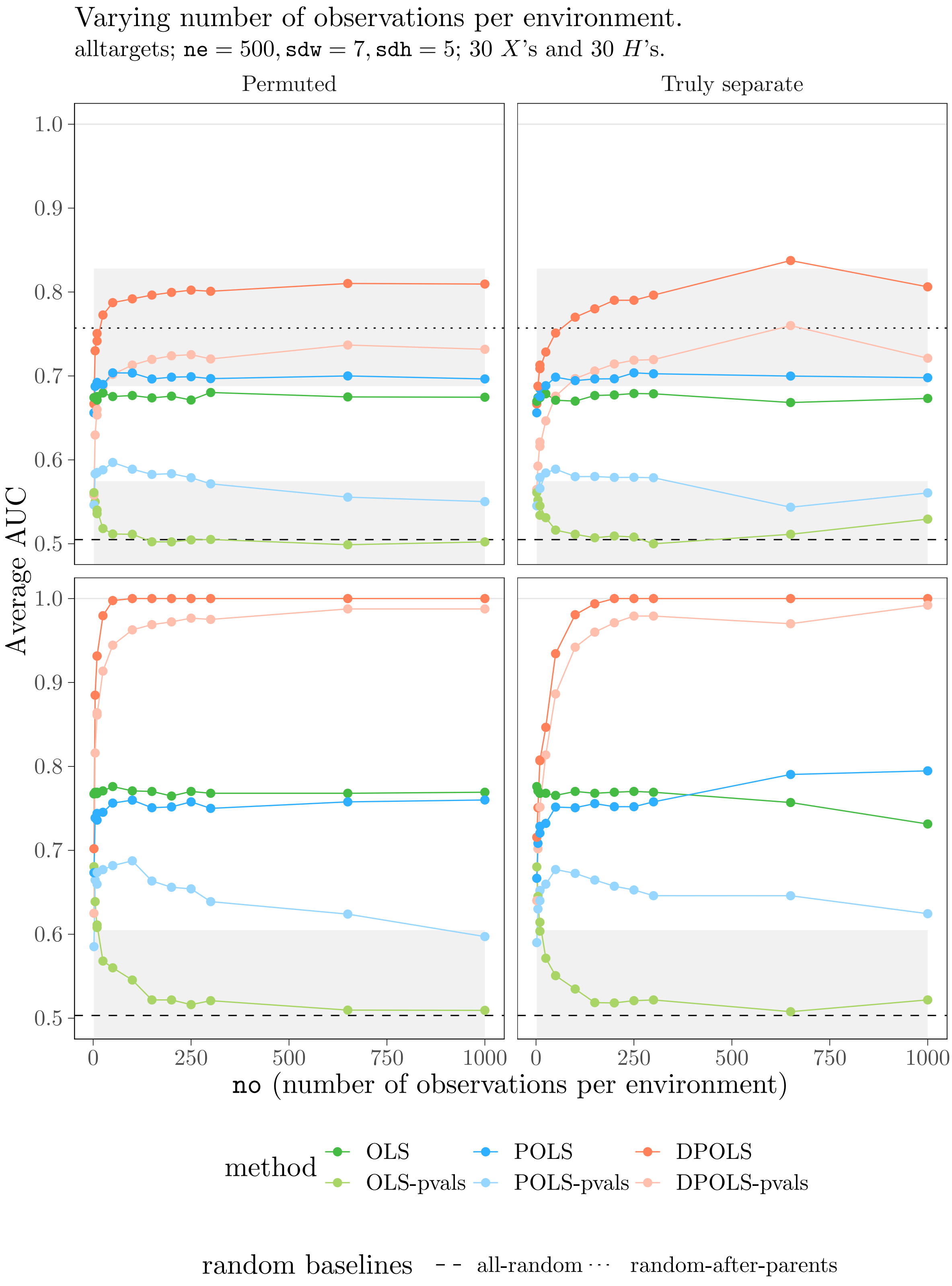


Figure 5: Our proposed methods, POLS and DPOLS, beat the baseline method "all-random". DPOLS converges to selecting the right parents for both permuted- and truly separate data; at the same time, it beats "random-after-parents", which means that it is able to select some extra ancestors, even after selecting all parents. POLS beats OLS at selecting ancestors.

