Weighting-Based Identification and Estimation in Graphical Models of Missing Data

Anna Guo and Razieh Nabi

Department of Biostatistics and Bioinformatics, Emory University

February 11, 2026

Abstract

We propose a constructive algorithm for identifying complete data distributions in graphical models of missing data. The complete data distribution is unrestricted, while the missingness mechanism is assumed to factorize according to a conditional directed acyclic graph. Our approach follows an interventionist perspective in which missingness indicators are treated as variables that can be intervened on. A central challenge in this setting is that sequences of interventions on missingness indicators may induce and propagate selection bias, so that identification can fail even when a propensity score is invariant to available interventions. To address this challenge, we introduce a tree-based identification algorithm that explicitly tracks the creation and propagation of selection bias and determines whether it can be avoided through admissible intervention strategies. The resulting tree provides both a diagnostic and a constructive characterization of identifiability under a given missingness mechanism. Building on these results, we develop recursive inverse probability weighting procedures that mirror the intervention logic of the identification algorithm, yielding valid estimating equations for both the missingness mechanism and functionals of the complete data distribution. Simulation studies and a real-data application illustrate the practical performance of the proposed methods. An accompanying `R` package, `flexMissing`, implements all proposed procedures.

Keywords: Missing not at random, selection bias, missing data DAGs, causal inference, inverse probability weighting, estimating equations.

1 Introduction

Data analysis across scientific disciplines is frequently complicated by systematically missing observations. While missingness is often assumed to be missing-completely-at-random (MCAR) or missing-at-random (MAR), these assumptions are frequently violated in practice, as the probability of missingness may depend on partially observed or unobserved variables. Such missing-not-at-random (MNAR) mechanisms can lead to biased inference if not properly accounted for (Rubin 1976, Little & Rubin 2002).

Despite their prevalence, MNAR mechanisms are difficult to handle because the underlying complete data distribution cannot, in general, be expressed as a function of the observed data distribution. A canonical example is self-censoring or self-masking, where a variable directly influences its own probability of being missing, rendering the target distribution non-identifiable without further assumptions. In non-identified settings, common approaches include imposing parametric or semiparametric structures (Wu & Carroll 1988, Little & Rubin 2002, Wang et al. 2014, Sun et al. 2018, Sportisse et al. 2020, Guo et al. 2023), conducting sensitivity analyses (Rotnitzky et al. 1998, Scharfstein & Irizarry 2003, Nabi et al. 2024), or deriving partial identification bounds (Horowitz & Manski 2000, Manski 2005).

At the same time, a growing body of work has identified MNAR mechanisms by imposing restrictions on the missingness selection mechanism. Examples include the permutation model (Robins 1997), the block-conditional MAR model (Zhou et al. 2010), itemwise conditionally independent nonresponse and no self-censoring models (Sadinle & Reiter 2017, Shpitser 2016), and discrete choice models (Tchetgen et al. 2018). These models differ in the specific restrictions imposed on the missingness mechanism, while the corresponding identification arguments are agnostic to assumptions on the complete data distribution.

A recent line of work studies missing data models by drawing parallels with causal graphical

models with hidden variables (Tian & Pearl 2002, Shpitser & Pearl 2006, Bhattacharya et al. 2022, Richardson et al. 2023). In this framework, a directed acyclic graph encodes assumptions about the full law, including both the complete data law, referred to as the target law, and the missingness mechanism (Glymour 2006, Daniel et al. 2012, Martel García 2013, Mohan et al. 2013, Thoemmes & Rose 2014, Shpitser et al. 2015, Bhattacharya et al. 2019, Nabi et al. 2020, Mohan & Pearl 2021, Nabi et al. 2023). Identification is then formulated from an interventionist perspective, in which missingness indicators are treated as intervention nodes. Target law identification then is expressed as a sequence of reweighting operations that adjust for selection bias induced by conditioning on observed cases.

Bhattacharya et al. (2019) showed that causal identification strategies are insufficient for missing data models, because they do not adequately account for selection bias induced by interventions. They instead characterized identification via admissible sequences of indicator interventions defined by a partial order, rather than the total order common in causal settings. While this framework clarifies key identification requirements and structural barriers due to selection bias, it offers no constructive method for assessing identifiability in a given model or for developing estimation and inference procedures.

In this paper, we develop a general and tractable framework for identification, estimation, and inference in a broad class of missing data models. We impose no restrictions on the target law and instead model the missingness mechanism using a conditional directed acyclic graph. Our approach yields a constructive identification strategy that explicitly accounts for selection bias induced by conditioning on observed cases and clarifies when such bias can be avoided through admissible intervention strategies on missingness indicators.

Our central contribution is a tree-based identification algorithm that tracks the creation and propagation of selection bias across sequences of interventions on missingness indicators.

Rather than returning a binary identifiability verdict, the algorithm produces explicit identification rules for propensity scores, potentially on restricted evaluation sets, and determines whether these rules jointly suffice to identify the target distribution. When target law identification is not achievable, the same tree structure can be queried to specify which subsets of propensity scores, together with their required evaluations, are sufficient to identify and estimate particular functionals of the target law.

Building on these identification results, we develop general estimation and inference procedures that mirror the intervention logic of the identification trees. For functionals identified by the identifiable components of the missingness mechanism, we propose recursive inverse probability weighting estimators that respect the same admissibility constraints required for identification and yield valid estimating equations for both the missingness mechanism and target functionals. We establish large-sample properties and demonstrate practical performance through a real-data application and simulation studies, including empirical comparisons with classical approaches such as the EM algorithm and multiple imputation.

The paper is organized as follows. Section 2 introduces notation and the graphical framework for missing data models. Section 3 provides intuition for identification through examples. The identification algorithm is presented in Section 4, followed by weighting based estimation and inference in Section 5. Simulation studies appear in Section 6, and Section 7 presents an application to survey data. Section 8 concludes, with proofs in the appendix.

2 Problem Setup and Notation

We begin by introducing notation and missing data assumptions. We omit variables that are always observed; all results extend directly to settings with fully observed covariates.

Let $X = (X_1, \ldots, X_K)^T$ be a random vector with joint distribution $p(X)$, termed the *target*

law, belonging to a model $\mathcal{M}_X$. Let $R = (R_1, \ldots, R_K)^T$ denote binary missingness indicators, where $R_k = 1$ if X_k is observed and $R_k = 0$ if X_k is missing, with conditional distribution $p(R \mid X)$, termed the *missingness mechanism*, in model $\mathcal{M}_{R \mid X}$. The joint distribution $p(X, R)$, termed the *full law*, lies in the product model $\mathcal{M} = \mathcal{M}_X \otimes \mathcal{M}_{R \mid X}$. The observed data are a coarsened version of X: define $X^* = (X_1^*, \ldots, X_K^*)^T$ by $X_k^* = X_k$ if $R_k = 1$ and $X_k^* =$ "?" otherwise. The distribution $p(X^*, R)$ is called the *observed data law*. Following (Nabi et al. 2023), we view each X_k as the counterfactual value of X_k^* had it been fully observed, or equivalently under the intervention $R_k = 1$.

We consider missing data models $\mathcal{M} = \mathcal{M}_X \otimes \mathcal{M}_{R \mid X}$, where the target model $\mathcal{M}_X$ is unrestricted and the missingness mechanism $\mathcal{M}_{R \mid X}$ is constrained by graphical assumptions. In particular, we assume $p(R \mid X)$ factorizes according to a conditional directed acyclic graph (DAG) $\mathcal{G}$, that is, $p(R \mid X) = \prod_{R_k \in R} p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k))$, where $\mathrm{pa}_{\mathcal{G}}(R_k)$ denotes the parents of R_k in $\mathcal{G}$. No graphical assumptions are imposed on the target law $p(X)$; $\mathcal{M}_X$ may correspond to a hidden-variable DAG, a Markov random field, or any other model that need not admit a graphical representation. Thus, we use $\mathcal{G}$ solely to encode restrictions on $p(R \mid X)$.

The DAG $\mathcal{G}$ follows standard missing data DAG (mDAG) conventions: it is acyclic and contains no edges from missingness indicators R to variables in X (Mohan et al. 2013). For each $R_k \in R$, let $\mathrm{de}_{\mathcal{G}}(R_k)$ denote its descendants in $\mathcal{G}$, including R_k, and define the non-descendants as $\mathrm{nd}_{\mathcal{G}}(R_k) = X \cup R \setminus \mathrm{de}_{\mathcal{G}}(R_k)$. Standard graphical d-separation rules apply (Pearl 2009). All mDAGs in the remainder of the paper are assumed to satisfy these conventions unless stated otherwise.

This paper focuses on identification and inference in graphical models of missing data, without committing to a specific choice of the functional of the target law. We analyze identifiability by characterizing which components of the missingness mechanism are identifiable from the

observed data law $p(X^*, R)$ under $\mathcal{M}$. In some settings this yields identification of the target law $p(X)$, while in others it suffices to identify particular functionals, denoted by $\theta(p(X))$. We also outline how such identification can be used to conduct inference on θ.

The central role of the missingness mechanism follows from the identity $p(X) = p(X, R = 1) \,/\, p(R = 1 \,|\, X)$, which shows that recovery of the target law, when possible, depends on identification of $p(R = 1 \,|\, X)$. This representation makes explicit the connection between identification in missing data models and inverse probability weighting, in which the target law is expressed as a reweighted version of the complete-case distribution.

Under the graphical restriction on the missingness mechanism, the conditional distribution $p(R = 1 \,|\, X)$ factorizes into a product of conditional probabilities

$$p(X) = p(X, R = 1) \Big/ \Big\{ \prod_{R_k \in R} p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k)) \Big|_{R=1} \Big\}, \tag{1}$$

where $p(.)|_{R_j=1}$ denotes the evaluation of $p(.)$ at $R_j = 1$. We define the propensity score of R_k as $\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k)) = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))$. Our identification strategy centers on when these scores are identifiable and how they enable recovery of the target law or functionals thereof.

We assume $\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k)) > \sigma > 0$ a.s. for a fixed positive constant σ, and for all $R_k \in R$. This condition ensures nonparametric identification of the target law and its smooth functionals, and finite asymptotic variance of the proposed weighting estimators (Robins et al. 2000).

3 Building Blocks for Propensity Score Identification

Identification of the propensity score π_k requires expressing this conditional probability as a unique functional of the observed data law $p(X^*, R)$. The main difficulty arises when $\mathrm{pa}_{\mathcal{G}}(R_k)$ contains variables in X that are subject to missingness.

When a parent $X_j \in \mathrm{pa}_{\mathcal{G}}(R_k)$ is subject to missingness, any representation of π_k in terms of observed data must replace X_j by its observed counterpart on rows where $R_j = 1$, thereby introducing selection through the corresponding missingness indicators. To keep track of this bookkeeping requirement, for each $R_k \in R$ we define the *counterfactual-induced selection set*

$$\mathcal{S}_k^x = \{R_j \in R : X_j \in \mathrm{pa}_{\mathcal{G}}(R_k)\}. \qquad \text{(counterfactual-induced selection set for } R_k\text{)} \quad (2)$$

The set $\mathcal{S}_k^x$ records indicators that become relevant solely because counterfactual variables appear among the parents of R_k. Its role is descriptive, not prescriptive: the presence of $R_j \in \mathcal{S}_k^x$ indicates that substituting X_j with its observed counterpart is unavoidable when forming candidate representations of π_k, but does not determine whether such substitutions are valid for identification. Whether this substitution can be justified depends on the assumptions encoded in the missingness mechanism.

For identification of the target law, it is not necessary that π_k be identifiable as a full conditional distribution. It suffices that π_k be identifiable at the evaluation under which all missingness indicators equal one, as required by (1). We therefore define the *indicator-induced selection set* S_k^r as the set of indicators whose evaluation at one is required in order for π_k to be identified at this evaluation. By construction, $\mathcal{S}_k^r \subseteq R \cap \mathrm{pa}_{\mathcal{G}}(R_k)$.

The sets $\mathcal{S}_k^x$ and $\mathcal{S}_k^r$ capture conceptually distinct, though potentially overlapping, sources of selection in identifying π_k. The former records indicators implicated by counterfactual substitution due to missing parents of R_k, while the latter records indicators whose evaluation at one is required for π_k to be usable in recovering the target law. These sets play a common role in identification; both determine which missingness indicators may induce selection in candidate representations of π_k. It is therefore convenient to collect them into a single object that summarizes all selection relevant for identifying the propensity score of R_k. For each

$R_k \in R$, we thus define the overall *selection set*

$$\mathcal{S}_k = \mathcal{S}_k^x \cup \mathcal{S}_k^r. \qquad \text{(selection set for } R_k\text{)} \quad (3)$$

A central obstacle in identifying π_k arises when some indicators in $\mathcal{S}_k^x$ are descendants of R_k. Presence of such indicators obstructs identification through purely associational arguments. For each $R_k \in R$, we collect these into the *problematic set*

$$\mathcal{R}_k^p = \mathcal{S}_k^x \cap \mathrm{de}_{\mathcal{G}}(R_k). \qquad \text{(problematic set for } R_k\text{)} \quad (4)$$

3.1 Associational and causal irrelevancy

Our identification arguments rely on two distinct notions of irrelevancy. The first is *associational irrelevancy*, which follows from the local Markov property of mDAGs. In particular, for each $R_k \in R$, $R_k \perp\!\!\!\perp \mathrm{nd}_{\mathcal{G}}(R_k) \setminus \mathrm{pa}_{\mathcal{G}}(R_k) \mid \mathrm{pa}_{\mathcal{G}}(R_k)$.

Example 1 *(Associational irrelevancy) Consider an mDAG in which $X_2 \to R_1$ and $X_1 \to R_2$, with no edges between R_1 and R_2. Here, $\pi_1(\mathrm{pa}_{\mathcal{G}}(R_1)) = p(R_1 = 1 \mid X_2)$ and the counterfactual-induced selection set is $S_1^x = \{R_2\}$. Since R_2 is a non-descendant non-parent of R_1, the local Markov property implies $R_1 \perp\!\!\!\perp R_2 \mid X_2$. Consequently, π_1 is fully identified via $p(R_1 = 1 \mid X_2, R_2 = 1)$, and $S_1^r = \emptyset$. An analogous argument applies to identification of π_2.*

More generally, associational irrelevancy allows us to append $R_j = 1$ to the conditioning set of π_k whenever $R_j \in \mathcal{S}_k^x \cap \{\mathrm{nd}_{\mathcal{G}}(R_k) \setminus \mathrm{pa}_{\mathcal{G}}(R_k)\}$, yielding

$$\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k)) := p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k)) = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k), R_j = 1). \quad (5)$$

In contrast, when $R_j \in \mathcal{S}_k^x \cap \mathrm{pa}_{\mathcal{G}}(R_k)$, the propensity score π_k is identifiable only when $R_j = 1$, and thus R_j necessarily belongs to the indicator-induced selection set $\mathcal{S}_k^r$. Such parent indicators form a collider structure $X_j \to R_k \leftarrow R_j$, termed a *colluder* (Bhattacharya et al. 2019). Although π_k is not fully identifiable in this case, the target law may still be

identifiable, since only π_k evaluated at $R = 1$ is required for target law identification.

Associational irrelevancy implies that if the problematic set $\mathcal{R}_k^p$ is empty, all selection relevant for identifying π_k arises from parent indicators, so π_k evaluated at $\mathcal{S}_k^r = 1$ is identifiable. If instead $\mathcal{R}_k^p \neq \emptyset$, associational irrelevancy fails for indicators in $\mathcal{S}_k^x \cap \mathrm{de}_{\mathcal{G}}(R_k)$, since $R_k \not\perp\!\!\!\perp \mathrm{de}_{\mathcal{G}}(R_k) \mid \mathrm{pa}_{\mathcal{G}}(R_k)$. However, identification may still be achieved via *causal irrelevancy*.

The key observation underlying causal irrelevancy is that the propensity score π_k is invariant to interventions on missingness indicators other than R_k itself. This invariance property, also known as autonomy, modularity, and stability (Haavelmo 1944, Spirtes et al. 2001, Dawid & Didelez 2010, Pearl 2009), allows us to search for post-intervention distributions in which problematic descendant relationships with R_k are broken.

We define an intervention on $R_j \in R$ as an operation that fixes $R_j = 1$. At the graphical level, this removes all incoming edges into R_j and replaces X_j by its observed counterpart, since $R_j = 1$. At the probabilistic level, this corresponds to a truncated factorization of the full law in which $p(X, R)$ is first evaluated at $R_j = 1$ and then renormalized by the propensity score of R_j, yielding a post-intervention distribution denoted by $p(X, R \setminus R_j \mid \mathrm{do}(R_j = 1))$. Section 4 gives the explicit form of this operator and its use in constructing identifying functionals.

According to causal irrelevancy, for any $R^* \subseteq R \setminus \{R_k\}$,

$$\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k)) := p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k)) = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k), \mathrm{do}(R^* = 1)), \tag{6}$$

where $\mathrm{do}(R^* = 1)$ denotes an intervention that sets all indicators in R^* to one. Identification of π_k via causal irrelevancy proceeds by intervening on indicators in the problematic set $\mathcal{R}_k^p$ in order to sever descendant relationships with R_k.

Example 2 *(Causal irrelevancy) Suppose* $X_1 \to R_2$ *and* $R_2 \to R_1$*, so that* $S_2^x = \{R_1\}$ *and* R_1 *is a descendant of* R_2*. Associational irrelevancy fails, but since* π_2 *is invariant*

to interventions on R_1 we can write π_2 as $p(R_2 = 1 \mid X_1) = p(R_2 = 1 \mid X_1, \mathrm{do}(R_1 = 1))$. The post-intervention distribution induced by $\mathrm{do}(R_1 = 1)$ is $p(X_1, X_2, R_2 \mid \mathrm{do}(R_1 = 1)) \coloneqq p(X_1, X_2, R_1 = 1, R_2)/p(R_1 = 1 \mid R_2)$. Using simple probability rules, π_2 is thus identified from the identified margin $p(X_1, R_2 \mid \mathrm{do}(R_1 = 1)) = p(X_1, R_1 = 1, R_2)/p(R_1 = 1 \mid R_2)$.

3.2 Selection bias and propagation

Identification via causal irrelevancy hinges on the ability to intervene on indicators in the problematic set $\mathcal{R}_k^p$ without inducing selection that obstructs identification of π_k. While such interventions break descendant relationships with R_k, each intervention may itself induce selection through the selection set of the intervened indicator. Selection on non-descendants of R_k is typically benign: if such indicators lie in $\mathrm{pa}_{\mathcal{G}}(R_k)$, they can be incorporated into the indicator-induced selection set $\mathcal{S}_k^r$, while selection on non-parent non-descendants is independent of R_k by the local Markov property. Difficulties arise when an intervention induces selection on R_k or its descendants, as such selection may render π_k unidentifiable, and in some settings this induced selection can be accommodated by subsequent interventions, while in others it is unavoidable and obstructs identification altogether.

When multiple interventions are performed sequentially, selection induced by earlier interventions may propagate to later ones. Specifically, if R_j is intervened on prior to an intervention on R_k, the portion of the selection induced by $\mathrm{do}(R_j = 1)$ that propagates through R_k is

$$\mathcal{S}_{j \downarrow k} \coloneqq \mathcal{S}_j \cap \mathrm{pa}_{\mathcal{G}}(R_k), \qquad \text{(selection propagation rule)} \quad (7)$$

which we refer to as the *selection propagation rule*. Propagation of selection through parent relationships can obstruct identification, while absence of such propagation permits sequential identification, as we illustrate via examples below.

Example 3 *(Admissible and inadmissible descendant interventions) In Figure 1(a), iden-*

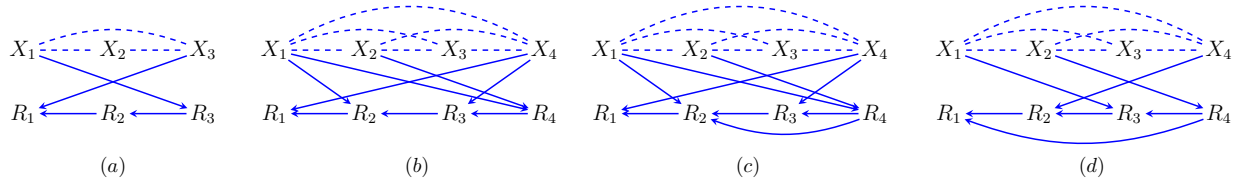


Figure 1: mDAGs illustrating selection behavior under interventions: (a) (in)admissible interventions; (b) non-propagating selection; (c) propagating selection; (d) identification may require interventions on descendants outside the causal path between R_k and $R_j \in \mathcal{R}_k^p$.

tification of $\pi_3(X_1) \coloneqq p(R_3 = 1 \,|\, X_1)$ *requires addressing the problematic set* $\mathcal{R}_3^p = \{R_1\}$. *Although* π_3 *is invariant to interventions on either* R_1 *or* R_2, *intervening on* R_1 *induces selection on* R_3, *since* $\mathcal{S}_1 = \{R_3\}$, *yielding a post-intervention distribution available only at* $R_3 = 1$ *and hence insufficient for identification. In contrast, intervening on* R_2 *induces no selection on* R_3, *as* $\mathcal{S}_2 = \emptyset$, *and the resulting post-intervention distribution admits a margin that identifies* π_3. *This example illustrates that although multiple descendant interventions may leave* π_k *invariant, only a subset yield admissible post-intervention distributions for identification. If* $X_3 \to R_1$ *is replaced by* $X_2 \to R_1$, *then intervening on* R_1 *induces selection on* R_2, *which cannot be resolved by conditional independence alone. Identification of* π_3 *therefore requires intervening on* R_2, *either alone or jointly with* R_1.

The above example shows that selection from an inadmissible intervention can sometimes be repaired by alternative or additional interventions, though not always.

Example 4 *(Unavoidable selection) Consider an mDAG with* $X_1 \to R_2$, $X_2 \to R_1$, *and* $R_2 \to R_1$. *Identifying* $\pi_2(X_1) \coloneqq p(R_2 = 1 \,|\, X_1)$ *involves the problematic set* $\mathcal{R}_2^p = \{R_1\}$. *Although* π_2 *is invariant to intervening on* R_1, *such an intervention induces selection on* R_2 *(*$\mathcal{S}_1 = \{R_2\}$*), leaving no post-intervention distribution that identifies* π_2. *This unavoidable selection bias blocks identification, as shown in (Nabi & Bhattacharya 2022, Guo et al. 2023).*

Selection induced by one intervention may also constrain subsequent intervention choices through propagation. Whether such propagation occurs depends on the graphical structure.

Example 5 *(Sequential interventions and propagation) In Figure 1(b), π_2 is identified by intervening on R_1, which induces selection on R_4. This selection does not obstruct identification of π_2, since R_4 is a non-descendant of R_2. Crucially, this induced selection does not propagate through the subsequent intervention on R_2, as $\mathcal{S}_{1\downarrow 2} = \emptyset$. As a result, intervening on R_2 yields an admissible post-intervention distribution from which π_4 is identified. In contrast, in Figure 1(c), an intervention on R_1 again induces selection on R_4, but now $R_4 \in \mathrm{pa}_{\mathcal{G}}(R_2)$. Consequently, selection propagates according to $\mathcal{S}_{1\downarrow 2} = \{R_4\}$. The subsequent intervention on R_2 then induces selection on the target indicator itself, and identification of π_4 fails.*

Identification may require intervening on descendants of R_k that lie outside causal paths from R_k to $\mathcal{R}_k^p$. Moreover, interventions used to identify one propensity score may constrain others due to induced and propagating selection.

Example 6 *(Interventions beyond causal pathways) In Figure 1(d), π_3 admits multiple admissible strategies. Intervening on R_1 alone separates R_3 from its problematic set and identifies π_3 without inducing selection on R_4. Intervening on R_2 alone or together with R_2 also identifies π_3, but only at $\mathcal{S}_3^r = \{R_4\}$, which remains sufficient for target law identification but introduces additional selection. For π_4, $\mathcal{R}_4^p = \{R_2\}$ and π_4 is invariant to interventions on R_3. However, intervening on R_3 alone induces selection on R_1, which remains a descendant of R_4 in $p(\cdot \mid \mathrm{do}(R_3 = 1))$. Thus identifying π_4 requires intervention on both R_1 and R_3. Consequently, π_3 must be identified via intervention only on R_1; intervening on R_2 instead induces selection on R_4 that propagates through R_3 and obstructs identification of π_4.*

Selection bias is central to identifiability under causal irrelevancy. Identification requires a post-intervention distribution with no selection on R_k or its non-intervened descendants and with $R_k \perp\!\!\!\perp R_j$ for all $R_j \in \mathcal{R}_k^p$. This calls for a constructive strategy that jointly accounts for induced and propagating selection and intervention order.

4 Identification Algorithm

In this section, we formalize the identification logic developed in Section 3 into a constructive procedure, summarized in Algorithm 1. The input to the algorithm is an mDAG $\mathcal{G}$ encoding restrictions on the missingness mechanism $p(R \mid X)$. The algorithm outputs a forest $\mathbb{F}$ collecting trees $\mathbb{T}_k$ associated with each missingness indicator $R_k \in R$, together with a set $\mathcal{D}$ of indicators whose propensity scores cannot be identified.

For each $R_k \in R$, the algorithm attempts to identify the propensity score $\pi_k(\text{pa}_{\mathcal{G}}(R_k))$ evaluated at an indicator-induced selection set $\mathcal{S}^r_k = 1$. If identification succeeds, the tree $\mathbb{T}_k$ encodes the sequence of interventions used to obtain an identified representation of π_k evaluated at $\mathcal{S}^r_k = 1$. If identification fails, R_k is added to $\mathcal{D}$, and the corresponding tree records where and why identification breaks down. The target law is identified if and only if $\mathcal{D} = \emptyset$, since (1) expresses $p(X)$ in terms of the collection $\{\pi_k|_{\mathcal{S}^r_k=1}\}_{k=1}^K$.

The algorithm processes missingness indicators sequentially according to a **valid reversed topological order** τ on the mDAG $\mathcal{G}$. For each indicator R_k, the algorithm attempts to identify π_k by exploiting either associational or causal irrelevancy. Trees constructed in identifying propensity scores of indicators earlier in the order are reused to guide intervention choices for identifying propensity scores of indicators later in the order.

For a fixed R_k, identification is immediate when the problematic set $\mathcal{R}^p_k$, defined in (4), is empty. In this case, associational irrelevancy arguments from Section 3.1 suffice, and the tree $\mathbb{T}_k$ consists only of the root node R_k. We therefore focus on the nontrivial case $\mathcal{R}^p_k \neq \emptyset$.

When $\mathcal{R}^p_k \neq \emptyset$, identification relies on causal irrelevancy. Conceptually, Algorithm 1 searches for a maximal admissible intervention strategy on descendants of R_k, that is, a collection of interventions that separates R_k from its problematic set $\mathcal{R}^p_k$ while avoiding selection that

propagates to R_k or obstructs subsequent identification steps. This is achieved by first considering a maximal set of admissible interventions and then pruning those that induce selection obstructing identification.

The remainder of this section describes the algorithm in detail. We first introduce the tree construction procedure used to encode candidate intervention strategies, then formalize the identification status check, and finally describe the pruning operation that removes unnecessary interventions. We conclude by characterizing the identified functionals for propensity scores whenever identification succeeds.

Trees and candidate interventions [tree-construction]. We now describe the tree construction for a fixed indicator R_k. The first step identifies which descendants of R_k can be intervened on without immediately obstructing identification of π_k.

Recall that indicators whose propensity scores are not identifiable, collected in $\mathcal{D}$, cannot be intervened on in any admissible strategy. In addition, some descendants of R_k cannot be intervened on because doing so would induce selection on R_k itself. Specifically, these are descendants of R_k that have X_k as a parent, referred to as *colluder descendants* of R_k

$$\mathcal{C}^{\mathrm{dir}}_{k,k} \coloneqq \{R_j \in \mathrm{de}_{\mathcal{G}}(R_k) \,:\, X_k \in \mathrm{pa}_{\mathcal{G}}(R_j)\}. \qquad \text{(colluder descendants of } R_k\text{)} \tag{7}$$

Thus, the initial set of indicators eligible for intervention when identifying π_k is

$$R^* \coloneqq \mathrm{de}_{\mathcal{G}}(R_k) \setminus \{\mathcal{C}^{\mathrm{dir}}_{k,k}, \mathcal{D}\}. \qquad \text{(candidate intervention set for } R_k\text{)} \tag{8}$$

Given a candidate intervention set R^* for R_k defined in (8), the algorithm constructs a provisional intervention tree $\mathbb{T}_k$ by attaching each $R_i \in R^*$ as a child of R_k, and augmenting this child with its previously constructed tree $\mathbb{T}_i$ from $\mathbb{F}$. The interventions encoded by $\mathbb{T}_k$ induce selection through two mechanisms. First, counterfactual substitution arising from

missing parents of R_k, i.e., $\mathcal{S}_k^x$ defined in (2). Second, intervening on children of R_k induces additional selection captured by the selection sets associated with each $R_j \in \mathrm{ch}_{\mathbb{T}_k}(R_k)$, i.e., $\mathcal{S}_j$ defined in (3). Let $\mathcal{T}_k \coloneqq \mathrm{ch}_{\mathbb{T}_k}(R_k)$. We collect all such selection into the *pre-selection set*

$$\widetilde{\mathcal{S}}_k \coloneqq \mathcal{S}_k^x \cup \bigcup_{R_j \in \mathcal{T}_k} \mathcal{S}_j. \qquad \text{(pre-selection set for } R_k\text{)} \quad (9)$$

The pre-selection set $\widetilde{\mathcal{S}}_k$ aggregates all selection imposed while attempting to identify π_k and is used exclusively to assess identifiability at the current stage. In contrast, the selection set $\mathcal{S}_k$ for R_k, defined in (3), characterizes only the portion of this selection that propagates through R_k when R_k itself is intervened on to identify subsequent propensity scores, according to the propagation rule in (7). By construction, $\mathcal{S}_k$ is always a subset of $\widetilde{\mathcal{S}}_k$, reflecting the fact that not all selection induced during identification propagates forward in the algorithm.

Identification check and failure modes [id-status]. Given $\mathbb{T}_k$ and $\widetilde{\mathcal{S}}_k$, the propensity score π_k is identifiable under evaluation $\mathcal{S}_k^r = 1$ if

$$R_k \perp\!\!\!\perp \widetilde{\mathcal{S}}_k \setminus \mathrm{pa}_{\mathcal{G}}(R_k) \mid \mathrm{pa}_{\mathcal{G}}(R_k) \;\text{ in }\; p(\cdot \mid \mathrm{do}(\mathcal{T}_k = 1)), \qquad \text{(identification criterion)} \quad (10)$$

where $p(\cdot \mid \mathrm{do}(\mathcal{T}_k = 1))$ corresponds to the post-intervention distribution induced by intervening on the current children of R_k in $\mathbb{T}_k$. The set $\widetilde{\mathcal{S}}_k \setminus \mathrm{pa}_{\mathcal{G}}(R_k)$ may include R_k itself, and by convention R_k is not independent of itself.

If (10) fails, collect indicators in $\widetilde{\mathcal{S}}_k \setminus \mathrm{pa}_{\mathcal{G}}(R_k)$ for which (10) fails to hold into the set

$$\mathcal{R}_k^d \coloneqq \Big\{ R_j \in \widetilde{\mathcal{S}}_k \setminus \mathrm{pa}_{\mathcal{G}}(R_k) \;:\; R_k \not\perp\!\!\!\perp R_j \mid \mathrm{pa}_{\mathcal{G}}(R_k) \text{ in } p(\cdot \mid \mathrm{do}(\mathcal{T}_k = 1)) \Big\}. \quad (11)$$

If $\mathcal{R}_k^d \cap \mathcal{R}_k^p \neq \emptyset$, then no admissible intervention can separate R_k from its problematic set R_k^p, and π_k cannot be identifiable (even at evaluation $R = 1$). In such cases, R_k is added to $\mathcal{D}$, and the algorithm returns non-identifiability of both π_k and the target law. Intuitively, identification fails because all possible interventions have already been applied in an attempt

to establish independence between R_k and the problematic indicators in $\mathcal{R}_k^p$ given $\text{pa}_{\mathcal{G}}(R_k)$.

In contrast if $\mathcal{R}_k^d \cap \mathcal{R}_k^p = \emptyset$, the dependence between R_k and $\mathcal{R}_k^p$ has already been resolved, but identification fails due to selection induced by unnecessary interventions. In this case, identification may still be achievable by pruning parts of the intervention trees, as described next. An example of pruning an unnecessary intervention was described in Example 6, where identifying π_4 required identifying π_3 without intervening on R_2.

Pruning unnecessary interventions [tree-prune]. If $\mathcal{R}_k^d \cap \mathcal{R}_k^p = \emptyset$, identification may still be possible by pruning interventions that introduce selection which propagates to R_k.

Let $\mathcal{C}_k$ denote the collection of descendants of R_k such that intervening on them induces selection on at least one indicator in $\mathcal{R}_k^d$. Specifically, for each $R_j \in \mathcal{R}_k^d$, define $\mathcal{C}_{k,j}^{\text{dir}} \coloneqq \{R_i \in \text{de}_{\mathcal{G}}(R_k) \mid X_j \in \text{pa}_{\mathcal{G}}(R_i)\}$, the set of descendants of R_k whose intervention induces selection on R_j. We then define

$$\mathcal{C}_k \coloneqq \bigcup_{R_j \in \mathcal{R}_k^d} \mathcal{C}_{k,j}^{\text{dir}}. \tag{12}$$

Any indicator in $\mathcal{C}_k$ cannot be intervened on without obstructing identification of π_k.

The pruning procedure proceeds as follows. For each candidate intervention $R_i \in R^*$: If $R_i \in \mathcal{C}_k$, then intervening on R_i necessarily induces selection on an indicator in $\mathcal{R}_k^d$, and R_i is removed from the candidate set R^*. Otherwise if $R_i \notin \mathcal{C}_k$, consider the subtree $\mathbb{T}_i$ retrieved from $\mathbb{F}$. If $\mathbb{T}_i$ contains children whose interventions induce selection on some $R_j \in \mathcal{R}_k^d$ that propagates through R_i according to the propagation rule in (7), then the corresponding branches are pruned from $\mathbb{T}_i$. After this pruning, the algorithm checks whether the propensity score π_i remains identifiable under its updated evaluation set. If so, the pruned tree $\mathbb{T}_i$ is appended to $\mathbb{T}_k$; otherwise, R_i is removed from R^*. Note that if any child of R_i in $\mathbb{T}_i$, say R_m, is also a child of R_k in $\mathbb{T}_k$, and the subtree $\mathbb{T}_m$ has already been pruned, then the corresponding subtree of R_m in $\mathbb{T}_i$ must be updated to match the pruned version. This

operation enables us to restrict attention to pruning the children of R_k, or their children, without needing to consider further descendants. See Appendix Figure E.2(c) for an example, with detailed identification discussion given in Appendix Subsection B.1.4. All pruned branches are recorded locally in a set $\mathcal{B}$, and the final $\mathbb{T}_k$ is added to $\mathbb{F}$.

Once all candidates in R^* have been examined, the identification criterion (10) is evaluated again using the updated tree $\mathbb{T}_k$. If it is satisfied, identification of π_k succeeds. Otherwise, $\mathcal{R}_k^d$ is updated and the pruning procedure is repeated. Since the set of indicators is finite, this iterative process terminates.