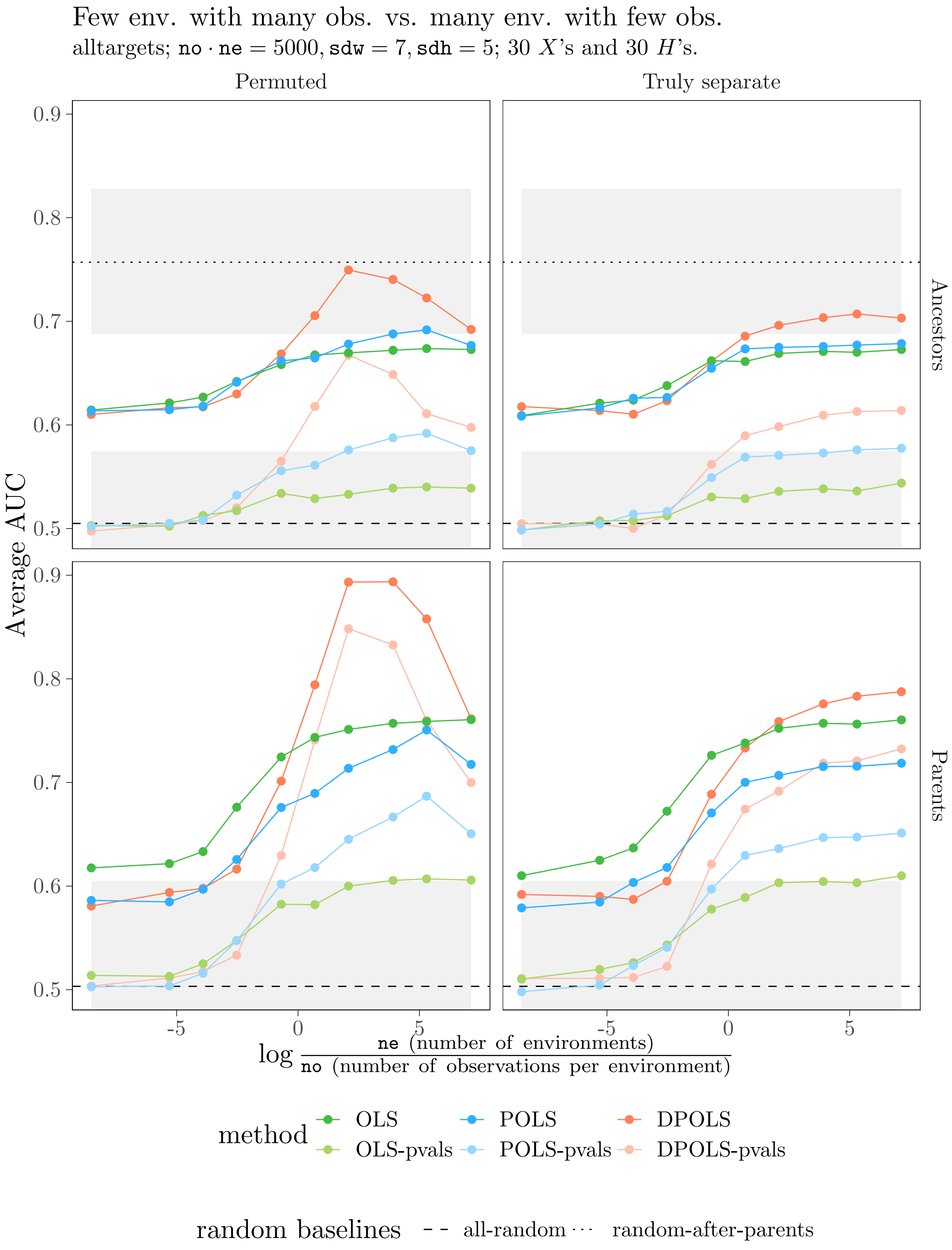


Figure 6: Our proposed methods, POLS and DPOLS, beat random guessing at all combinations of number of environments and number of observations per environment. The total number of observations is held constant at 5000, but it is varied between many environments with few observations per environment, and few environments with many observations per environment. For truly separate data, more environments are better; for permuted data, performance drops for the highest number of environments.