When identification of π_k succeeds, the algorithm constructs the indicator-induced selection set for R_k as follows and proceed to identify the next propensity score according to τ:

$$\mathcal{S}_k^r := \widetilde{\mathcal{S}}_k \cap \mathrm{pa}_{\mathcal{G}}(R_k) = \{\mathcal{S}_k^x \cap \mathrm{pa}_{\mathcal{G}}(R_k)\} \cup \{\cup_{R_j \in \mathrm{ch}_{\mathbb{T}_k}(R_k)} \mathcal{S}_{j\downarrow k}\}.$$

(indicator-induced selection set for R_k) (13)

This set has two components. First, any indicator in $\mathcal{S}_k^x \cap \mathrm{pa}_{\mathcal{G}}(R_k)$ must be set to one to render the corresponding counterfactual parents observed. Second, interventions used to identify π_k may induce selection that propagates through R_k via (7). Under associational irrelevancy, discussed in Section 3.1, this reduces to $\mathcal{S}_k^x \cap \mathrm{pa}_{\mathcal{G}}(R_k)$, since $\mathcal{T}_k = \emptyset$.

The selection set for R_k in (3) can therefore be explicitly defined as

$$\mathcal{S}_k := \mathcal{S}_k^x \cup \mathcal{S}_k^r \ = \mathcal{S}_k^x \cup \bigcup_{R_j \in \mathcal{T}_k} \mathcal{S}_{j\downarrow k}. \tag{14}$$

Identification functional induced by the intervention tree $\mathbb{T}_k$. The preceding construction shows that whenever Algorithm 1 succeeds, the intervention tree $\mathbb{T}_k$ encodes a concrete strategy for intervening on missingness indicators so as to obtain a post-intervention distribution in which R_k is independent of its problematic set, without inducing selection on

R_k or on its non-intervened descendants.

To construct an identification functional for π_k given $\mathbb{T}_k$, we define graphical and probabilistic fixing operators. For $R_i \in R$, let $\phi^{\mathcal{G}}_{R_i}$ denote the graphical fixing operation on an mDAG $\mathcal{G}$ that removes all incoming edges into R_i and replaces the variable X_i by its observed counterpart X_i^*, corresponding to the intervention $\mathrm{do}(R_i = 1)$. Further, we define the probabilistic fixing operator $\phi^p_{R_i}\{p\}$ that maps $p(X, R)$ to a law on $(X, R \setminus \{R_i\})$ defined by

$$\phi^p_{R_i}\{p\}(X, R \setminus R_i) \;\coloneqq\; \frac{p(X, R \setminus R_i, R_i = 1)}{p(R_i = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_i))}, \tag{15}$$

For a sequence of indicators $\sigma = (s_1, \ldots, s_m)$, we define the composed fixing operator $\phi^p_\sigma \;\coloneqq\; \phi^p_{s_m} \circ \cdots \circ \phi^p_{s_1}$, and interpret ϕ^p_σ as the post-intervention distribution induced by intervening sequentially on the indicators in σ.

Given R_k and $\mathbb{T}_k$, let $\sigma_k = (s_1, \ldots, s_m)$ be any ordering of $\mathcal{T}_k$ consistent with the reverse topological order used by the algorithm. The post-intervention distribution induced by $\mathbb{T}_k$ is $\phi^p_{\sigma_k}\{p\}$. The following theorem shows that whenever the identification criterion holds in $\phi^p_{\sigma_k}\{p\}$ (i.e., condition (10)), the tree $\mathbb{T}_k$ yields an explicit identifying functional for π_k.

Theorem 1 (Identification functional induced by $\mathbb{T}_k$) *Assume identification criterion (10) holds for R_k using the tree $\mathbb{T}_k$ in the post-intervention law $p_{\mathbb{T}_k} = \phi^p_{\sigma_k}\{p\}$. Then $\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))$, evaluated at $\mathcal{S}^r_k = 1$, is identified from the observed data law via*

$$\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1} \;=\; p_{\mathbb{T}_k}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1}. \tag{16}$$

Theorem 1 makes explicit the role of the intervention tree in identification. The tree $\mathbb{T}_k$ determines a post-intervention distribution in which the conditional distribution of R_k given its parents is well-defined and identifiable, while the evaluation set $\mathcal{S}^r_k$ records the minimal indicator restrictions under which this functional can be recovered from observed data.

Together, $(\mathbb{T}_k, \mathcal{S}_k^r)$ characterize how π_k is identified. The identifying functional in Theorem 1 admits a representation in terms of observed data and inverse-probability weights.

Corollary 2 (Observed-data representation) *Under the conditions of Theorem 1, define* $W_k := \prod_{R_i \in \mathcal{T}_k} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i))^{-1}$, *where each* π_i *is evaluated at its own identified evaluation set as returned by Algorithm 1. Then*

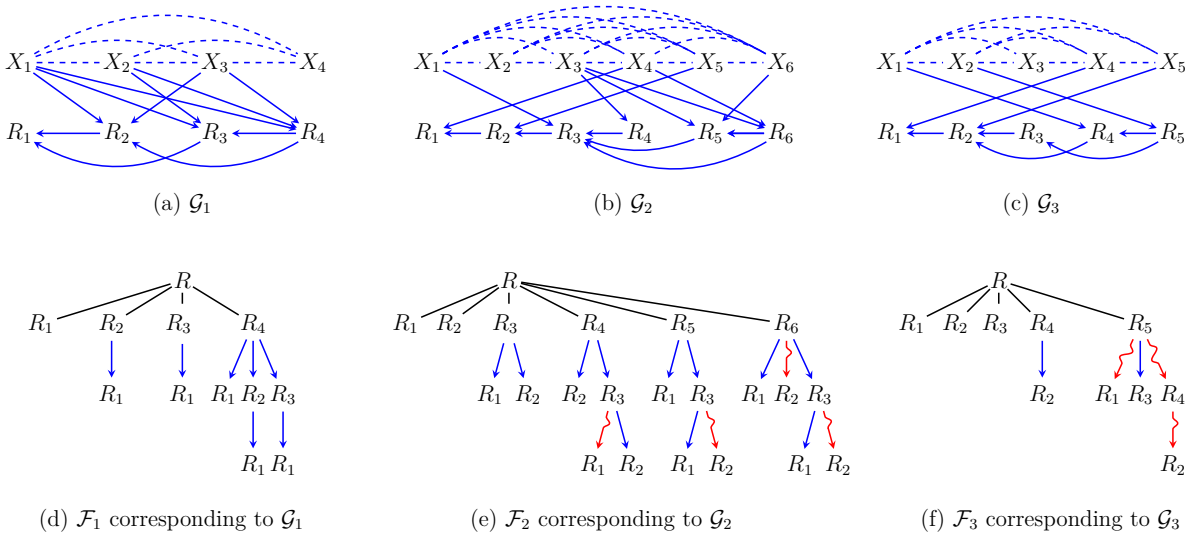
$$\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}_k^r=1} = \frac{\mathbb{E}\Big[\mathbb{I}(R_k=1)\, W_k \;\Big|\; \mathrm{pa}_{\mathcal{G}}(R_k),\ \mathcal{T}_k = 1,\ \widetilde{\mathcal{S}}_k = 1\Big]}{\mathbb{E}\Big[W_k \;\Big|\; \mathrm{pa}_{\mathcal{G}}(R_k),\ \mathcal{T}_k = 1,\ \widetilde{\mathcal{S}}_k = 1\Big]}\Big|_{\mathcal{S}_k^r=1}. \tag{17}$$

If $\mathcal{T}_k = \emptyset$, *then* $W_k \equiv 1$ *and the expression reduces to the associational identification formula.*

We illustrate the tree-construction of Algorithm 1 through the following examples.

Example 7 *(Identification without pruning) The mDAG* $\mathcal{G}_1$ *in Figure 2(a) illustrates a setting in which no pruning is required. The algorithm begins with* R_1, *whose propensity score* $\pi_1 = p(R_1 = 1 \,|\, R_2, R_3)$ *is directly observed. The algorithm then proceeds to* R_2 *and* R_3 *in either order, since neither is a descendant of the other. For concreteness, consider* R_2, *whose propensity score is* $\pi_2 = p(R_2 = 1 \,|\, X_1, X_3, R_4)$. *Here,* $\mathcal{R}_2^p = \{R_1\}$, *so an intervention on* R_1 *is required to separate* R_2 *from its problematic set. Since* $\mathcal{S}_1 = \emptyset$, *this intervention induces no additional selection, and* π_2 *is fully identified. An analogous argument applies to* R_3. *Finally, for* R_4, *we have* $\pi_4 = p(R_4 = 1 \,|\, X_1, X_2, X_3)$ *with* $\mathcal{R}_4^p = \{R_1, R_2, R_3\}$. *Intervening on all three indicators yields separation without inducing selection, so* π_4 *is identified. As all propensity scores are identified without evaluation restrictions, the target law is identified. Key quantities for this example are summarized in Appendix Table 4, with* $\mathbb{F}_1$ *shown in Figure 2(d).*

Example 8 *(Identification with pruning) The mDAG* $\mathcal{G}_2$ *in Figure 2(b) demonstrates how pruning removes unnecessary interventions that would otherwise obstruct identification. The propensity scores for* R_1 *and* R_2 *are identified via associational irrelevancy. Consider* R_3,



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\mathcal{G}_3$

(d) $\mathcal{F}_1$ corresponding to $\mathcal{G}_1$ (e) $\mathcal{F}_2$ corresponding to $\mathcal{G}_2$ (f) $\mathcal{F}_3$ corresponding to $\mathcal{G}_3$

Figure 2: (a, b, c) Examples used to illustrate the identification Algorithm 1; (d, e, f) The corresponding constructed trees.

whose propensity score $\pi_3 = p(R_3 = 1 \,|\, X_1, R_4, R_5, R_6)$ *has* $\mathcal{R}_3^p = \{R_1\}$. *Intervening on both* R_1 *and* R_2 *yields identification of* π_3 *evaluated at* $\mathcal{S}_3^r = \{R_4, R_5\}$. *When identifying* $\pi_4 = p(R_4 = 1 \,|\, X_3)$, *however, intervening on* R_3 *using its full tree induces selection on* R_4 *itself. The algorithm detects this via* $\mathcal{R}_4^d$ *and prunes the branch corresponding to the unnecessary intervention on* R_2 *within* $\mathbb{T}_3$. *After pruning,* π_3 *remains identifiable under a reduced evaluation set, and* π_4 *becomes identifiable. A similar pruning step is required when identifying* π_6, *where selection induced indirectly through* $\mathbb{T}_3$ *must again be removed. In this graph, all problematic selections can be eliminated through pruning, and the target law is identified. Details are summarized in Appendix Table 5, with* $\mathbb{F}_2$ *shown in Figure 2(e).*

Example 9 *(Non-identification despite pruning) The mDAG* $\mathcal{G}_3$ *in Figure 2(c) illustrates a setting in which pruning cannot eliminate all selection bias. We focus on* R_5, *whose propensity score is* $\pi_5 = p(R_5 = 1 \,|\, X_2)$ *with* $\mathcal{R}_5^p = \{R_2\}$ *and* $\mathcal{C}_{5,5}^{dir} = \{R_2\}$. *Interventions on* R_1, R_3, *and* R_4 *are initially admissible. Pruning removes the subtree for* R_4, *but this renders* π_4 *itself non-identifiable. Subsequent pruning eliminates* R_1, *leaving only* R_3. *An intervention on* R_3 *alone,*

however, cannot separate R_5 from R_2, yielding $\mathcal{R}_5^d = \mathcal{R}_5^p$ and triggering non-identification. Thus, even though pruning removes unnecessary interventions, all admissible intervention strategies induce unavoidable selection on the problematic set. The algorithm concludes that π_5, and hence the target law, is not identifiable. Appendix Table 6 provides a summary of this example, with $\mathbb{F}_3$ shown in Figure 2(f).

5 Estimation Procedure

The identification algorithm in Section 4 returns, for each $R_k \in R$, an intervention tree $\mathbb{T}_k$ and an evaluation set $\mathcal{S}_k^r$ such that $\pi_k(\text{pa}_\mathcal{G}(R_k))|_{\mathcal{S}_k^r=1}$ is identified as a functional of the observed data law. Given n i.i.d. copies of (X^*, R) drawn from the observed data law induced by the full law $p(X, R) \in \mathcal{M}$, this section develops estimation and inference procedures that mirror that construction. We first describe *recursive inverse probability weighted* estimating euqations for the identified propensity scores $\{\pi_k|_{\mathcal{S}_k^r=1}\}_{k=1}^K$. We then describe estimation of generic parameters defined through moment conditions under the target law. These procedures are summarized in Appendices Algorithms E1 and E2.

5.1 Estimation of propensity scores

Estimation proceeds sequentially in the reversed topological order τ used by the identification algorithm. Let $\mathcal{P}$ and $\mathcal{E}$ collect the fitted propensity score models and the estimating equations, respectively, initialized as $\emptyset$. For each R_k, fix a parametric model for $\pi_k(\text{pa}_\mathcal{G}(R_k); \theta_k)|_{\mathcal{S}_k^r=1}$ indexed by a finite dimensional parameter θ_k.

Let

$$W_k(\theta_{\mathcal{T}_k}) \coloneqq \prod_{R_i \in \mathcal{T}_k} \frac{\mathbb{I}(R_i = 1)}{\pi_i(\text{pa}_\mathcal{G}(R_i); \theta_i)|_{\mathcal{S}_i^r=1}}, \tag{18}$$

Algorithm 1 TREE-BASED IDENTIFICATION ALGORITHM($\mathcal{G}(X, R, X^*)$)

1: **Identification Procedure:**

- Index R according to a valid reversed topological order on R, denoted by τ
- Let $\mathbb{F}$ be the list of all constructed trees; initialize it to an empty list
- Let $\mathcal{D} = \{R_j \in R \mid \pi_{j|R=1} \text{ is not ID}\}$; initialize it to an empty list
- **For** each $R_k \in R$, let:
  - Let $\mathbb{T}_k$ be a tree data structure with R_k as the root node
  - $\mathcal{S}_k^x \coloneqq \{R_j \in R \mid X_j \in \text{pa}_{\mathcal{G}}(R_k)\}$ and $\mathcal{R}_k^p \coloneqq \mathcal{S}_k^x \cap \text{de}_{\mathcal{G}}(R_k)$
  - **If** $\mathcal{R}_k^p = \emptyset$: $*$ Add $\mathbb{T}_k$ to $\mathbb{F}$ $*$ $\tilde{\mathcal{S}}_k = \mathcal{S}_k = \mathcal{S}_k^x$ $*$ $\mathcal{S}_k^r = \tilde{\mathcal{S}}_k \cap \text{pa}_{\mathcal{G}}(R_k)$
    Else: $*$ $(\mathbb{T}_k, \tilde{\mathcal{S}}_k, \mathcal{S}_k, \mathcal{S}_k^r, \mathbb{F}, \mathcal{D}) \leftarrow$ **tree-construction**$(R_k, \mathbb{F}, \mathcal{D})$
- **If** $\mathcal{D} = \emptyset$: Print *target law is identified*
- RETURN $\mathcal{D}, \mathbb{F}$

2: **tree-construction**$(R_k, \mathbb{F}, \mathcal{D})$

- Let $\mathcal{C}_{k,k}^{\text{dir}} \coloneqq \{R_i \in \text{de}_{\mathcal{G}}(R_k) \mid X_k \in \text{pa}_{\mathcal{G}}(R_i)\}$
- **For** $R_i \in \text{de}_{\mathcal{G}}(R_k) \setminus \{\mathcal{C}_{k,k}^{\text{dir}}, \mathcal{D}\}$: – Add $R_k \to R_i$ in $\mathbb{T}_k$, augment R_i in $\mathbb{T}_k$ with $\mathbb{T}_i$ in $\mathbb{F}$
- Let $\tilde{\mathcal{S}}_k = \mathcal{S}_k^x \cup_{R_i \in \text{ch}_{\mathbb{T}_k}(R_k)} \mathcal{S}_i$
- (id_status, $\tilde{\mathcal{S}}_k, \mathbb{T}_k$) $\leftarrow$ **id-status**$(R_k, \tilde{\mathcal{S}}_k, \mathbb{T}_k)$
- $\mathcal{S}_k^r = \tilde{\mathcal{S}}_k \cap \text{pa}_{\mathcal{G}}(R_k)$ and $\mathcal{S}_k = \mathcal{S}_k^x \cup \mathcal{S}_k^r$
- **If** id_status=T: $*$ Add $\mathbb{T}_k$ to $\mathbb{F}$ **Else:** $*$ Add R_k to $\mathcal{D}$
- RETURN $\mathbb{T}_k, \tilde{\mathcal{S}}_k, \mathcal{S}_k, \mathcal{S}_k^r, \mathbb{F}, \mathcal{D}$

3: **tree-prune**$(R_i, \mathbb{T}_i, \mathbb{T}_k, \mathcal{C}_k, \mathcal{B})$

- **For** all $R_m \in \text{ch}_{\mathbb{T}_i}(R_i)$:
  - **If** $R_m \in \mathcal{B}$: $*$ Update $\mathbb{T}_m$ with the one in $\mathcal{B}$
  - **If** $R_m \in \mathcal{C}_k$: Prune the node R_m and its corresponding tree from the subtree $\mathbb{T}_i$
- Update: $*$ $\tilde{\mathcal{S}}_i = \mathcal{S}_i^x \cup_{R_m \in \text{ch}_{\mathbb{T}_i}(R_i)} \mathcal{S}_m$ $*$ $\mathcal{S}_i^r = \tilde{\mathcal{S}}_i \cap \text{pa}_{\mathcal{G}}(R_i)$ $*$ $\mathcal{S}_i = \mathcal{S}_i^x \cup \mathcal{S}_i^r$
- (id_status, $\tilde{\mathcal{S}}_i, \mathbb{T}_i$) $\leftarrow$ **id-status**$(R_i, \tilde{\mathcal{S}}_i, \mathbb{T}_i)$
- **If** id_status=T: – Update $\mathbb{T}_i$ in $\mathcal{B}$. Replace $\mathbb{T}_i$ augmented to R_i in $\mathbb{T}_k$ by $\mathbb{T}_i$ from $\mathcal{B}$
  Else: Prune R_i and augmented $\mathbb{T}_i$ from $\mathbb{T}_k$
- RETURN $(\mathbb{T}_i, \tilde{\mathcal{S}}_i, \mathbb{T}_k, \mathcal{B})$

4: **id-status**$(R_k, \tilde{\mathcal{S}}_k, \mathbb{T}_k)$

- Let $\mathcal{R}_k^d \coloneqq \left\{R_j \in \tilde{\mathcal{S}}_k \setminus \text{pa}_{\mathcal{G}}(R_k) \middle| R_k \not\perp\!\!\!\perp R_j \mid \text{pa}_{\mathcal{G}}(R_k) \text{ in } p(. \mid \text{do}\{\text{ch}_{\mathbb{T}_k}(R_k)\})\right\}$
- **If** $\mathcal{R}_k^d = \emptyset$: – id_status = T – RETURN id_status, $\tilde{\mathcal{S}}_k, \mathbb{T}_k$
- **Else If**: $\mathcal{R}_k^p \cap \mathcal{R}_k^d \neq \emptyset$: – id_status = F – RETURN id_status, $\tilde{\mathcal{S}}_k, \mathbb{T}_k$
- **Else**:
  - Let $\mathcal{B}$ collect updated branches for $\mathbb{T}_k$ and the corresponding indicators; initialized to an empty list.
  - Let $\mathcal{C}_k \coloneqq \cup_{R_j \in \mathcal{R}_k^d} \mathcal{C}_{k,j}^{\text{dir}}$ where $\mathcal{C}_{k,j}^{\text{dir}} \coloneqq \{R_i \in \text{de}_{\mathcal{G}}(R_k) \mid X_j \in \text{pa}_{\mathcal{G}}(R_i)\}$
  - **For** $R_i \in \text{ch}_{\mathbb{T}_k}(R_k)$: (consistent with the order τ)
    - $*$ **If** $R_i \in \mathcal{C}_k$: Prune the node R_i and its corresponding tree from $\mathbb{T}_k$
    - $*$ **If** $R_i \notin \mathcal{C}_k$, and either (i) $\text{ch}_{\mathbb{T}_i}(R_i) \cap \mathcal{B} \neq \emptyset$, or (ii) $\exists R_j \in \mathcal{R}_k^d$ s.t. $\text{ch}_{\mathbb{T}_i}(R_i) \cap \mathcal{C}_{k,j}^{\text{dir}} \neq \emptyset$ and $R_j \in \text{pa}_{\mathcal{G}}(R_k)$:
      - $(\mathbb{T}_i, \mathcal{S}_i, \mathbb{T}_k, \mathcal{B}) \leftarrow$ **tree-prune**$(R_i, \mathbb{T}_i, \mathbb{T}_k, \mathcal{C}_k, \mathcal{B})$
  - Update: $*$ $\tilde{\mathcal{S}}_k = \mathcal{S}_k^x \cup_{R_i \in \text{ch}_{\mathbb{T}_k}(R_k)} \mathcal{S}_i$
  - RETURN **id-status**$(R_k, \tilde{\mathcal{S}}_k, \mathbb{T}_k)$

be the inverse propensity weight with the convention $W_k(\theta_{\mathcal{T}_k}) \equiv 1$ if $\mathcal{T}_k = \emptyset$, where $\mathcal{T}_k$ denotes the children of R_k in $\mathbb{T}_k$. The product in (18) implements the same intervention logic used for identification, namely, intervening on the indicators in $\mathcal{T}_k$, by setting them to one and reweighting by the product of their propensity scores.

Recall the pre-selection set $\widetilde{\mathcal{S}}_k$ in (9), and let $f_k(\mathrm{pa}_{\mathcal{G}}(R_k))$ be a vector whose dimension matches that of θ_k. We estimate θ_k by solving the empirical estimating equation

$$P_n \Psi_k(X^*, R; \theta_k, \hat{\theta}_{\mathcal{T}_k}) \;=\; 0, \tag{19}$$

where

$$\Psi_k(X^*, R; \theta_k, \hat{\theta}_{\mathcal{T}_k}) \;:=\; \mathbb{I}(\widetilde{\mathcal{S}}_k = 1)\, W_k(\hat{\theta}_{\mathcal{T}_k})\, f_k(\mathrm{pa}_{\mathcal{G}}(R_k)) \left\{ R_k - \pi_k(\mathrm{pa}_{\mathcal{G}}(R_k); \theta_k) \right\}. \tag{20}$$

The restriction $\mathbb{I}(\widetilde{\mathcal{S}}_k = 1)$ ensures that the covariates in $\mathrm{pa}_{\mathcal{G}}(R_k)$ and $\mathrm{pa}_{\mathcal{G}}(R_i)$ for all $R_i \in \mathcal{T}_k$ are observed and that each propensity score appearing in the weight $W_k(\theta_{\mathcal{T}_k})$ is evaluated under its required evaluation $\mathcal{S}_i^r = 1$. When $\mathcal{T}_k = \emptyset$, (20) reduces to a standard estimating equation computed on the subset $\widetilde{\mathcal{S}}_k = 1$. When $\mathcal{T}_k \neq \emptyset$, the weighting term implements the post-intervention distribution used to justify identification of π_k.

The pruning step in Algorithm 1 may modify a previously stored subtree $\mathbb{T}_i$ when appending it to $\mathbb{T}_k$. This changes the child set $\mathcal{T}_i = \mathrm{ch}_{\mathbb{T}_i}(R_i)$ and therefore changes the estimating equation (20) used to estimate θ_i. In this case, θ_i must be re-estimated using the pruned version of $\mathbb{T}_i$. We use the same notation Ψ_i for the modified estimating functions, but to distinguish them, we collect them in $\mathcal{E}_k$, with $\mathcal{P}_k$ collecting the propensity score fits after all re-estimation steps needed for identifying π_k; the subscript k indicates they are specifically tailored for estimating π_k. Re-estimation is also performed in the order τ, ensuring that any parameters required to form weights in descendant equations are available when needed.

The estimator $\hat{\theta}_k$ is consistent for θ_k if the models for π_k and all propensity scores for $R_i \in \mathcal{T}_k$ are correctly specified and standard regularity conditions hold (Liang & Zeger 1986, Robins

et al. 1995). Its asymptotic variance has a sandwich form, with contributions from the estimating equation for θ_k and the associated propensity score equations, detailed below.

Let $\boldsymbol{\theta}_k$ denote the stacked parameter vector consisting of θ_k and the parameters associated with all indicators appearing in the (possibly pruned) intervention tree $\mathbb{T}_k$, ordered so that θ_k is last. Let $\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)$ be the corresponding stacked estimating function obtained by collecting the estimating equations (20) for R_k, for its children in $\mathbb{T}_k$, and recursively for all descendants of those children in $\mathbb{T}_k$. By construction, $\boldsymbol{\Psi}_k$ includes exactly the estimating equations whose solutions are required to form the inverse propensity weights used in estimating π_k.

Theorem 3 (Asymptotic normality of recursive propensity score estimators) *Assume that each propensity score model appearing in $\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)$ is correctly specified, and that standard regularity conditions for M-estimation hold, including differentiability of $\boldsymbol{\Psi}_k$, finiteness of second moments, and nonsingularity of $A_k \coloneqq \mathbb{E}\{\partial \boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k\}$. Then $\hat{\boldsymbol{\theta}}_k$, the solution to $P_n \boldsymbol{\Psi}_k(\boldsymbol{\theta}_k) = 0$, satisfies*

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) \;\rightsquigarrow\; N(0,\, V_k), \qquad V_k \;=\; A_k^{-1} B_k (A_k^{-1})', \tag{21}$$

where $B_k \coloneqq \mathbb{E}\{\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)'\}$. A consistent estimator of V_k is $\hat{V}_k = \hat{A}_k^{-1}\hat{B}_k(\hat{A}_k^{-1})'$ with $\hat{A}_k = P_n\{\partial \boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k\}|_{\hat{\boldsymbol{\theta}}_k}$ and $\hat{B}_k = P_n\{\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)'\}|_{\hat{\boldsymbol{\theta}}_k}$.

The asymptotic variance of $\hat{\theta}_k$ is the bottom-right block of V_k (and similarly for $\hat{V}_k$).

In settings where parametric models for π_k are difficult to specify, one may replace $\pi_k(\cdot;\theta_k)$ by a flexible regression estimator of $p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))$ fit on the subset $\widetilde{\mathcal{S}}_k = 1$, using observation weights W_k when $\mathcal{T}_k \neq \emptyset$. Many flexible machine learning and statistical models support weights, including tree-based methods and boosting, and can flexibly capture nonlinearities and high-order interactions among covariates. Inference in this setting typically relies on the bootstrap, since closed form variance expressions are generally unavailable; formal asymptotic

normality requires additional conditions (e.g., sample splitting or complexity restrictions) and is not pursued here.

5.2 Statistical analyses for functionals of the target law

Let θ denote a target parameter defined as a functional of the target law $p(X)$. We assume θ is characterized by a moment condition

$$\mathbb{E}\{\, M(\tilde{X};\, \theta)\,\} = 0, \tag{22}$$

where $\tilde{X} \subseteq X$ is the subset of variables required to evaluate M and the dimension of M matches that of θ. Let $\tilde{R} \subseteq R$ denote the missingness indicators for $\tilde{X}$.

Because $M(\tilde{X};\theta)$ is only observed when $\tilde{R} = 1$, estimation must adjust for selection. Moreover, each propensity score appearing in the required inverse probability weights may itself only be evaluable on a further restricted subset determined by its selection set. This motivates the following closure construction. Define the operator $\mathrm{cl}(\cdot)$ on sets of indicators by

$$\mathrm{cl}(A) \;\coloneqq\; A \cup \bigcup_{R_i \in A} \mathcal{S}_i, \tag{23}$$

where $\mathcal{S}_i$ is the selection set for R_i produced by the identification algorithm. Starting from $A_0 \coloneqq \tilde{R}$, define the recursion $A_{\ell+1} \coloneqq \mathrm{cl}(A_\ell)$. Since R is finite, there exists ℓ^* such that $A_{\ell^*+1} = A_{\ell^*}$. We define the smallest set of indicators containing $\tilde{R}$ that is closed under inclusion of the selection sets $\{\mathcal{S}_i\}$

$$\mathcal{R} \;\coloneqq\; A_{\ell^*}, \tag{24}$$

Given estimates of $\pi_i(\mathrm{pa}_{\mathcal{G}}(R_i))|_{\mathcal{S}_i^r=1}$ for $R_i \in \mathcal{R}$ from Section 5.1, define

$$\Psi(X^*, R; \theta, \theta_{\mathcal{R}}) \;\coloneqq\; \mathbb{I}(\mathcal{R} = 1) \Big\{ \prod_{R_i \in \mathcal{R}} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r=1} \Big\}^{-1} M(\tilde{X}; \theta), \tag{25}$$

where $\theta_{\mathcal{R}} \coloneqq (\theta_i)_{R_i \in \mathcal{R}}$. The estimator $\hat{\theta}$ is obtained by solving

$$P_n \Psi(X^*, R;\, \theta, \hat{\theta}_{\mathcal{R}}) = 0, \tag{26}$$

and is consistent for θ if all the propensity scores for indicators in $\mathcal{R}$ are consistently estimated.

Theorem 4 (Asymptotic normality for target parameters) *Assume that all propensity score models associated with indicators in $\mathcal{R}$ are correctly specified, and that standard regularity conditions for the stacked estimating equations hold for the system consisting of the propensity score estimating equations used to fit $\{\theta_i : R_i \in \mathcal{R}\}$ together with (26). Then $\hat{\theta}$ is consistent and $\sqrt{n}\,(\hat{\theta} - \theta) \rightsquigarrow N(0,\, V)$, where V is obtained as the appropriate block of the sandwich covariance matrix for the stacked system $\boldsymbol{\Psi}(\boldsymbol{\theta})$, i.e., the estimating equations required to evaluate the inverse probability weights appearing in Ψ*

We briefly illustrate the estimation procedure for a simple functional of the target law, namely the mean of a partially observed variable.

Example 10 *Consider estimation of $\theta = \mathbb{E}(X_3)$ in the missing data model of Figure 2(c). Evaluating the estimating function $M(\tilde{X}; \theta) = X_3 - \theta$ requires observing X_3, so the initial indicator set is $\tilde{R} = \{R_3\}$. In this graph, the selection set associated with R_3 is empty, implying that the closure construction stabilizes immediately and yields $\mathcal{R} = \{R_3\}$. As a result, θ is identified even though the full target law is not, and can be estimated using a single inverse probability weighted estimating equation based on the propensity score for R_3. By contrast, attempting to estimate the mean of other variables in the same graph leads to a closure $\mathcal{R}$ that includes indicators whose propensity scores are not identified, revealing non-identifiability of the corresponding means.*

These examples highlight how the proposed framework distinguishes identifiable functionals from non-identifiable ones, even when full recovery of the target law is impossible. Detailed

derivations for these mean estimators, together with additional examples illustrating recursive closure, parametric regression, and causal effect estimation, are provided in Appendix B.2.

6 Simulations

We evaluated the finite-sample performance of the proposed estimation procedures across three statistical tasks of increasing complexity: (1) mean estimation, (2) parametric regression, and (3) causal effect estimation. We compared our approach with complete-case analysis and two widely used imputation-based methods implemented in R: Amelia, which relies on a multivariate normal model estimated via EM with bootstrapping, and multiple imputation by chained equations (MICE). The accompanying R package `flexMissing` implements the proposed methods. The simulation code is provided separately at `annaguo-bios/missing-tree-paper`.

Across all tasks, data were generated from four mDAGs of increasing complexity. These include one three-variable MAR model (Appendix Figure E.1(a)) and three MNAR models: a three-variable mDAG (Appendix Figure E.1(c)), a five-variable mDAG (Appendix Figure E.1(e)), and a ten-variable mDAG (Appendix Figure E.1(f)). We denote these by $\mathcal{G}_1$ through $\mathcal{G}_4$. For each mDAG, we considered sample sizes of 500, 1000, 2000, 4000, and 8000, with 500 Monte Carlo replicates per setting. Data-generating processes are detailed in Appendix D.2. Default settings were used for Amelia and MICE, including five imputations.

Task 1: mean estimation. For all mDAGs, we generated continuous data and targeted the mean of X_3. Estimation bias across simulation replicates is summarized using boxplots in Figure 3. Under the MAR model $\mathcal{G}_1$, all methods except complete-case analysis exhibited negligible bias as sample size increased. In contrast, complete-case analysis remained biased even at the largest sample size. Under the MNAR models $\mathcal{G}_2$ through $\mathcal{G}_4$, only the proposed tree-based method consistently recovered the true mean, while all competing methods exhib-

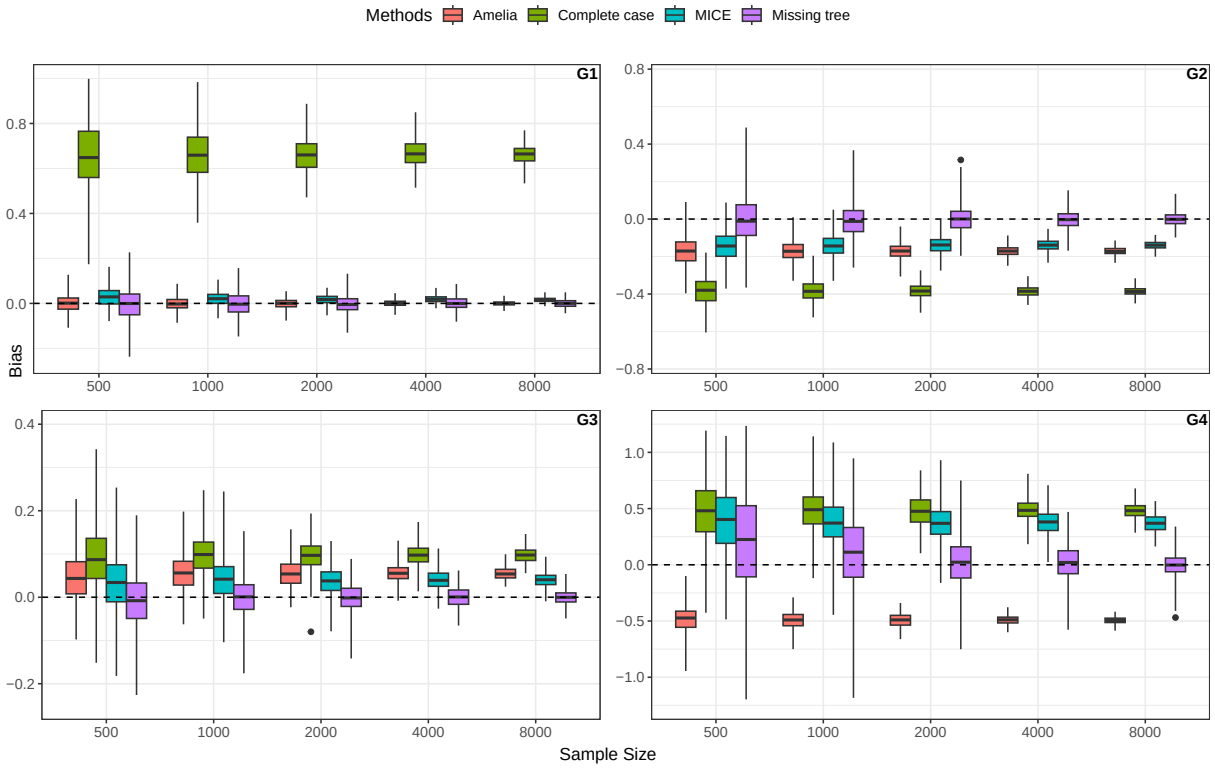


Figure 3: Simulation results for estimation of a mean using four missing data methods: Amelia, complete-case analysis, MICE, and the proposed tree-based method. Panels correspond to data generated under mDAGs $\mathcal{G}_1$ through $\mathcal{G}_4$.

ited substantial bias that did not attenuate with increasing sample size. See Appendix D.2.1 for the DGP and Appendices D.1 and D.3.1 for estimation details.