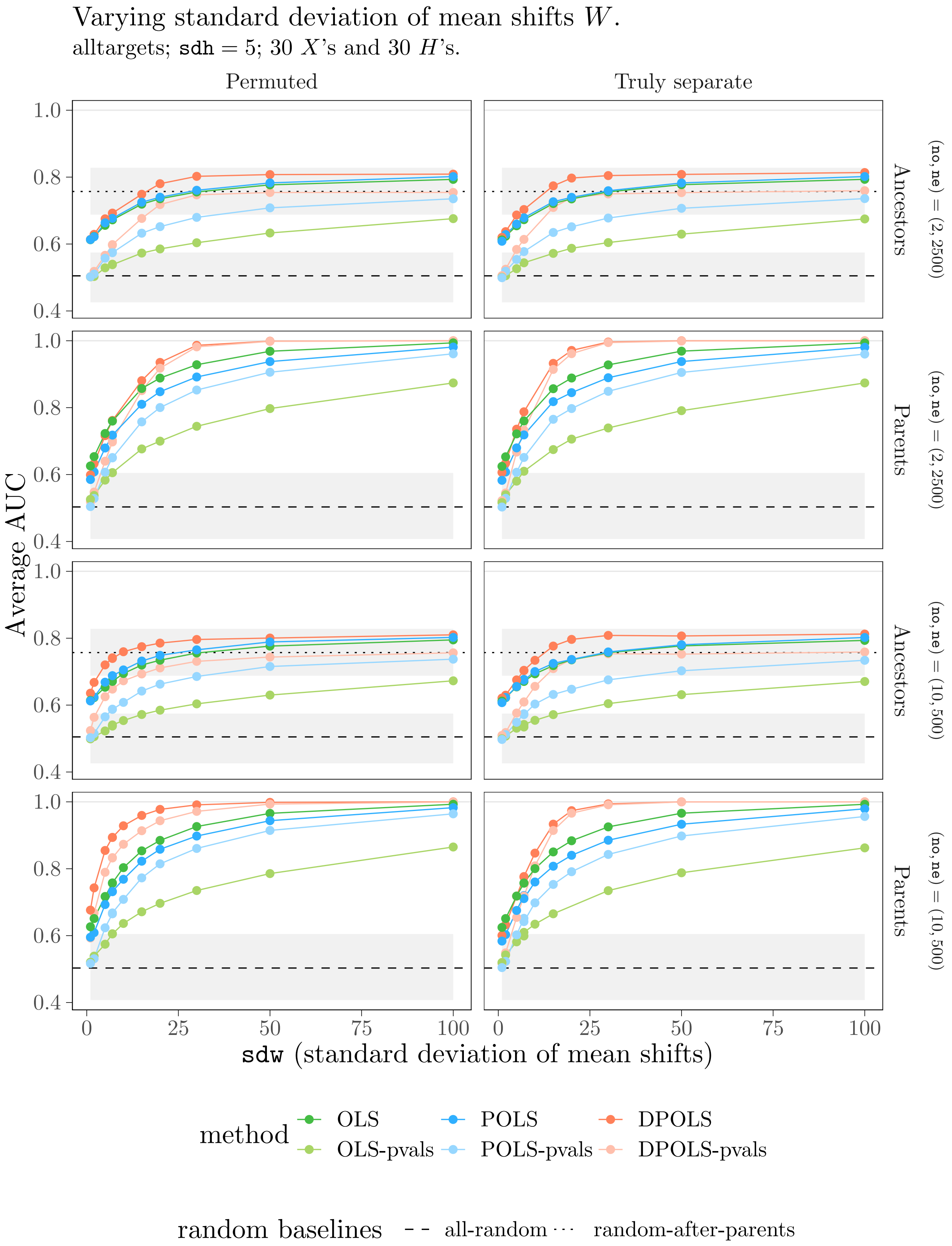


Figure 7: As expected, the methods perform better when the mean shifts have higher standard deviation. For sufficiently large `sdw`, they are all able to find the correct parents first, and are better at finding ancestors than if they started guessing randomly after having selected the parents.

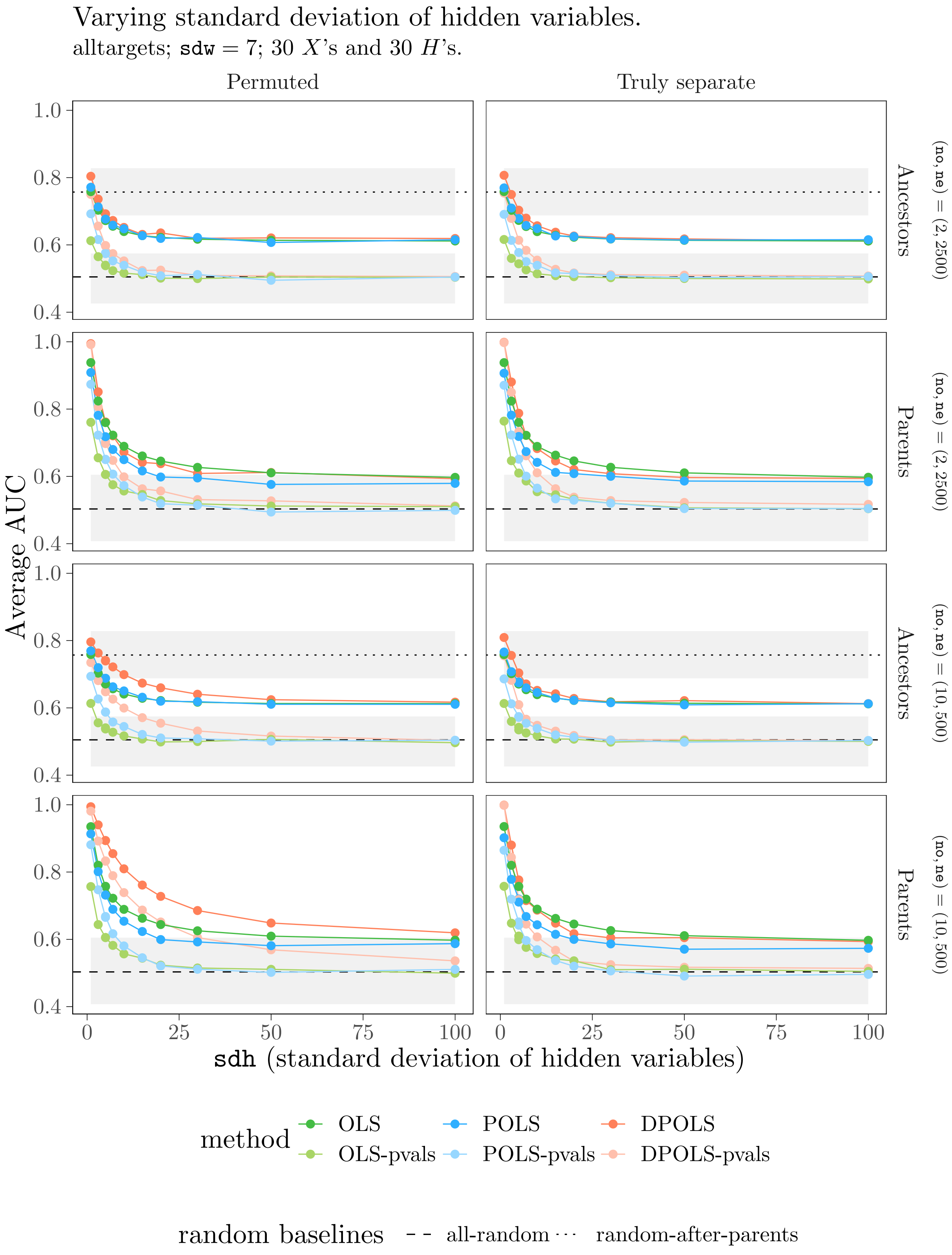


Figure 8: As expected, the methods perform worse when the hidden variables have higher standard deviation. Even for very large `sdh`, however, they all still perform better than random guessing. The differences in average AUC between OLS, POLS, and DPOLS seem to converge to 0 for `sdh` → ∞.