Task 2: parametric regression. We next assessed inference for regression parameters under missing data. The data-generating process from Task 1 was modified by removing the direct effect of X_1 on X_3, inducing the conditional independence $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ in all mDAGs. We tested this null hypothesis using Wald tests in a correctly specified linear model for X_3 given X_1 and X_2. Performance was evaluated using bias, root mean squared error, type I error, and 95% confidence interval coverage, with results summarized in Table 1. For detailed DGP and estimation derivations see Appendices D.2.2, D.1 and D.3.2, respectively.

Under MAR $\mathcal{G}_1$, all methods achieved nominal performance. Under MNAR $\mathcal{G}_2$, only the

Table 1: Results for conditional independence tests across methods and graphs.

		G1					G2					G3					G4				
	n	500	1000	2000	4000	8000	500	1000	2000	4000	8000	500	1000	2000	4000	8000	500	1000	2000	4000	8000
Amelia	Bias	0.005	-0.006	0.003	0.004	0	-0.157	-0.164	-0.162	-0.157	-0.164	-0.121	-0.134	-0.132	-0.125	-0.128	-0.083	-0.085	-0.082	-0.081	-0.084
	RMSE	0.092	0.059	0.042	0.031	0.021	0.208	0.191	0.178	0.165	0.168	0.204	0.175	0.157	0.137	0.134	0.209	0.163	0.121	0.102	0.095
	Type I Error	0.06	0.054	0.058	0.062	0.056	0.198	0.372	0.604	0.866	0.992	0.106	0.216	0.362	0.526	0.79	0.076	0.086	0.112	0.172	0.342
	Coverage (%)	94	94.6	94.2	93.8	94.4	80.2	62.8	39.6	13.4	0.8	89.4	78.4	63.8	47.4	21	92.4	91.4	88.8	82.8	65.8
Complete case	Bias	0.005	-0.007	0.003	0.004	0	0.04	0.03	0.036	0.043	0.04	0.002	-0.001	-0.003	0	0.004	-0.01	-0.008	0.012	0.003	0.003
	RMSE	0.089	0.06	0.041	0.03	0.02	0.135	0.089	0.073	0.061	0.049	0.15	0.107	0.069	0.049	0.037	0.231	0.158	0.116	0.071	0.055
	Type I Error	0.06	0.058	0.038	0.044	0.054	0.072	0.07	0.098	0.17	0.274	0.056	0.058	0.05	0.042	0.054	0.064	0.052	0.07	0.032	0.058
	Coverage (%)	94	94.2	96.2	95.6	94.6	92.8	93	90.2	83	72.6	94.4	94.2	95	95.8	94.6	93.6	94.8	93	96.8	94.2
MICE	Bias	0.018	0.001	0.009	0.008	0.003	-0.16	-0.188	-0.193	-0.197	-0.209	-0.204	-0.236	-0.256	-0.258	-0.265	-0.221	-0.221	-0.228	-0.239	-0.242
	RMSE	0.094	0.061	0.043	0.031	0.02	0.224	0.218	0.21	0.208	0.216	0.271	0.267	0.272	0.265	0.269	0.394	0.312	0.279	0.262	0.252
	Type I Error	0.088	0.054	0.068	0.06	0.046	0.248	0.406	0.62	0.78	0.938	0.296	0.546	0.818	0.974	1	0.088	0.15	0.23	0.442	0.726
	Coverage (%)	91.2	94.6	93.2	94	95.4	75.2	59.4	38	22	6.2	70.4	45.4	18.2	2.6	0	91.2	85	77	55.8	27.4
Missing tree	Bias	0.003	-0.007	0.004	0.003	0	0.001	-0.006	0	0.007	0.002	0.009	0.005	-0.003	0	0.005	-0.015	-0.007	0.011	0.002	0.004
	RMSE	0.098	0.066	0.048	0.034	0.022	0.142	0.09	0.068	0.047	0.034	0.206	0.148	0.104	0.072	0.054	0.257	0.173	0.127	0.079	0.061
	Type I Error	0.076	0.058	0.056	0.064	0.046	0.07	0.05	0.07	0.058	0.062	0.078	0.066	0.054	0.052	0.054	0.058	0.048	0.06	0.032	0.05
	Coverage (%)	92.2	94.2	94.4	93.6	95.4	93	95	93	94.2	93.8	92.2	93.4	94.6	94.4	94.4	94	95.2	94	96.8	95

proposed method maintained negligible bias, near-nominal type I error, and correct coverage. Competing methods exhibited inflated type I error and poor coverage due to spurious associations induced by conditioning on observed data. For the more complex MNAR models $\mathcal{G}_3$ and $\mathcal{G}_4$, complete-case analysis and the proposed method performed well, while imputation-based methods failed in a model-dependent manner. The unbiasedness of complete-case analysis follows from the imposed graphical structure; see Subsection D.3.2 for a detailed discussion. Overall, while the validity of competing approaches depended on the specific missingness structure, the proposed method consistently achieved correct inference across all settings considered.

Task 3: causal effect estimation. Finally, we evaluated estimation of the causal effect of a binary treatment X_2 on a continuous outcome X_3 using a g-formula (back-door adjustment) controlling for X_1. In this task, X_2 was generated as a binary variable, violating the multivariate normal assumption underlying Amelia. Results are summarized in Figure 4. For detailed DGP and estimation derivations see Appendices D.2.3, D.1 and D.3.3, respectively.

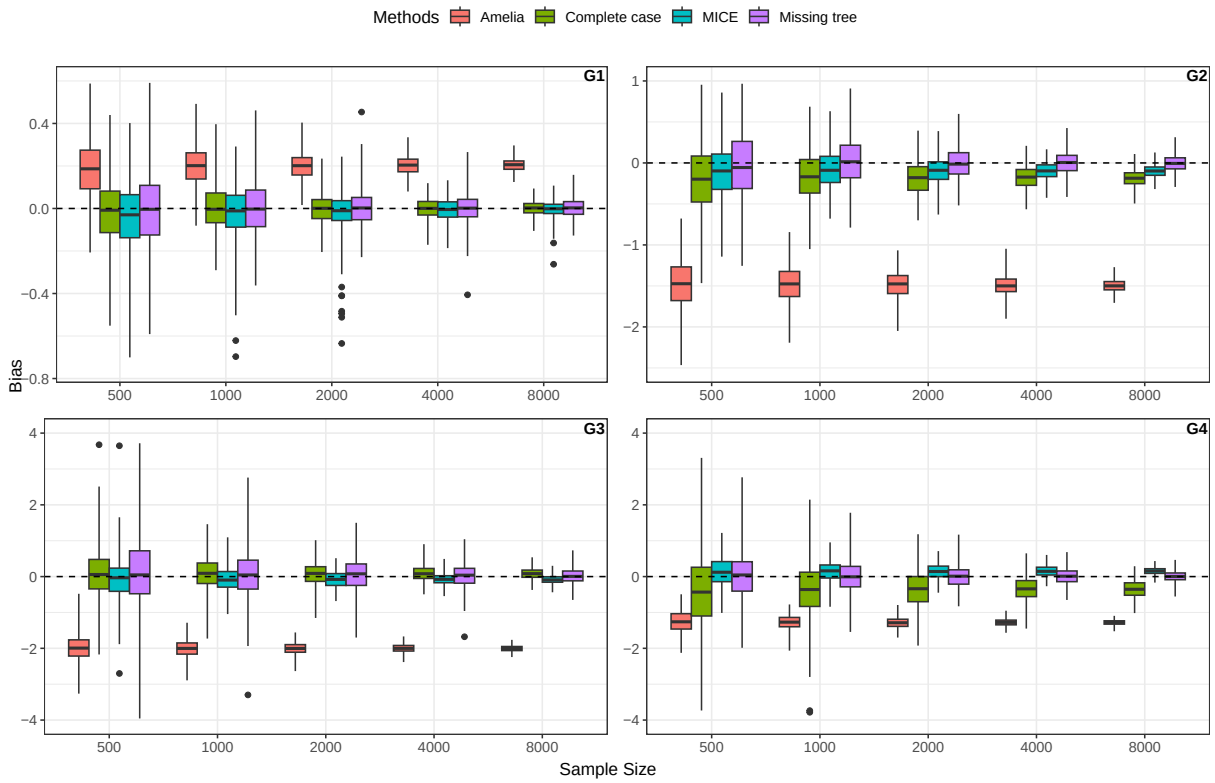


Figure 4: Simulation results for estimating an average causal effect. Missing data are handled using four methods: Amelia, complete-case analysis, MICE, and the proposed tree-based method. Panels corresponds to data generated under mDAGs $\mathcal{G}_1$ through $\mathcal{G}_4$.

Across all mDAGs, the proposed method yielded estimates with negligible bias. Amelia produced biased estimates under all models, with bias persisting as sample size increased. Complete-case analysis and MICE were unbiased under the MAR model $\mathcal{G}_1$, but yielded biased estimates under the MNAR models $\mathcal{G}_2$ through $\mathcal{G}_4$.

7 Data application

We illustrate the proposed framework using data from the Student Feedback Survey for Bachelor Graduates 2016, a national survey of 11,708 Finnish university students who completed a Bachelor's degree or studied for three years in programs without one. The survey collects information on students' academic experiences, study financing, and well-being.

Table 2: Parameter estimates with 95% confidence intervals for the real data application. The estimate of $p(R_1 = 1)$ is $0.853\,(0.847, 0.860)$ and is omitted from the table for brevity.

	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$
$p(X_1 = x_1)$	0.157 (0.149, 0.164)	0.443 (0.433, 0.453)	0.401 (0.391, 0.410)
$p(X_2 = 1 \mid x_1)$	0.370 (0.344, 0.395)	0.432 (0.417, 0.446)	0.473 (0.458, 0.489)
$p(X_2 = 2 \mid x_1)$	0.538 (0.512, 0.564)	0.515 (0.515, 0.545)	0.439 (0.423, 0.454)
$p(X_2 = 3 \mid x_1)$	0.093 (0.078, 0.108)	0.039 (0.033, 0.044)	0.088 (0.079, 0.097)
$p(R_2 = 1 \mid R_1 = 1, x_1)$	0.902 (0.887, 0.916)	0.969 (0.963, 0.974)	0.995 (0.993, 0.997)

Our analysis focuses on two questions concerning study financing. Students were asked whether they funded their studies through student loans and through personal income from work, with response options indicating complete, partial, or no funding. These variables, denoted by X_1 and X_2, exhibit missingness rates of 14.7% and 8.2%, respectively. The relationship between (X_1, X_2) and their missingness indicators is represented by the mDAG shown in Appendix Figure E.2(b), which was used and justified by Tikka & Karvanen (2024). Under this graphical model, the target law is identified. Our objective is to estimate the parameters of the target law, as well as the missingness mechanism. Since both variables are discrete, all relevant components are correctly specified using saturated regressions.

Table 2 reports point estimates and 95% confidence intervals for the target law and missingness parameters, closely matching the maximum likelihood results of Tikka & Karvanen (2024).

8 Discussion

Commonly used complete-case, imputation-based, and EM-based methods can perform poorly under missing-not-at-random mechanisms because they are typically agnostic to the structure of the missingness process. In contrast, the approach developed in this paper explicitly tailors identification and estimation to a specified missingness mechanism. By encoding assumptions about missingness through a graphical model and using this structure

to guide both identification and weighting-based estimation, our framework makes transparent when nonparametric identification and valid inference is possible and why it may fail. This perspective highlights that robustness to MNAR mechanisms does not come from ignoring missingness structure, but rather from leveraging it directly in a principled way.

From an identification standpoint, several important directions for future work remain. First, while this paper focuses on identification of the target law, full law identification (via our intervention/weighted based arguments) remains an open problem in general graphical missing data models. Second, our framework treats the target law as unrestricted; when additional assumptions such as independence constraints on the target law are available, they can facilitate identification and should be incorporated directly into the algorithm. Finally, the proposed identification algorithm is sound but not complete: there exist settings in which the algorithm concludes non-identification even though identification is possible via alternative arguments. For example, in the mDAG shown in Appendix Figure E.2(a), our algorithm fails to identify the target law due to non-identification of a propensity score, yet it can be identified using an odds ratio parameterization, as shown by Nabi et al. (2020). At the same time, odds ratio parameterizations are known to fail in the presence of colluder structures, which are naturally accommodated by the present framework. An important open direction is therefore to study whether different parameterizations and identification strategies can be systematically combined to obtain procedures that are both sound and complete.

A second set of open directions concerns estimation and model selection. While this paper develops recursive inverse probability weighted estimators based on parametric models for the missingness mechanism, future work should investigate more flexible semiparametric and machine learning–based estimators, together with influence function–based inference. In addition, the current framework assumes that the missingness mechanism is specified a

priori that is Markov relative to a conditional mDAG. Developing methods that combine identification and estimation with model selection or discovery of the missingness structure from data is an important and challenging direction. Progress along these lines would further broaden the practical applicability of graphical approaches to missing data, especially in complex observational studies where the missingness mechanism is only partially understood.

References

Bhattacharya, R., Nabi, R. & Shpitser, I. (2022), 'Semiparametric inference for causal effects in graphical models with hidden variables', *Journal of Machine Learning Research* **23**, 1–76.

Bhattacharya, R., Nabi, R., Shpitser, I. & Robins, J. (2019), Identification in missing data models represented by directed acyclic graphs, *in* 'Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI-35th)', AUAI Press.

Daniel, R. M., Kenward, M. G., Cousens, S. N. & De Stavola, B. L. (2012), 'Using causal diagrams to guide analysis in missing data problems', *Statistical Methods in Medical Research* **21**(3), 243–256.

Dawid, A. P. & Didelez, V. (2010), 'Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview', *Statistics Surveys* **4**, 184–231.

Glymour, M. M. (2006), 'Using causal diagrams to understand common problems in social epidemiology', *Methods in social epidemiology* pp. 393–428.

Guo, A., Zhao, J. & Nabi, R. (2023), Sufficient identification conditions and semiparametric estimation under missing not at random mechanisms, *in* 'Uncertainty in Artificial Intelligence', PMLR, pp. 777–787.

Haavelmo, T. (1944), 'The probability approach in econometrics', *Econometrica: Journal of the Econometric Society* pp. iii–115.

Horowitz, J. L. & Manski, C. F. (2000), 'Nonparametric analysis of randomized experiments with missing covariate and outcome data', *Journal of the American statistical Association* **95**(449), 77–84.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

Little, R. & Rubin, D. (2002), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, Wiley.

Manski, C. F. (2005), 'Partial identification with missing data: concepts and findings',

International Journal of Approximate Reasoning **39**(2-3), 151–165.

Martel García, F. (2013), 'Definition and diagnosis of problematic attrition in randomized controlled experiments', *Available at SSRN 2302735* .

Mohan, K. & Pearl, J. (2021), 'Graphical models for processing missing data', *Journal of the American Statistical Association* pp. 1–16.

Mohan, K., Pearl, J. & Tian, J. (2013), Graphical models for inference with missing data, *in* 'Advances in Neural Information Processing Systems 26', Curran Associates, Inc., pp. 1277–1285.

Nabi, R. & Bhattacharya, R. (2022), 'On testability and goodness of fit tests in missing data models', *arXiv preprint arXiv:2203.00132* .

Nabi, R., Bhattacharya, R. & Shpitser, I. (2020), Full law identification in graphical models of missing data: Completeness results., *in* 'Proceedings of the Twenty Seventh International Conference on Machine Learning (ICML-20)'.

Nabi, R., Bhattacharya, R., Shpitser, I. & Robins, J. M. (2023), 'Causal and counterfactual views of missing data models', *Statistica Sinica* .

Nabi, R., Bonvini, M., Kennedy, E. H., Huang, M.-Y., Smid, M. & Scharfstein, D. O. (2024), 'Semiparametric sensitivity analysis: unmeasured confounding in observational studies', *Biometrics* **80**(4), ujae106.

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

Richardson, T. S., Evans, R. J., Robins, J. M. & Shpitser, I. (2023), 'Nested markov properties for acyclic directed mixed graphs', *The Annals of Statistics* **51**(1), 334–361.

Robins, J. M. (1997), 'Non-response models for the analysis of non-monotone non-ignorable missing data', *Statistics in Medicine* **16**, 21–37.

Robins, J. M., Rotnitzky, A. & Scharfstein, D. O. (2000), Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models, *in* 'Statistical models in epidemiology, the environment, and clinical trials', Springer, pp. 1–94.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the american statistical association* **90**(429), 106–121.

Rotnitzky, A., Robins, J. M. & Scharfstein, D. O. (1998), 'Semiparametric regression for repeated outcomes with nonignorable nonresponse', *Journal of the american statistical association* **93**(444), 1321–1339.

Rubin, D. B. (1976), 'Causal inference and missing data (with discussion)', *Biometrika* **63**, 581–592.

Sadinle, M. & Reiter, J. P. (2017), 'Itemwise conditionally independent nonresponse modelling

for incomplete multivariate data', *Biometrika* **104**(1), 207–220.

Scharfstein, D. O. & Irizarry, R. A. (2003), 'Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data', *Biometrics* **59**(3), 601–613.

Shpitser, I. (2016), Consistent estimation of functions of data missing non-monotonically and not at random, *in* 'Proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems (NIPS-16)', Curran Associates, Inc.

Shpitser, I., Mohan, K. & Pearl, J. (2015), Missing data as a causal and probabilistic problem, *in* 'Proceedings of the Thirty First Conference on Uncertainty in Artificial Intelligence (UAI-15)', AUAI Press, pp. 802–811.

Shpitser, I. & Pearl, J. (2006), Identification of joint interventional distributions in recursive semi-Markovian causal models, *in* 'Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)', AAAI Press, Palo Alto.

Spirtes, P., Glymour, C. & Scheines, R. (2001), *Causation, Prediction, and Search*, 2 edn, Springer Verlag, New York.

Sportisse, A., Boyer, C. & Josse, J. (2020), 'Imputation and low-rank estimation with missing not at random data', *Statistics and Computing* **30**(6), 1629–1643.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. & Tchetgen, E. J. T. (2018), 'Semiparametric estimation with data missing not at random using an instrumental variable', *Statistica Sinica* **28**(4), 1965.

Tchetgen, E. J. T., Wang, L. & Sun, B. (2018), 'Discrete choice models for nonmonotone nonignorable missing data: Identification and inference', *Statistica Sinica* **28**(4), 2069.

Thoemmes, F. & Rose, N. (2014), 'A cautious note on auxiliary variables that can increase bias in missing data problems', *Multivariate Behavioral Research* **49**(5), 443–459.

Tian, J. & Pearl, J. (2002), A general identification condition for causal effects, *in* 'Eighteenth National Conference on Artificial Intelligence', pp. 567–573.

Tikka, S. & Karvanen, J. (2024), 'Full law identification under missing data with categorical variables', *arXiv preprint arXiv:2402.05633* .

Wang, S., Shao, J. & Kim, J. K. (2014), 'An instrumental variable approach for identification and estimation with nonignorable nonresponse', *Statistica Sinica* pp. 1097–1116.

Wu, M. C. & Carroll, R. J. (1988), 'Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process', *Biometrics* pp. 175–188.

Zhou, Y., Little, R. J. A. & D., K. J. (2010), 'Block-conditional missing at random models for missing data', *Statistical Science* **25**(4), 517–532.

Appendix

The appendix is organized as follows. Appendix A introduces a glossary of notation and terminology used throughout the paper. Appendix B presents additional examples and technical details that supplement the identification results in Sections 3 and 4, and the estimation results in Section 5. Appendix B.1 contains worked examples of the tree-based identification algorithm. Appendix B.2 contains worked examples with step-by-step derivations for example estimation of means, regression coefficients, and causal effects. Appendix B.3 provides pseudocode for the proposed estimation algorithms for propensity score estimations outlined in Section 5.1 and estimation of functionals of the target law outlined in Section 5.2. Appendix C contains proofs of the main theoretical results. Appendix D provides full details of the simulation study, including data-generating processes for all graphs and tasks, additional figures, and supplementary tables.

A Glossary of terms and notations

Table 3: Glossary of terms and notations

Symbol	Definition
X, X_k	Vector of variables, k-th element of X
$p(X), \mathcal{M}_X$	Target law, model of $p(X)$
R, R_k	Missingness indicators of X, X_k
$R_k = 1 / R_k = 0$	X_k missing/observed
$p(R \mid X), \mathcal{M}_{R\|X}$	Missingness mechanism, model of $p(R \mid X)$
$p(X, R)$	Full law
$\mathcal{M} = \mathcal{M}_X \otimes \mathcal{M}_{R\|X}$	Model of $p(X, R)$
X^*, X_k^*	Coarsened version of X, X_k
$p(X^*, R)$	Observed data law
$\mathrm{pa}_{\mathcal{G}}(R_k)$	Parents of R_k in $\mathcal{G}$
$\mathrm{de}_{\mathcal{G}}(R_k)$	Descendants of R_k in $\mathcal{G}$, including R_k
$\mathrm{nd}_{\mathcal{G}}(R_k)$	Non-descendants of R_k, defined as $X \cup R \setminus \mathrm{de}_{\mathcal{G}}(R_k)$
$\theta(p(X))$	Particular functional of $p(X)$
$p(\cdot)\|_{R_j=1}$	Evaluation of $p(\cdot)$ at $R_j = 1$
$\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))$	Propensity score of R_k, defined as $p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))$
$\mathcal{S}_k^x$	Counterfactual-induced selection set, defined as $\{R_j \in R : X_j \in \mathrm{pa}_{\mathcal{G}}(R_k)\}$
$\mathcal{S}_k^r$	Indicator-induced selection set
$\mathcal{S}_k = \mathcal{S}_k^x \cup \mathcal{S}_k^r$	Selection set
$\mathcal{R}_k^p = \mathcal{S}_k^x \cap \mathrm{de}_{\mathcal{G}}(R_k)$	Problematic set
$\mathrm{do}(R_j = 1)$	An intervention on R_j setting it to 1
$\mathcal{S}_{j\downarrow k} \coloneqq \mathcal{S}_j \cap \mathrm{pa}_{\mathcal{G}}(R_k)$	Selection propagated from $\mathrm{do}(R_j)$ to R_k
$\mathbb{T}_k$	Identification tree associated with R_k
$\mathbb{F}$	Forest that collects all trees
$\mathcal{D}$	Collection of indicators whose propensity score is not identified
τ	A valid reversed topological order on the mDAG $\mathcal{G}$
$\mathcal{C}_{k,k}^{\mathrm{dir}}$	Colluder descendants of R_k, defined as $\{R_j \in \mathrm{de}_{\mathcal{G}}(R_k) \,:\, X_k \in \mathrm{pa}_{\mathcal{G}}(R_j)\}$
R^*	Candidate intervention set for R_k
$\mathcal{T}_k \coloneqq \mathrm{ch}_{\mathbb{T}_k}(R_k)$	Children of R_k in $\mathbb{T}_k$
$\widetilde{\mathcal{S}}_k \coloneqq \mathcal{S}_k^x \cup \bigcup_{R_j \in \mathcal{T}_k} \mathcal{S}_j$	Pre-selection set for R_k
$p(\cdot \mid \mathrm{do}(\mathcal{T}_k = 1))$	Post-intervention distribution where indicators in $\mathcal{T}_k$ are intervened on
$\mathcal{R}_k^d$	Indicator in $\widetilde{\mathcal{S}}_k \setminus \mathrm{pa}_{\mathcal{G}}(R_k)$ which is dependent on R_k given $\mathrm{pa}_{\mathcal{G}}(R_k)$ in $p(\cdot \mid \mathrm{do}(\mathcal{T}_k = 1))$
$\mathcal{C}_{k,j}^{\mathrm{dir}}$	Descendants of R_k that selection on R_j, defined as $\{R_i \in \mathrm{de}_{\mathcal{G}}(R_k) \mid X_j \in \mathrm{pa}_{\mathcal{G}}(R_i)\}$
$\mathcal{C}_k \coloneqq \bigcup_{R_j \in \mathcal{R}_k^d} \mathcal{C}_{k,j}^{\mathrm{dir}}$	Descendants of R_k that select on $\mathcal{R}_k^d$
$\mathcal{B}$	Collection of pruned branches
$\phi_{R_i}^{\mathcal{G}}$	Graphical fixing operation applied to R_i on an mDAG $\mathcal{G}$
$\phi_{R_i}^{p}\{p\}$	Probabilistic fixing operator applied to R_i in $p(X, R)$
$\sigma_k = (s_1, \ldots, s_m)$	Any ordering of $\mathcal{T}_k$ consistent with τ
$p_{\mathbb{T}_k} = \phi_{\sigma_k}^{p}\{p\}$	The post-intervention distribution induced by $\mathbb{T}_k$
$\mathcal{P}$	Collection of fitted propensity score models
$\mathcal{E}$	Collection of estimating equations
$W_k(\theta_{\mathcal{T}_k})$	Inverse propensity weight for estimating π_k, defined in (18)
$\mathcal{P}_k$	Collection of re-fitted propensity score models tailored for estimating π_k
$\mathcal{E}_k$	Collection of re-constructed estimating equations tailored for estimating π_k
$\boldsymbol{\theta}_k$	Stacked parameter vector relevant for estimating θ_k
$\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)$	Stacked estimating function corresponding to $\boldsymbol{\theta}_k$
$M(\tilde{X}; \theta)$	Estimating function for θ
$\tilde{X}$	Collection of variables required to evaluate M
$\tilde{R}$	Missingness indicators for $\tilde{X}$
$\mathrm{cl}(\cdot)$	Closure operation defined as $\mathrm{cl}(A) \;\coloneqq\; A \cup \bigcup_{R_i \in A} \mathcal{S}_i$
$\mathcal{R}$	Smallest set containing $\tilde{R}$ and is closed under inclusion of $\mathcal{S}_i$
$\theta_{\mathcal{R}} \coloneqq (\theta_i)_{R_i \in \mathcal{R}}$	Collection of parameters indexing propensity score $R_i \in \mathcal{R}$

Table 4: Key definitions used in the tree-based identification algorithm, illustrated using the mDAG in Figure 2(a).

R_k	**Propensity scores**	$\mathcal{S}_k^x$	$\mathcal{R}_k^p$	$C_{k,k}^{\text{dir}}$	$\text{ch}_{\mathbb{T}_k}(R_k)$	$\tilde{\mathcal{S}}_k$	$\mathcal{S}_k^r$	$\mathcal{S}_k$
R_1	$p(R_1 \mid R_2, R_3)$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
R_2	$p(R_2 \mid X_1, X_3, R_4)$	$\{R_1, R_3\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$	$\{R_1, R_3\}$	$\emptyset$	$\{R_1, R_3\}$
R_3	$p(R_3 \mid X_1, X_2, R_4)$	$\{R_1, R_2\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$	$\{R_1, R_2\}$	$\emptyset$	$\{R_1, R_2\}$
R_4	$p(R_4 \mid X_1, X_2, X_3)$	$\{R_1, R_2, R_3\}$	$\{R_1, R_2, R_3\}$	$\emptyset$	$\{R_1, R_2, R_3\}$	$\{R_1, R_2, R_3\}$	$\emptyset$	$\{R_1, R_2, R_3\}$

B Additional examples and details

B.1 Examples illustrating key identification concepts

B.1.1 Identification example via Figures 2(a, d)

The mDAG $\mathcal{G}_1$ in Figure 2(a) is an example where the tree-prune procedure never gets executed. The algorithm begins with R_1, for which $\pi_1 \coloneqq p(R_1 = 1 \,|\, R_2, R_3)$ is directly observed and therefore identified. It then proceeds to either R_2, followed by R_3, or vice versa. Since neither variable is a descendant of the other, both orders are valid. Their propensity scores are identified analogously, so we focus on R_2 for illustration. $\pi_2 \coloneqq p(R_2 = 1 \,|\, X_1, X_3, R_4)$, with $\mathcal{S}_2^x = \{R_1, R_3\}$ and $\mathcal{R}_2^p = \{R_1\}$. An intervention on R_1 is applied to invoke $R_2 \perp R_1 \,|\, \text{pa}_{\mathcal{G}}(R_2)$ in the resulting post-intervention distribution. This intervention does not impose any additional selection ($\mathcal{S}_1 = \emptyset$), thus $\widetilde{\mathcal{S}}_2 = \mathcal{S}_2^x$, and consequently $\mathcal{S}_2^r = \widetilde{\mathcal{S}}_2 \cap \text{pa}_{\mathcal{G}}(R_2) = \emptyset$. Thus, π_2 is identified as a full conditional distribution. Finally, the algorithm considers R_4, for which $\pi_4 \coloneqq p(R_4 \,|\, X_1, X_2, X_3)$, with $\mathcal{S}_4^x = \mathcal{R}_4^p = \{R_1, R_2, R_3\}$. π_4 is identified by intervening on all indicators in $\mathcal{R}_4^p$. Since $\mathcal{R}_k^d = \emptyset$, no pruning is needed when appending constructed subtrees to the tree for R_4. As a result, all propensity scores, each evaluated at its corresponding indicator evaluation set which in this case is empty, are identified, and hence the target law is identified. Key definitions of the identification procedure, including $\mathcal{S}_k^x$, $\mathcal{R}_k^p$, $\mathcal{C}_{k,k}^{\text{dir}}$, $\text{ch}_{\mathbb{T}_k}(R_k)$, $\widetilde{\mathcal{S}}_k$, and $\mathcal{S}_k$ for each indicator R_k, are summarized in Appendix Table 4.

B.1.2 Identification example via Figures 2(b, e)

The mDAG $\mathcal{G}_2$ in Figure 2(b) is an example where tree-prune gets executed. With the pruning, all selection biases can be avoided and all the propensity scores under proper evaluation are identified. The identification algorithm begins with R_1, followed by R_2, whose propensity scores are easily identified via the associational irrelevancy criterion. We then focus on R_3, for which $\pi_3 \coloneqq p(R_3 = 1 \,|\, X_1, R_4, R_5, R_6)$, with $\mathcal{R}_3^p = \{R_1\}$. Intervening on both R_1 and R_2 makes π_3 evaluated at $\mathcal{S}_3^r = \{R_4, R_5\}$ identified. In this evaluation, R_4 results from the intervention on R_1, whereas R_5 results from the intervention on R_2. The algorithm can then proceed to either R_4 or R_5, and then to the other. For illustration, we first consider R_4, for which $\pi_4 \coloneqq p(R_4 = 1 \,|\, X_3)$ with $\mathcal{R}_4^p = \{R_3\}$ and $\mathcal{C}_{4,4}^{\text{dir}} = \{R_1\}$. Thus, only R_2 and R_3 are qualified candidates for intervention. The subtree for R_2 attaches directly to the tree of R_4, whereas the subtree for R_3 requires pruning. This is because, in $\mathbb{T}_3$, the branch $R_1 \in \mathcal{C}_{4,4}^{\text{dir}}$ induces selection on R_4, which propagates through R_3 since $R_4 \in \text{pa}_{\mathcal{G}}(R_3)$. Consequently, R_1 must be pruned from $\mathbb{T}_3$ before $\mathbb{T}_3$ is appended to $\mathbb{T}_4$. After pruning, the selection set of R_3, $\mathcal{S}_3$, is updated from $\{R_1, R_4, R_5\}$ to $\{R_1, R_5\}$, and π_4 becomes identified. The propensity score of R_5, evaluated at $\mathcal{S}_5^r = \{R_6\}$, is identified analogously as that of R_6, and is therefore omitted. Finally, the algorithm turns to R_6, for which $\pi_6 \coloneqq p(R_6 = 1 \,|\, X_3, X_4)$, with $\mathcal{R}_6^p = \{R_3\}$ and $\mathcal{C}_{6,6}^{\text{dir}} = \{R_5\}$. The qualified candidates for intervention are therefore $\{R_1, R_2, R_3\}$ (note that R_4 is not a descendant of R_6 and thus excluded). In the initial assessment, identification is hindered by selection on $\mathcal{R}_6^d = \{R_5\}$, which arises from branches R_2 and R_3. Intervening on R_2 directly selects on R_5, and thus R_2 must be pruned from the tree $\mathbb{T}_6$. In contrast, the selection of R_5 induced by intervening on R_3 arises because identifying π_3 involves an unnecessary intervention on R_2. As a result, R_2 is pruned in the subtree $\mathbb{T}_3$. After pruning, π_6 becomes identified. Appendix Table 5 summarizes key definitions for the identification procedure.

Table 5: Key definitions used in the tree-based identification algorithm, illustrated using the mDAG in Figure 2(b).

R_k	**Propensity scores**	$\mathcal{S}_k^x$	$\mathcal{R}_k^p$	$C_{k,k}^{\text{dir}}$	$\text{ch}_{\mathbb{T}_k}(R_k)$	$\tilde{\mathcal{S}}_k$	$\mathcal{S}_k^r$	$\mathcal{S}_k$
R_1	$p(R_1 \mid X_4, R_2)$	$\{R_4\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{R_4\}$	$\emptyset$	$\{R_4\}$
R_2	$p(R_2 \mid X_5, R_3)$	$\{R_5\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{R_5\}$	$\emptyset$	$\{R_5\}$
R_3	$p(R_3 \mid X_1, R_4, R_5, R_6)$	$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_1, R_2\}$	$\{R_1, R_4, R_5\}$	$\{R_4, R_5\}$	$\{R_1, R_4, R_5\}$
R_3 ($R_3 \to R_1$ pruned)		$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_2\}$	$\{R_1, R_5\}$	$\{R_5\}$	$\{R_1, R_5\}$
R_3 ($R_3 \to R_2$ pruned)		$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$	$\{R_1, R_4\}$	$\{R_4\}$	$\{R_1, R_4\}$
R_4	$p(R_4 \mid X_3)$	$\{R_3\}$	$\{R_3\}$	$\{R_1\}$	$\{R_2, R_3\}$	$\{R_3, R_5\}$	$\emptyset$	$\{R_3\}$
R_5	$p(R_5 \mid X_3, X_6, R_6)$	$\{R_3, R_6\}$	$\{R_3\}$	$\{R_2\}$	$\{R_1, R_3\}$	$\{R_1, R_3, R_4, R_6\}$	$\{R_6\}$	$\{R_3, R_6\}$
R_6	$p(R_6 \mid X_3, X_4)$	$\{R_3, R_4\}$	$\{R_3\}$	$\{R_5\}$	$\{R_1, R_3\}$	$\{R_1, R_3, R_4\}$	$\emptyset$	$\{R_3, R_4\}$

B.1.3 Identification example via Figures 2(c, f)

The mDAG $\mathcal{G}_3$ in Figure 2(c) is an example where tree-prune gets executed. However, some selections cannot be avoided. Thus, the algorithm concludes the target law is not identified. We omit the discussion of R_1 through R_4, as the technique for identifying their propensity scores is the same as described above. We now focus on R_5, whose propensity score is $\pi_5 \coloneqq p(R_5 = 1 \,|\, X_2)$, with $\mathcal{R}_5^p = \{R_2\}$ and $\mathcal{C}_{5,5}^{\text{dir}} = \{R_2\}$. The candidates for intervention are therefore $\{R_1, R_3, R_4\}$. The subtrees for R_1 and R_3 can be attached directly, whereas the subtree for R_4 is pruned at R_2 during the initial assessment of identifiability, where $\mathcal{R}_5^d = \{R_5\}$. This pruning is necessary as intervenetion on R_2 selects on R_5, and this selection progagates through R_4 given that $R_5 \in \text{pa}_{\mathcal{G}}(R_4)$. After pruning, the propensity score for R_4 is no longer identified, thus R_4 is pruned from $\mathbb{T}_5$. The second assessment returns $\mathcal{R}_5^d = \{R_4\}$, which arises from the intervention on R_1. This leads to pruning R_1, leaving R_3 as the only indicator on which an intervention is applied. However, an intervention on R_3 alone is insufficient to identify π_5, because in the third assessment we obtain $\mathcal{R}_5^d = \{R_2\} = \mathcal{R}_5^p$, which implies that identification fails. Details are provided in Appendix Table 6.