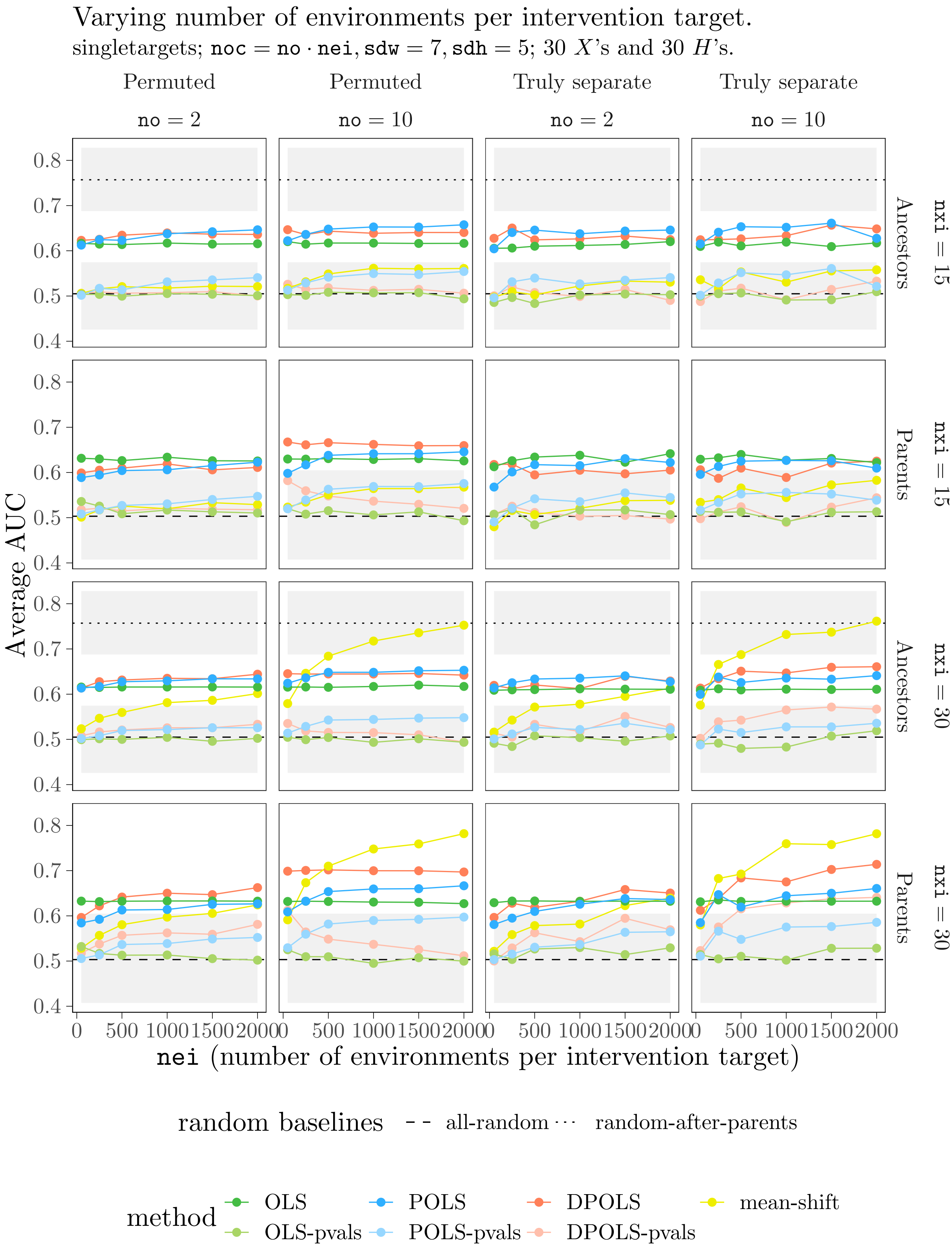


Figure 9: Both of our proposed methods, POLS and DPOLS, beat random guessing. OLS, POLS, and DPOLS have similar performance, and all perform worse in these singletargets settings than in the alltargets settings in Fig. 3. They beat the "mean-shift" baseline in many settings, for instance when not all X's are intervened on, or with two observations per environment.