Table 6: Key definitions used in the tree-based identification algorithm, illustrated using the mDAG in Figure 2(c).

R_k	**Propensity scores**	$\mathcal{S}_k^x$	$\mathcal{R}_k^p$	$\mathcal{C}_{k,k}^{\text{dir}}$	$\text{ch}_{\mathbb{T}_k}(R_k)$	$\tilde{\mathcal{S}}_k$	$\mathcal{S}_k^r$	$\mathcal{S}_k$
R_1	$p(R_1 \mid X_4, R_2)$	$\{R_4\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{R_4\}$	$\emptyset$	$\{R_4\}$
R_2	$p(R_2 \mid X_5, R_3, R_4)$	$\{R_5\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{R_5\}$	$\emptyset$	$\{R_5\}$
R_3	$p(R_3 \mid R_5)$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
R_4	$p(R_4 \mid X_1, R_5)$	$\{R_1\}$	$\{R_1\}$	$\{R_1\}$	$\{R_2\}$	$\{R_1, R_5\}$	$\{R_5\}$	$\{R_1, R_5\}$
R_5	$p(R_5 \mid X_2)$ not ID	$\{R_2\}$	$\{R_2\}$	$\{R_2\}$	-	-	-	-

B.1.4 Identification example via Figures E.2(c, d)

The mDAG in Appendix Figure E.2(c) provides an example in which pruning requires updating a grown subtree to match its pruned version. We omit the discussion of R_1 through R_4, as constructing their trees is straightforward and does not involve pruning. We now focus on R_5, whose propensity score is $\pi_5 \coloneqq p(R_5 = 1 \mid X_2)$, with $\mathcal{R}_5^p = \{R_3\}$ and $\mathcal{C}_{5,5}^{\text{dir}} = \{R_2\}$. The candidates for intervention are therefore $\{R_1, R_3, R_4\}$. The initial assessment of identifiability yields $\mathcal{R}_5^d = R_5$, indicating that intervening on these candidates induce selection on R_5 itself. This selection arises from intervening on R_2. To remove this selection, the edge $R_3 \to R_2$ in subtree $\mathbb{T}_3$ is pruned, and π_3 remains identifiable. The resulting pruned $\mathbb{T}_3$ is added to the pruned-tree collection $\mathcal{B}$. Similarly, the edge $R_4 \to R_2$ in subtree $\mathbb{T}_4$ is pruned. In addition, the subtree $\mathbb{T}_3$ attached to R_3 in $\mathbb{T}_4$ is replaced to match its pruned version in $\mathcal{B}$. Otherwise, selection induced by R_2 would propagate to R_3 and further through R_4, thereby obstructing identification of π_5. Collecting pruned subtrees and reusing them when needed allows the algorithm to restrict attention to pruning children of R_k, or their children, without requiring consideration of deeper descendants. Selection arising from further descendants is handled automatically by replacing the relevant subtrees with their pruned versions when available in $\mathcal{B}$. This design substantially improves the feasibility of the identification algorithm, as exhaustively tracing selection through all descendants is computationally expensive, especially

Table 7: Key definitions used in the tree-based identification algorithm, illustrated using the mDAG in Figure E.2(c).

R_k	**Propensity scores**	$\mathcal{S}_k^x$	$\mathcal{R}_k^p$	$\mathcal{C}_{k,k}$	$\mathrm{ch}_{\mathbb{T}_k}(R_k)$	$\tilde{\mathcal{S}}_k$	$\mathcal{S}_k^r$	$\mathcal{S}_k$
R_1	$p(R_1 \mid R_2)$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
R_2	$p(R_2 \mid X_5, R_3)$	$\{R_5\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{R_5\}$	$\emptyset$	$\{R_5\}$
R_3	$p(R_3 \mid X_1, R_4, R_5)$	$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_1, R_2\}$	$\{R_1, R_5\}$	$\{R_5\}$	$\{R_1, R_5\}$
R_3	($R_3 \to R_2$ pruned)	$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$
R_4	$p(R_4 \mid X_3, R_5)$	$\{R_3\}$	$\{R_3\}$	$\emptyset$	$\{R_1, R_2, R_3\}$	$\{R_1, R_3, R_5\}$	$\{R_5\}$	$\{R_1, R_3, R_5\}$
R_4	($R_4 \to R_2$ & $\mathbb{T}_3$ pruned)	$\{R_3\}$	$\{R_3\}$	$\emptyset$	$\{R_1, R_3\}$	$\{R_1, R_3\}$	$\{R_5\}$	$\{R_1, R_3\}$
R_5	$p(R_5 \mid X_3)$	$\{R_3\}$	$\{R_3\}$	$\{R_2\}$	$\{R_1, R_3, R_4\}$	$\{R_1, R_3\}$	$\emptyset$	$\{R_3\}$

in large graphs. Details are provided in Appendix Table 7.

B.2 Examples illustrating key estimation concepts

We provide detailed illustrations of the estimation procedures developed in Section 5.2. These examples illustrate the closure construction for $\mathcal{R}$, the associated inverse probability weighted estimating equations, and the stacked variance calculations within concrete graphical models. All examples rely on the intervention trees and selection sets produced by Algorithm 1; no additional identification arguments are introduced.

We focus on three inferential tasks: (i) **estimation of means**, (ii) **parametric regression coefficients**, and (iii) **causal effects**. The first two examples illustrate estimation of means in settings where the full target law is not identifiable, highlighting both cases in which a mean remains identifiable and cases in which identification fails due to unavoidable selection. The remaining examples demonstrate how the same framework extends to regression and causal parameters once the appropriate closure of missingness indicators is taken into account.

Throughout, $\tilde{R}$ denotes the set of indicators corresponding to variables required to evaluate a target estimating function, and $\mathcal{R}$ denotes its closure under the selection sets $\{\mathcal{S}_i\}$ as defined in (24). The estimating function Ψ is constructed according to (25), and asymptotic variances

are obtained from the stacked estimating equations described in Section 5.2.

B.2.1 Estimation task: Mean of a partially observed variable

We begin with estimation of the mean $\theta = \mathbb{E}(X_3)$ in the mDAG shown in Figure 2(c). The parameter θ is defined as the unique solution to the moment condition

$$\mathbb{E}\{M(\tilde{X}; \theta)\} = 0, \qquad M(\tilde{X}; \theta) = X_3 - \theta.$$

Although the target law $p(X)$ is not identified in this mDAG because the propensity score π_5 is not identifiable, the mean $\mathbb{E}(X_3)$ remains identifiable. Evaluating $M(\tilde{X}; \theta)$ requires observing X_3, thus we initialize $\tilde{R} = \{R_3\}$. The selection set associated with R_3 is empty, implying that the closure construction in (24) stabilizes immediately and yields $\mathcal{R} = \{R_3\}$.

The estimating function for θ therefore takes the inverse probability weighted form

$$\Psi(X, R; \theta, \theta_3) = \frac{\mathbb{I}(R_3 = 1)}{\pi_3(\mathrm{pa}_{\mathcal{G}}(R_3); \theta_3)|_{\mathcal{S}_3^r = 1}} (X_3 - \theta), \tag{27}$$

where π_3 denotes the propensity score for R_3. The corresponding propensity score estimating equation is

$$\Psi_3(X, R; \theta_3) = R_3 - \pi_3(\mathrm{pa}_{\mathcal{G}}(R_3); \theta_3). \tag{28}$$

An estimator $\hat{\theta}$ is obtained by solving $P_n \Psi(X, R, \theta, \hat{\theta}_3) = 0$, where $\hat{\theta}_3$ solves $P_n \Psi_3(X, R; \theta_3) = 0$. The asymptotic variance of $\hat{\theta}$ follows from the stacked estimating equations $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_3, \Psi)'$ as described in Section 5.2, where $\boldsymbol{\Psi}_3 = \{\Psi_3\}$.

By contrast, identification of the mean of other variables fails in this mDAG. For example,

consider $\theta = \mathbb{E}(X_1)$. The initial indicator set is $A_0 = \tilde{R} = \{R_1\}$. According to Appendix Table 6, π_1 can be evaluated only on $\mathcal{S}_1 = \{R_4\}$, yielding $A_1 = \text{cl}(A_0) = \{R_1, R_4\}$. Next, π_4 can be evaluated only on $\mathcal{S}_4 = \{R_1, R_5\}$, yielding $A_2 = \text{cl}(A_1) = \{R_1, R_4, R_5\}$. Since π_5 is not identified, the closure includes an indicator whose propensity score cannot be estimated, and no unbiased estimating equation for θ can be constructed. Thus $\mathbb{E}(X_1)$ is not identifiable in this graph.

B.2.2 Estimation task: Mean requiring recursive closure

Consider the modification of the mDAG in Figure 2(c), shown in Appendix Figure E.1(e), in which the edge $R_4 \to R_2$ is replaced by $R_4 \to R_3$. The corresponding intervention trees are shown in Appendix Figure E.1(f).

Under this modification, additional propensity scores become identifiable. In particular, π_4, evaluated at $R_5 = 1$, is identified by intervening on R_2 and R_3, with the resulting selection set remaining $\mathcal{S}_4 = \{R_1, R_5\}$. Moreover, π_5 becomes identifiable by intervening on R_1, R_3, and R_4, with pruning of R_2 from the subtree of R_4 to avoid inducing selection on R_5. The resulting selection set for R_5 is $\mathcal{S}_5 = \{R_2\}$.

We again consider estimation of $\theta = \mathbb{E}(X_1)$. Starting from $A_0 = \tilde{R} = \{R_1\}$, the closure construction proceeds as follows:

$$A_1 = \text{cl}(A_0) = \{R_1, R_4\}, \qquad A_2 = \text{cl}(A_1) = \{R_1, R_4, R_5\},$$

and incorporating $\mathcal{S}_5 = \{R_2\}$ yields

$$A_3 = \text{cl}(A_2) = \{R_1, R_2, R_4, R_5\}.$$

Since $\mathcal{S}_2 = \{R_5\} \subseteq A_3$, the recursion stabilizes and $\mathcal{R} = \{R_1, R_2, R_4, R_5\}$.

The estimating function for θ is therefore

$$\Psi(X^*, R; \theta, \theta_{\mathcal{R}}) = \frac{\mathbb{I}(\mathcal{R} = 1)}{\prod_{R_i \in \mathcal{R}} \pi_i(\text{pa}_{\mathcal{G}}(R_i); \theta_i)} (X_1 - \theta), \tag{29}$$

with estimating functions for the propensity scores of $R_i \in \mathcal{R}$, defined by the intervention trees in Appendix Figure E.1(f), given by

$$\Psi_1(X^*, R; \theta_1) = \mathbb{I}(R_4 = 1)\ (R_1 - \pi_1(X_4, R_2; \theta_1)). \tag{30}$$

$$\Psi_2(X^*, R; \theta_2) = \mathbb{I}(R_5 = 1)\ (R_2 - \pi_2(X_5, R_3; \theta_2)). \tag{31}$$

$$\Psi_4(X^*, R; \theta_4, \theta_2, \theta_3) = \frac{\mathbb{I}(R_1 = 1, R_5 = 1)\ \mathbb{I}(R_2 = 1, R_3 = 1)}{\pi_2(X_5, R_3; \theta_2)\ \pi_3(R_4, R_5; \theta_3)} (R_4 - \pi_4(X_1, R_5; \theta_4)). \tag{32}$$

$$\Psi_5(X^*, R; \theta_5, \theta_1, \theta_3, \theta_4) = \frac{\mathbb{I}(R_2 = 1)\ \mathbb{I}(R_1 = 1, R_3 = 1, R_4 = 1)}{\pi_1(X_4, R_2; \theta_1)\ \pi_3(R_5; \theta_3)\ \pi_4(X_1, R_5; \theta_4)} (R_5 - \pi_5(X_2; \theta_5)). \tag{33}$$

The asymptotic variance of $\hat{\theta}$ is obtained from the stacked estimating equations $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \boldsymbol{\Psi}_4, \boldsymbol{\Psi}_5, \Psi)'$, where $\boldsymbol{\Psi}_k$ denotes the stacked estimating functions associated with Ψ_k. The propensity scores π_1 and π_2 are identified via associational irrelevancy, so that $\boldsymbol{\Psi}_1 = \{\Psi_1\}$ and $\boldsymbol{\Psi}_2 = \{\Psi_2\}$. In contrast, π_4 and π_5 are identified via causal irrelevancy, and their stacked systems additionally include estimating functions for intervened indicators and their descendants. Specifically, $\boldsymbol{\Psi}_4 = \{\Psi_2, \Psi_3, \Psi_4\}$ and $\boldsymbol{\Psi}_5 = \{\Psi_1, \Psi_3, \Psi_4, \Psi_5\}$, where Ψ_3 is given by (28). The Ψ_4 appearing in $\boldsymbol{\Psi}_4$ is defined in (32). In contrast, the Ψ_4 included in $\boldsymbol{\Psi}_5$, given by (34), is different and is specifically constructed for estimating π_5. In particular, π_2 is dropped from the inverse weight, as dictated by the intervention tree in Appendix Figure E.1(f), where R_2 is pruned from the subtree of R_4 before it is appended to the tree of R_5.

$$\Psi_4(X, R; \theta_4, \theta_3) = \frac{\mathbb{I}(R_1 = 1)\ \mathbb{I}(R_3 = 1)}{\pi_3(R_4, R_5; \theta_3)} (R_4 - \pi_4(X_1, R_5; \theta_4)). \quad \text{(tailored for estimating } \pi_5) \tag{34}$$

B.2.3 Estimation task: Parametric regression with missing covariates

Using the mDAG in Appendix Figure E.1(e), consider estimation of the regression of X_3 on (X_1, X_2) under the linear mean model

$$\mathbb{E}(X_3 \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

with target parameter $\theta = (\beta_0, \beta_1, \beta_2)$. The corresponding moment condition is

$$M(\tilde{X}; \theta) = f(X_1, X_2)\,\{X_3 - (\beta_0 + \beta_1 X_1 + \beta_2 X_2)\}, \qquad f(X_1, X_2) = (1, X_1, X_2)'. \tag{35}$$

Evaluating M requires observing (X_1, X_2, X_3), so the initial indicator set is $\tilde{R} = \{R_1, R_2, R_3\}$. Applying the closure construction yields $\mathcal{R} = \{R_1, R_2, R_3, R_4, R_5\}$. The estimating function Ψ follows from (25), and the asymptotic variance of $\hat{\theta}$ is obtained by stacking the estimating equations for all propensity scores in $\mathcal{R}$ together with Ψ. That is, $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \boldsymbol{\Psi}_3, \boldsymbol{\Psi}_4, \boldsymbol{\Psi}_5, \Psi)'$, where specifications of $\boldsymbol{\Psi}_1$ through $\boldsymbol{\Psi}_5$ are given in Subsections B.2.1 and B.2.2.

B.2.4 Estimation task; Average causal effect

Finally, we consider estimation of the average causal effect of a binary treatment X_1 on an outcome X_3. For clarity, we focus on the counterfactual mean $\theta = \mathbb{E}(X_3^{x_1})$ for a fixed treatment level $x_1 \in \{0, 1\}$. In the absence of missing data, θ is identified under standard causal assumptions using adjustment for the confounder X_2.

We assume that θ, together with nuisance parameters θ_{trt} and θ_{or} indexing a treatment model

and an outcome regression, respectively, is characterized by the moment condition

$$M(\tilde{X};\theta,\theta_{trt},\theta_{or}) = \begin{pmatrix} f_{trt}(X_2)\,\{X_1 - p(X_1 \mid X_2;\theta_{trt})\} \\ f_{or}(X_1,X_2)\,\{X_3 - \mathbb{E}(X_3 \mid X_1,X_2;\theta_{or})\} \\ \frac{\mathbb{I}(X_1=x_1)}{p(X_1|X_2;\theta_{trt})}\{X_3 - \mathbb{E}(X_3 \mid x_1,X_2;\theta_{or})\} + \mathbb{E}(X_3 \mid x_1,X_2;\theta_{or}) - \theta \end{pmatrix}. \quad (36)$$

In the function M, $f_{trt}(X_2)$ and $f_{or}(X_1,X_2)$ have dimensions matching θ_{trt} and θ_{or}, respectively. We leave their specific forms unspecified, assuming only that they follow some parametric structure. For example, if the relationship between X_3 and the covariates is fully captured by a linear regression with main terms only, the second row of M follows the form in (35), and f_{or} can follow the form of f in Subsection B.2.3 or any three-dimensional function of X_1 and X_2 that is feasible.

The missingness-adjusted estimating function Ψ is obtained by substituting (36) into (25), with $\mathcal{R}$ determined by the closure construction for the indicators required to evaluate M. Estimation and inference then follow from the general theory in Section 5.2.

B.3 Estimation algorithm

See Algorithm E1 for a summary description of our proposed *recursive inverse probability weighted* estimators for the identified propensity scores $\{\pi_k|_{S_k^r=1}\}_{k=1}^K$.

See Algorithm E2 for a summary description of estimation of generic parameters defined through moment conditions under the target law.