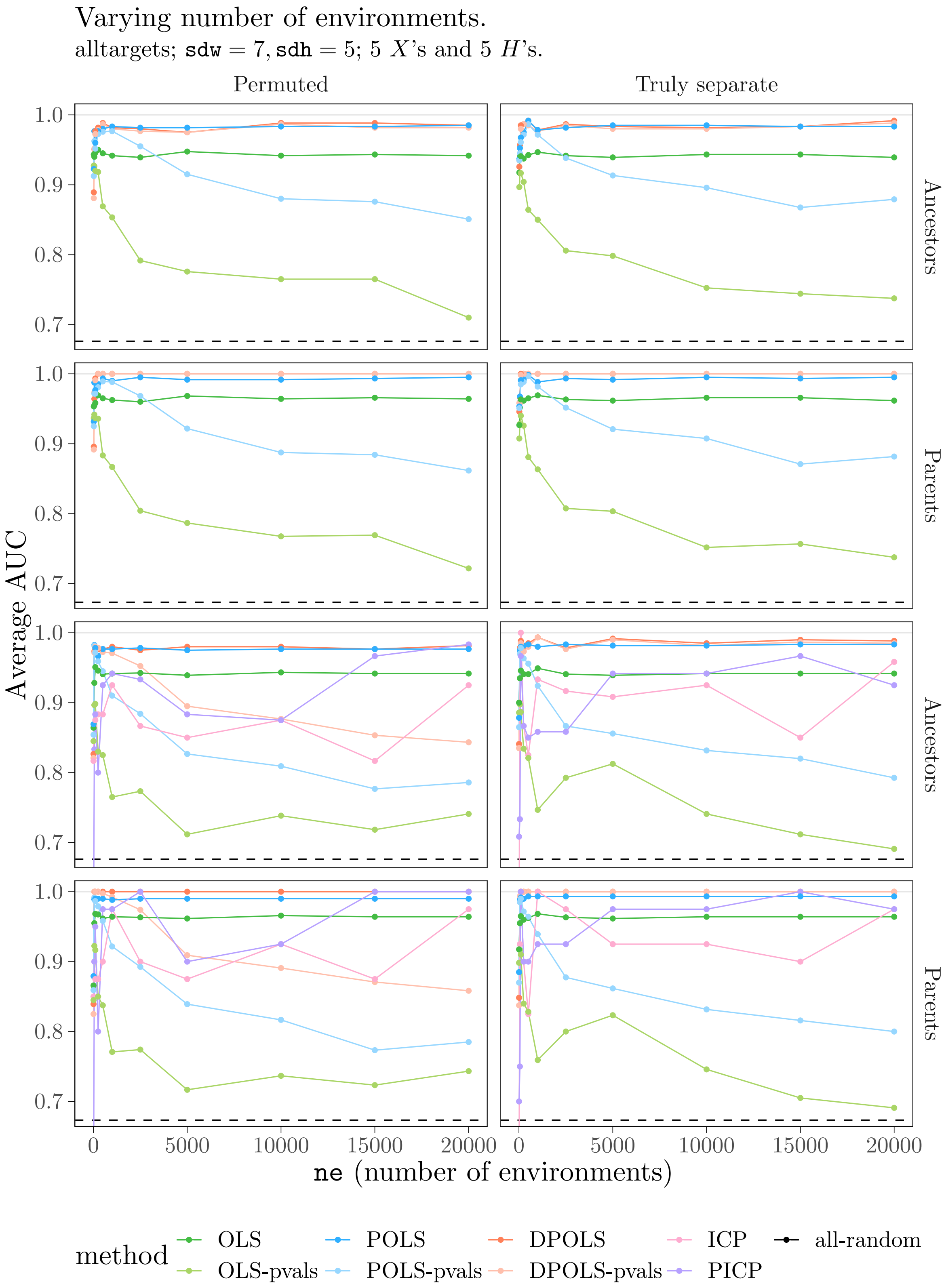


Figure 10: Both of our proposed methods, POLS and DPOLS, beat both baseline methods (OLS and "all-random"). They have average AUC around 1, but note that these are very small settings (where there is often only a single parent). PICP performs better than ICP.

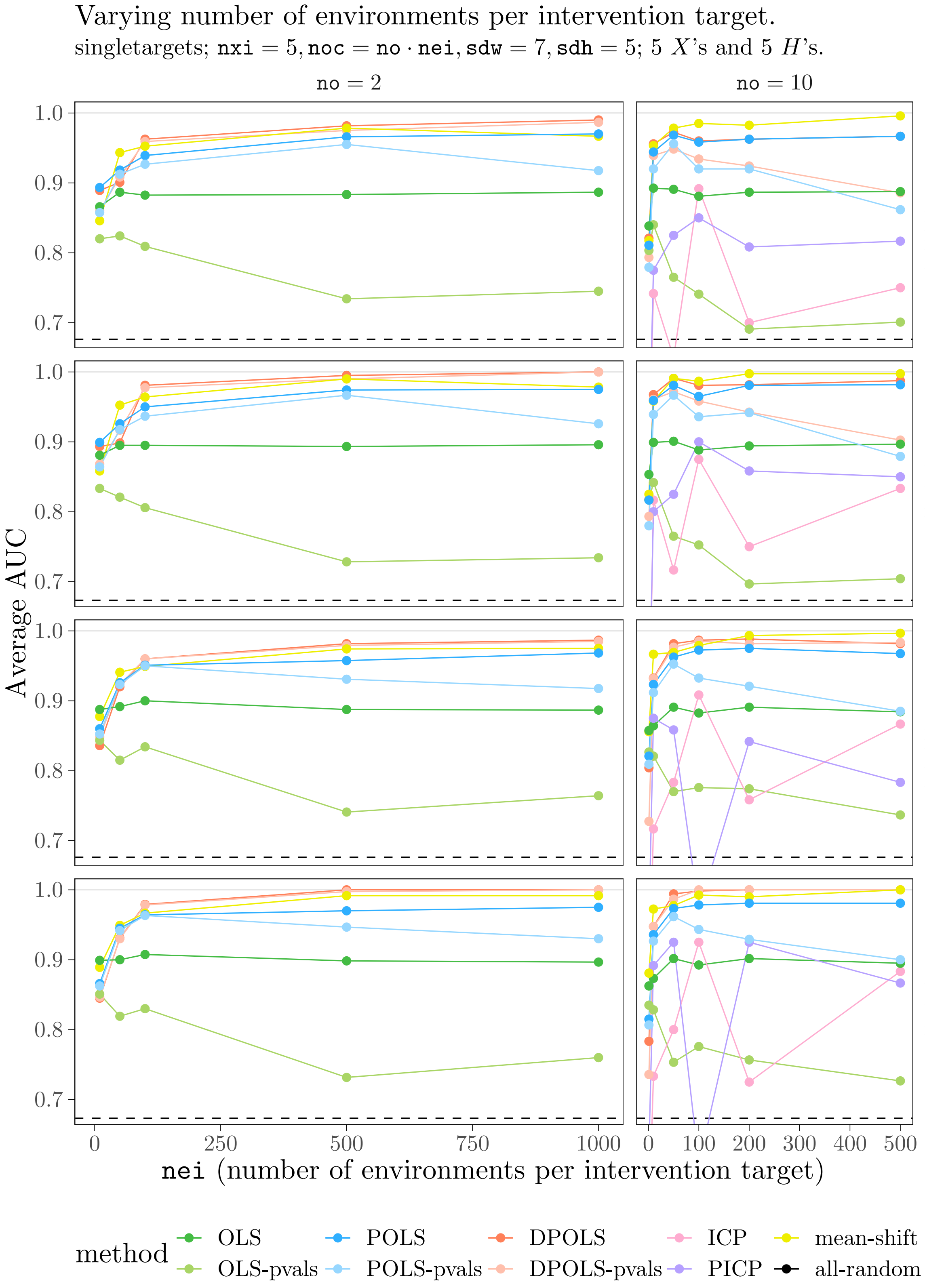


Figure 11: Both of our proposed methods, POLS and DPOLS, beat the baseline methods (OLS and "all-random"), while their performance is comparable to that of "mean-shift".

5 Discussion

We have seen that, in the alltargets setting, DPOLS will asymptotically select all or almost all parents, and that it is able to select some extra ancestors after this. Hence, it is a viable method for Problems B (causal discovery from truly separate data) and C (causal discovery from permuted separate data); it allows us to discover all direct causes in the presence of hidden variables and with unknown intervention targets without any randomization. Future work should provide more extensive theoretical guarantees for DPOLS, in particular in the setting of permuted separate data; for small sample sizes, we have seen that it performs better on permuted separate data than on truly separate data.

We have also seen that POLS is a viable method for Problem A (causal discovery from unpaired data), since it performs better than guessing at random. Future work should provide theoretical guarantees, but also consider whether there are other methods that may beat it in Problem A, perhaps using the idea of plugging $(\mathbf{X}, \widetilde{\mathbf{Y}})$ into a well-known method.

Finally, our presentation of the methods does not lend itself to practical use in real situations, since we have not given any way of deciding how many variables to select; we only provide an ordering from most likely to be parent or ancestor to least likely. We have not spent much time on this problem, but a few unincluded exploratory simulations indicate that the p-values we have provided do not lead to the desired Type I error rate. Future work should find a way to decide significance in practice. Another discussion of practical importance is whether the assumptions are reasonable in real-world scenarios. For instance, is it reasonable that we only intervene on observed variables, and that all experiments in the same environment have the exact same mean shifts?

Appendix A Population least squares

In this appendix we show that $\beta^{\mathrm{OLS}} = \arg\min_\beta \sum_{k=1}^K E(Y^k - \beta^t X^k)^2$. Exchanging differentiation and expectation, and using the chain rule, yields

$$\begin{aligned} D_\beta \sum_{k=1}^K E(Y^k - \beta^t X^k)^2 &= \sum_{k=1}^K E(2(Y^k - \beta^t X^k)(-X^t)) \\ &= 2\sum_{k=1}^K (\beta^t E(X^k (X^k)^t) - E(Y^k (X^k)^t)) \\ &= 2\beta^t \sum_{k=1}^K \mathrm{cov}(X^k) - 2\sum_{k=1}^K \mathrm{cov}(X^k, Y^k). \end{aligned}$$

Hence, the only solution to $D_\beta \sum_{k=1}^K E(Y^k - \beta^t X^k)^2 = 0$ is

$$\beta = \left(\sum_{k=1}^K \mathrm{cov}(X^k)\right)^{-1} \sum_{k=1}^K \mathrm{cov}(X^k, Y^k).$$

Since

$$D_\beta^2 \sum_{k=1}^K E(Y^k - \beta^t X^k)^2 = 2\sum_{k=1}^K \mathrm{cov}(X^k) \succ 0$$

the function

$$\beta \mapsto \sum_{k=1}^K E(Y^k - \beta^t X^k)^2$$

is strictly convex, so it attains its unique minimum in the unique stationary point

$$\beta^{\mathrm{OLS}} := \left(\sum_{k=1}^K \mathrm{cov}(X^k)\right)^{-1} \sum_{k=1}^K \mathrm{cov}(X^k, Y^k).$$

Appendix B Modified strong law of large numbers for strong consistency

In this appendix we show that $\hat{\beta}^{\mathrm{OLS}} \overset{\text{a.s}}{\to} \beta^{\mathrm{OLS}}$ for $K \to \infty$ in the alltargets setting with truly permuted data (and the same results for POLS and DPOLS follow by similar arguments). Let $a, b \in \{1, \dots, d\}$ be arbitrary, and let $Z_k := \frac{1}{n_k} \sum_{i=1}^{n_k} X_a^{k,i} X_b^{k,i}$ for all $k \in \mathbb{N}$. We need that

$$\frac{1}{K} \sum_{k=1}^K Z_k \overset{\text{a.s}}{\to} \mathrm{cov}(X_a^1, X_b^1) \quad \text{for } K \to \infty. \tag{9}$$

By Eq. (4), this will give that $\mathbf{X}^t\mathbf{X} \overset{\text{a.s}}{\to} \mathrm{cov}(X^1)$ and a similar argument (where X_b is replaced by Y in (9)) gives $\mathbf{X}^t\mathbf{Y} \overset{\text{a.s}}{\to} \mathrm{cov}(X^1, Y^1)$, and thus, by continuity, that $\hat{\beta}^{\mathrm{OLS}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \overset{\text{a.s}}{\to} (\mathrm{cov}(X^1))^{-1}\mathrm{cov}(X^1, Y^1)$ for $K \to \infty$, as desired.

So let us prove (9). If all n_k are equal (that is, if there is $\phi \in \mathbb{R}$ such that $n_k = \phi$ for all $k \in \mathbb{N}$), then $Z_1, Z_2, \dots$ are iid. so the strong law of large numbers directly gives (9). If n_k are not all equal, then $Z_1, Z_2, \dots$ are still independent, but not identically distributed, so we have to do a small variation of the strong law of large numbers, as follows.

We will modify a proof of the strong law of large numbers from Hansen (2020) to suit our needs. Let $U_k := Z_k - \text{cov}(X_a^1, X_b^1)$ for all $k \in \mathbb{N}$. If we can prove that

$$\sum_{k=1}^{\infty} V\left(\frac{U_k}{k}\right) < \infty \tag{10}$$

then the Khintchine-Kolmogorov Theorem gives

$$\sum_{k=1}^{K} \frac{U_k}{k} \overset{\text{a.s}}{\to} V \quad \text{for } K \to \infty$$

for some limit variable V. Kronecker's Lemma then gives

$$\frac{1}{K}\sum_{k=1}^{K} U_k \overset{\text{a.s}}{\to} 0 \quad \text{for } K \to \infty$$

which gives (9).

We now prove (10). We have

$$\begin{aligned}\sum_{k=1}^{\infty} V\left(\frac{U_k}{k}\right) &= \sum_{k=1}^{\infty} \frac{1}{k^2 n_k^2} V\left(\sum_{i=1}^{n_k} X_a^{k,i} X_b^{k,i}\right) \\ &= \sum_{k=1}^{\infty} \frac{1}{k^2 n_k^2}\left(\sum_{i=1}^{n_k} V(X_a^{k,i} X_b^{k,i}) + \sum_{i\neq j} \text{cov}(X_a^{k,i} X_b^{k,i}, X_a^{k,j} X_b^{k,j})\right).\end{aligned}$$

Since $X^{k,1}, \dots, X^{k,d}$ are identically distributed we get

$$\sum_{i=1}^{n_k} V(X_a^{k,i} X_b^{k,i}) = n_k V(X_a^{k,1} X_b^{k,1}).$$

Since $(X^{k,i}, X^{k,j}) \overset{\mathcal{D}}{=} (X^{k,i'}, X^{k,j'})$ for all choices of $i \neq j$ and $i' \neq j'$ we get

$$\sum_{i\neq j} \text{cov}(X_a^{k,i} X_b^{k,i}, X_a^{k,j} X_b^{k,j}) = (n_k^2 - n_k)\text{cov}(X_a^{k,1} X_b^{k,1}, X_a^{k,2} X_b^{k,2}).$$

Since $X^{1,i}, X^{2,i}, \dots$ are identically distributed we get

$$\frac{1}{n_k^2}\left| n_k V(X_a^{k,1}, X_b^{k,1}) + (n_k^2 - n_k)\text{cov}(X_a^{k,1} X_b^{k,1}, X_a^{k,2} X_b^{k,2})\right| \leq V(X_a^{1,1} X_b^{1,1}) + |\text{cov}(X_a^{1,1} X_b^{1,1}, X_a^{1,2} X_b^{1,2})|$$

which doesn't depend on k, so

$$\sum_{k=1}^{\infty} V\left(\frac{U_k}{k}\right) = \sum_{k=1}^{\infty} \frac{1}{k^2 n_k^2}\left(n_k V(X_a^{k,1} X_b^{k,1}) + (n_k^2 - n_k)\text{cov}(X_a^{k,1} X_b^{k,1}, X_a^{k,2} X_b^{k,2})\right) < \infty.$$

This finishes the proof for OLS. A similar argument works for POLS and DPOLS; you just have to insert $\widetilde{\mathbf{X}}$, $\widetilde{X}$, $\widetilde{\mathbf{Y}}$, and $\widetilde{Y}$ in the appropriate places.

Appendix C Simulating data

This appendix contains algorithms illustrating a simplified version of how we simulate truly separate data and permuted separate data.

Algorithm 1: Simulating truly separate data

Input :

- A list $\mathcal{B}$ of matrices giving path coefficients (each entry being a B-matrix for Eq. (8))
- The distributions of mean shifts (P_{W^k}) and the distributions of noise variables (P_{N^k}), for all environments $k \in \{1, \dots, K\}$.
- The number of observations $n_1, \dots, n_K$ for each environment.

Output: A list containing, for all B in $\mathcal{B}$, a realisation $(\mathbf{X}, \mathbf{Y}, \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})_B$ from the k W-bridged SCMs with path coefficients given by B, noise distributions P_{N^k} and mean shift distributions P_{W^k}, and with $n_1, \dots, n_K$ observations in each environment.

```
for $B$ in $\mathcal{B}$ do
    for $k = 1$ to $K$ do
        $w^k \leftarrow \texttt{sampleFrom}(P_{W^k})$;
        for $i = 1$ to $n_k$ do
            $n^{k,i} \leftarrow \texttt{sampleFrom}(P_{N^k})$;
            $(h^{k,i}, x^{k,i}, y^{k,i}) \leftarrow (I - B)^{-1}(n^{k,i} + (\mathbf{0}, w^k, 0))$;
            $\widetilde{n}^{k,i} \leftarrow \texttt{sampleFrom}(P_{N^k})$;
            $(\widetilde{h}^{k,i}, \widetilde{x}^{k,i}, \widetilde{y}^{k,i}) \leftarrow (I - B)^{-1}(\widetilde{n}^{k,i} + (\mathbf{0}, w^k, 0))$;
        end
        $(\mathbf{X}^k, \mathbf{Y}^k) \leftarrow \begin{pmatrix} x^{k,1} & y^{k,1} \\ \vdots & \vdots \\ x^{k,n_k} & y^{k,n_k} \end{pmatrix}$;
        $(\widetilde{\mathbf{X}}^k, \widetilde{\mathbf{Y}}^k) \leftarrow \begin{pmatrix} \widetilde{x}^{k,1} & \widetilde{y}^{k,1} \\ \vdots & \vdots \\ \widetilde{x}^{k,n_k} & \widetilde{y}^{k,n_k} \end{pmatrix}$;
    end
    $(\mathbf{X}, \mathbf{Y}, \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})_B \leftarrow \begin{pmatrix} \mathbf{X}^1 & \mathbf{Y}^1 & \widetilde{\mathbf{X}}^1 & \widetilde{\mathbf{Y}}^1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{X}^k & \mathbf{Y}^k & \widetilde{\mathbf{X}}^k & \widetilde{\mathbf{Y}}^k \end{pmatrix}$;
end
```

Algorithm 2: Simulating permuted separate data

```
Input   :
    • A list $\mathcal{B}$ of matrices giving path coefficients (each entry being a $B$-matrix for Eq. (8))
    • The distributions of mean shifts ($P_{W^k}$) and the distributions of noise variables ($P_{N^k}$),
      for all environments $k \in \{1, \dots, K\}$.
    • The number of observations $n_1, \dots, n_K$ for each environment.
Output: A list containing, for all $B$ in $\mathcal{B}$, a realisation $(\mathbf{X}, \mathbf{Y}, \check{\mathbf{X}}, \check{\mathbf{Y}})_B$ from the $k$
        emulated $W$-bridged SCMs with path coefficients given by $B$, noise
        distributions $P_{N^k}$ and mean shift distributions $P_{W^k}$, and with $n_1, \dots, n_K$
        observations in each environment.
for $B$ in $\mathcal{B}$ do
    for $k = 1$ to $K$ do
        $w^k \leftarrow$ sampleFrom($P_{W^k}$);
        for $i = 1$ to $n_k$ do
            $n^{k,i} \leftarrow$ sampleFrom($P_{N^k}$);
            $(h^{k,i}, x^{k,i}, y^{k,i}) \leftarrow (I - B)^{-1}(n^{k,i} + (\mathbf{0}, w^k, 0))$;
        end
        $(\mathbf{X}^k, \mathbf{Y}^k) \leftarrow \begin{pmatrix} x^{k,1} & y^{k,1} \\ \vdots & \vdots \\ x^{k,n_k} & y^{k,n_k} \end{pmatrix}$;
        $\sigma \leftarrow$ getRandomPermutation($\{1, \dots, n_k\}$);
        $(\check{\mathbf{X}}^k, \check{\mathbf{Y}}^k) \leftarrow$ permuteEachColumn($(\mathbf{X}^k, \mathbf{Y}^k), \sigma$);
    end
    $(\mathbf{X}, \mathbf{Y}, \check{\mathbf{X}}, \check{\mathbf{Y}})_B \leftarrow \begin{pmatrix} \mathbf{X}^1 & \mathbf{Y}^1 & \check{\mathbf{X}}^1 & \check{\mathbf{Y}}^1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{X}^K & \mathbf{Y}^K & \check{\mathbf{X}}^K & \check{\mathbf{Y}}^K \end{pmatrix}$;
end
```

References

Barnett, V. (1999). *Comparative Statistical Inference.* John Wiley & Sons, third edition.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of Structural Causal Models with Cycles and Latent Variables. *arXiv:1611.06221v5.*

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis.* CRC Press, third edition.

Hansen, E. (2020). *Stochastic Processes.* Institute of Mathematics, University of Copenhagen, first edition.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press.

Lauritzen, S. L. (2019). *Lectures on Graphical Models.* Polyteknisk Boghandel, third edition.

Lauritzen, S. L. (2021). *Basic Mathematical Statistics.* Polyteknisk Boghandel.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer Texts in Statistics. Springer, New York, second edition.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, second edition.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):947–1012.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press.

Reichenbach, H. (1956). *The Direction of Time.* University of California Press, Los Angeles.

Schopenhauer, A. (2004). *Essays of Schopenhauer.* Number 11945 in Project Gutenberg. September 26, 2020 edition.

Shao, J. (2003). *Mathematical Statistics.* Springer Texts in Statistics. Springer, New York, second edition.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search.* The MIT Press, second edition.