Algorithm E1 RECURSIVE IPW ESTIMATION OF PROPENSITY SCORES$(\mathcal{G}(X, R, X^*), \mathbb{F}, \mathcal{D})$

1: **Estimation Procedure:**

- Let $\mathcal{E}$ and $\mathcal{P}$ be the lists that collect the estimating equations and fitted models of propensity scores for all $R_k \notin \mathcal{D}$, respectively; initialize them to empty lists
- For $R_k \in R \setminus \mathcal{D}$ (following a valid reverse topological order τ):
  - $\hat{\pi}_k := \pi_k(\text{pa}_{\mathcal{G}}(R_k); \hat{\theta}_k)|_{\mathcal{S}_k^r=1} \leftarrow$ **estimate-propensity**$(R_k, \mathcal{P}, \mathbb{F})$
  - Add $\hat{\pi}_k$ to $\mathcal{P}$, and Ψ_k to $\mathcal{E}$
- RETURN $\mathcal{E}, \mathcal{P}$

2: **estimate-propensity**$(R_k, \mathcal{P}, \mathbb{F})$

- Assume the parameter θ_k indexing $\pi_k(\text{pa}_{\mathcal{G}}(R_k); \theta_k)|_{\mathcal{S}_k^r=1}$ is the unique solution to $\mathbb{E}(\Psi_k(X^*, R; \theta_k, \theta_{\mathcal{T}_k}) = 0$, where

$$\Psi_k(X^*, R; \theta_k, \theta_{\mathcal{T}_k}) := \mathbb{I}(\widetilde{\mathcal{S}}_k = 1)\, W_k(\theta_{\mathcal{T}_k})\, f_k(\text{pa}_{\mathcal{G}}(R_k))\, \{R_k - \pi_k(\text{pa}_{\mathcal{G}}(R_k); \theta_k)\}.$$

  - $f_k(\text{pa}_{\mathcal{G}}(R_k))$ is a function with the same dimension as θ_k
  - $W_k(\theta_{\mathcal{T}_k}) = \prod_{R_i \in \mathcal{T}_k} (\mathbb{I}(R_i = 1)/\pi_i(\text{pa}_{\mathcal{G}}(R_i); \theta_i))$ is the inverse propensity weight. By convention, $W_k(\theta_{\mathcal{T}_k}) = 1$ when $\mathcal{T}_k = \emptyset$.
- The estimator $\hat{\theta}_k$ is the solution to $P_n\ \Psi_k(X^*, R; \theta_k, \hat{\theta}_{\mathcal{T}_k}) = 0$
  - $\hat{\theta}_i \in \hat{\theta}_{\mathcal{T}_k}$ index the estimated propensity score for $R_i \in \text{ch}_{\mathbb{T}_k}(R_k)$
    - $\hat{\theta}_i$ is retrieved from $\mathcal{P}$ if $\mathbb{T}_i$ under $\mathbb{T}_k$ is not pruned, thus matches that in $\mathbb{F}$
    - Re-estimate $\hat{\theta}_i$ with Ψ_i that respect the structure of $\mathbb{T}_i$ if it is pruned
    - $\mathcal{E}_k$ and $\mathcal{P}_k$ collect updated estimating function and the re-fitted models
- The asymptotic variance of $\hat{\boldsymbol{\theta}}_k$ is $V_k = A_k^{-1}\ B_k\ (A_k^{-1})'$ and an estimator of it is $\hat{V}_k = \hat{A}_k^{-1}\ \hat{B}_k\ (\hat{A}_k^{-1})'$, where
  - $\boldsymbol{\theta}_k$ is the stacked parameter vector consisting of θ_k and the parameters associated with all indicators appearing in tree $\mathbb{T}_k$, ordered so that θ_k is last.
  - $\boldsymbol{\Psi}_t$ is the corresponding stacked estimating function, collecting Ψ_i for all indicator R_i appear in $\mathbb{T}_k$. Ψ_i is retrieved from $\mathcal{E}_k$ if $\Psi_i \in \mathcal{E}_k$, otherwise retrieve from $\mathcal{E}$.
  - $A_k := \mathbb{E}(\partial \boldsymbol{\Psi}_k / \partial \boldsymbol{\theta}_k)$, $B_k := \mathbb{E}(\boldsymbol{\Psi}_k\ \boldsymbol{\Psi}_k')$, $\hat{A}_k := P_n(\partial \boldsymbol{\Psi}_k / \partial \boldsymbol{\theta}_k|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k})$, and $\hat{B}_k := P_n(\boldsymbol{\Psi}_k\ \boldsymbol{\Psi}_k'|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k})$
- The asymptotic variance of $\hat{\theta}_k$ is the bottom-right block of V_k, with its estimator given by the corresponding entry of $\hat{V}_k$
- Alternatively, π_k can be estimated by fitting R_k on its parents $\text{pa}_{\mathcal{G}}(R_k)$ with weight equals to $\mathbb{I}(\mathcal{S}_k = 1)\, W_k(\hat{\theta}_{\mathcal{T}_k})$, using flexible machine learning methods.
- RETURN $\pi_k(\text{pa}_{\mathcal{G}}(R_k); \hat{\theta}_k), \Psi_k$.

Algorithm E2 IPW Estimation of target law functionals$(\mathcal{G}(X, R, X^*), \mathbb{F}, \mathcal{D}, M(\tilde{X}; \theta))$

1: **Estimation Procedure:**

- $M(\tilde{X}; \theta)$ is a user-specified function of the target law with the same dimension as the target parameter θ, satisfying θ as the unique solution of $\mathbb{E}(M(\tilde{X}; \theta)) = 0$, where $\tilde{X} \subseteq X$ are the variables involved in its construction
- Let $\mathcal{E}, \mathcal{P}$ be the output of Algorithm E1
- Let $\tilde{R} = \{R_i \in R : X_i \in \tilde{X}\}$ collect the missingness indicators associated with $\tilde{X}$
- Let $\mathcal{R}$ denote the set of indicators containing $\tilde{R}$ that is closed under inclusion of the selection sets $\{\mathcal{S}_i\}$. Define $\mathcal{R}$ as follows:
  - Starting from $A_0 = \tilde{R}$
  - Define the recursion $A_{\ell+1} \coloneqq \mathrm{cl}(A_\ell)$, where $\mathrm{cl}(A) \coloneqq A \cup \bigcup_{R_i \in A} \mathcal{S}_i$
  - For ℓ^* such that $A_{\ell^*+1} = A_{\ell^*}$, define $\mathcal{R} \coloneqq A_{\ell^*}$
- **If** $\mathcal{R} \cap \mathcal{D} \neq \emptyset$: *Fail to identify the target parameter*
- **Else:** define estimating function $\Psi(X^*, R; \theta, \theta_{\mathcal{R}})$ as in (37), where $\theta_{\mathcal{R}} \coloneqq (\theta_i)_{R_i \in \mathcal{R}}$:

$$\Psi(X^*, R; \theta, \theta_{\mathcal{R}}) \coloneqq \mathbb{I}(\mathcal{R} = 1) \Big\{ \prod_{R_i \in \mathcal{R}} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1} \Big\}^{-1} M(\tilde{X}; \theta). \tag{37}$$

  An estimator $\hat{\theta}$ is obtained by solving $P_n \Psi(X^*, R; \theta, \hat{\theta}_{\mathcal{R}}) = 0$, where $\hat{\theta}_{\mathcal{R}}$ is retrieved from $\mathcal{P}$
- Define $\boldsymbol{\Psi} \coloneqq \{\boldsymbol{\Psi}_i : R_i \in \mathcal{R}\}$, retrieved from $\mathcal{E}$. Append Ψ as its last element: $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}, \Psi\}$
- Let $\boldsymbol{\theta}$ collect the parameters in one-to-one correspondence with $\boldsymbol{\Psi}$, with its estimator denoted by $\hat{\boldsymbol{\theta}}_k$
- The asymptotic variance of $\hat{\boldsymbol{\theta}}$ is given by $V = A^{-1}B(A^{-1})'$, with an estimator $\hat{V} = \hat{A}^{-1}\ \hat{B}\ (\hat{A}^{-1})'$, where
  - $A \coloneqq \mathbb{E}(\partial \boldsymbol{\Psi}/\partial \boldsymbol{\theta})$, $B \coloneqq \mathbb{E}(\boldsymbol{\Psi}\ \boldsymbol{\Psi}')$, $\hat{A} \coloneqq P_n(\partial \boldsymbol{\Psi}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}})$, and $\hat{B} \coloneqq P_n(\boldsymbol{\Psi}\ \boldsymbol{\Psi}'|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}})$
- The asymptotic variance of $\hat{\theta}$ is the bottom right element of V, with its estimator at the same entry of $\hat{V}$
- RETURN $\hat{\theta}, \hat{V}$

C Proofs

Proof of Theorem 1. Let $\mathcal{T}_k$ be the set of indicators intervened on in $\mathbb{T}_k$, and let $\sigma_k = (s_1, \ldots, s_m)$ be an ordering of $\mathcal{T}_k$ consistent with the reverse topological order used by Algorithm 1. Write $\phi^p_{\sigma_k}\{p\} = \phi^p_{s_m} \circ \cdots \circ \phi^p_{s_1}\{p\}$. We emphasize that $\phi^p_{\sigma_k}\{p\}$ need not be identifiable as a full law. The algorithm is only required to identify the conditional kernel

$$\phi^p_{\sigma_k}\{p\}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k)) \quad \text{evaluated at} \quad \mathcal{S}^r_k = 1.$$

The tree induces a well-defined post-intervention kernel for R_k. Since $\mathcal{T}_k \subseteq R \setminus \{R_k\}$, the intervention sequence encoded by $\mathbb{T}_k$ does not intervene on R_k. By invariance of the missingness mechanism to interventions on indicators other than the target indicator, the propensity score of R_k is unchanged by intervening on $\mathcal{T}_k$. In particular, as kernels,

$$\phi^p_{\sigma_k}\{p\}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k)) = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k)) = \pi_k(\mathrm{pa}_{\mathcal{G}}(R_k)), \tag{38}$$

on the subset of the sample space where both sides are well-defined.

By assumption, the identification criterion (10) holds *for the post-intervention distribution induced by* $\mathbb{T}_k$, meaning that the conditional kernel $\phi^p_{\sigma_k}\{p\}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))$ is identifiable from the observed data law when evaluated at $\mathcal{S}^r_k = 1$. That is,

$$\phi^p_{\sigma_k}\{p\}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1} \tag{39}$$

admits an observed-data functional (possibly involving previously identified propensity scores along $\mathbb{T}_k$) under the model $\mathcal{M}$. Combining the statements yields $\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1} = \phi^p_{\sigma_k}\{p\}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1}$, which is precisely (16). ■

Proof of Corollary 2. Let $\mathcal{T}_k$ be the children of R_k in $\mathbb{T}_k$, and let $p_{\mathbb{T}_k} = \phi^p_{\sigma_k}\{p\}$ be the induced post-intervention law after intervening on all indicators in $\mathcal{T}_k$. Write $\mathcal{T}_k = 1$ for the event that all indicators in $\mathcal{T}_k$ equal one. The same convention applies to $\widetilde{\mathcal{S}}_k$. Repeated application of the fixing formula (15) implies that

$$p_{\mathbb{T}_k}(x, r) \;\propto\; \mathbb{I}(\mathcal{T}_k = 1, \widetilde{\mathcal{S}}_k = 1)\, W_k\, p(x, r), \tag{40}$$

where $W_k = \prod_{R_i \in \mathcal{T}_k} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i))^{-1}$ and each π_i is evaluated at its admissible evaluation set returned by Algorithm 1, which is ensured by $\mathbb{I}(\widetilde{\mathcal{S}}_k = 1)$. Expectations under p_{π_k} can be written as weighted expectations under p. Consequently, for any measurable function f,

$$\mathbb{E}_{p_{\mathbb{T}_k}}[f(X, R)] = \frac{\mathbb{E}\Big[f(X, R)\, \mathbb{I}(\mathcal{T}_k = 1, \widetilde{\mathcal{S}}_k = 1)\, W_k\Big]}{\mathbb{E}\Big[\mathbb{I}(\mathcal{T}_k = 1, \widetilde{\mathcal{S}}_k = 1)\, W_k\Big]}, \tag{41}$$

whenever the expectations exist, and conditioning preserves this weighted representation. Applying Theorem 1, we can write the conditional probability under $p_{\mathbb{T}_k}$ as

$$\begin{aligned}
\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1} &= p_{\mathbb{T}_k}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1} \\
&= p_{\mathbb{T}_k}(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k), \mathcal{T}_k = 1, \widetilde{\mathcal{S}}_k = 1)\Big|_{\mathcal{S}^r_k=1} \\
&= \mathbb{E}_{\mathbb{T}_k}(\mathbb{I}(R_k = 1) \mid \mathrm{pa}_{\mathcal{G}}(R_k), \mathcal{T}_k = 1, \widetilde{\mathcal{S}}_k = 1)\Big|_{\mathcal{S}^r_k=1}.
\end{aligned}$$

Substituting the weighted representation (40) yields

$$\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1} = \frac{\mathbb{E}\Big[\mathbb{I}(R_k = 1)\, W_k \;\Big|\; \mathrm{pa}_{\mathcal{G}}(R_k),\ \mathcal{T}_k = 1,\ \widetilde{\mathcal{S}}_k = 1\Big]}{\mathbb{E}\Big[W_k \;\Big|\; \mathrm{pa}_{\mathcal{G}}(R_k),\ \mathcal{T}_k = 1,\ \widetilde{\mathcal{S}}_k = 1\Big]}\Bigg|_{\mathcal{S}^r_k=1}, \tag{42}$$

which is (17). If $\mathcal{T}_k = \emptyset$, then $W_k \equiv 1$ and the expression reduces to the associational identification formula, namely

$$\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{\mathcal{S}^r_k=1} = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k) \backslash \mathcal{S}^x_k,\, \mathcal{S}^x_k = 1).$$

■

Proof of Theorem 3 (estimating equations for propensity scores). The estimating equation for unknown parameter θ_k indexing π_k is formulated as $P_n\Psi_k(X^*, R; \theta_k, \hat{\theta}_{\mathcal{T}_k})$, with

$$\Psi_k(X^*, R; \theta_k, \theta_{\mathcal{T}_k}) \;:=\; \mathbb{I}(\widetilde{\mathcal{S}}_k = 1)\, W_k(\hat{\theta}_{\mathcal{T}_k})\, f_k(\mathrm{pa}_{\mathcal{G}}(R_k)) \Big\{ R_k - \pi_k(\mathrm{pa}_{\mathcal{G}}(R_k); \theta_k) \Big\}$$

We establish the validity of this estimating equation by showing that $\mathbb{E}(\Psi_k(X^*, R; \theta_k, \theta_{\mathcal{T}_k})) = 0$. Let $\sigma_k = (s_1, \ldots, s_m)$ be an ordering of $\mathcal{T}_k$ consistent with τ.

$$\begin{aligned}
&\mathbb{E}(\Psi_k(X^*, R; \theta_k, \theta_{\mathcal{T}_k})) \\
&= \mathbb{E}\Big(\mathbb{I}(\widetilde{\mathcal{S}}_k = 1) \prod_{R_i \in \mathcal{T}_k} \frac{\mathbb{I}(R_i = 1)}{\pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1}} f_k(\mathrm{pa}_{\mathcal{G}}(R_k)) \left\{ R_k - \pi_k(\mathrm{pa}_{\mathcal{G}}(R_k); \theta_k) \right\} \Big) \\
&= \mathbb{E}\Big(\frac{\mathbb{E}(\mathbb{I}(s_1 = 1) \mid O, R \backslash s_1)}{\pi_{s_1}(\mathrm{pa}_{\mathcal{G}}(s_1); \theta_{s_1})} \mathbb{I}(\widetilde{\mathcal{S}}_k = 1) \prod_{R_i \in \mathcal{T}_k \backslash s_1} \frac{\mathbb{I}(R_i = 1)}{\pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1}} f_k(\mathrm{pa}_{\mathcal{G}}(R_k)) \left\{ R_k - \pi_k(\mathrm{pa}_{\mathcal{G}}(R_k); \theta_k) \right\} \Big) \\
&= \mathbb{E}\Big(\frac{\mathbb{E}(\mathbb{I}(s_1 = 1) \mid \mathrm{pa}_{\mathcal{G}}(s_1, \theta_{s_1}))}{\pi_{s_1}(\mathrm{pa}_{\mathcal{G}}(s_1); \theta_{s_1})} \mathbb{I}(\widetilde{\mathcal{S}}_k = 1) \prod_{R_i \in \mathcal{T}_k \backslash s_1} \frac{\mathbb{I}(R_i = 1)}{\pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1}} f_k(\mathrm{pa}_{\mathcal{G}}(R_k)) \left\{ R_k - \pi_k(\mathrm{pa}_{\mathcal{G}}(R_k); \theta_k) \right\} \Big) \\
&\vdots \\
&= \mathbb{E}(f_k(\mathrm{pa}_{\mathcal{G}}(R_k)) \left\{ R_k - \pi_k(\mathrm{pa}_{\mathcal{G}}(R_k); \theta_k) \right\}) = 0,
\end{aligned}$$

where π_{s_1} and θ_{s_1} denote the propensity score of s_1 and its indexing parameter.

The second equality follows by applying the tower rule, conditioning on (X, R) with s_1 excluded in the inner expectation, and the third equality follows from the local Markov property. The omitted steps apply the same argument to the indicators s_2 through s_m. ■

Proof of Theorem 4 (estimating equations for functionals of the target law).

The estimating equation for the target parameter θ is constructed as $P_n\Psi(X^*, R;\, \theta, \hat{\theta}_{\mathcal{R}}) = 0$,

where

$$\Psi(X^*, R; \theta, \theta_{\mathcal{R}}) \;\coloneqq\; \mathbb{I}(\mathcal{R} = 1) \Big\{ \prod_{R_i \in \mathcal{R}} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1} \Big\}^{-1} M(\tilde{X}; \theta),$$

We establish the validity of this estimating equation by showing that $\mathbb{E}(\Psi(X^*, R;\, \theta, \theta_{\mathcal{R}})) = 0$. Let $\sigma = (s_1, \cdots, s_l)$ be an ordering of $\mathcal{R}$ consistent with τ.

$$\begin{aligned}
&\mathbb{E}(\Psi(X^*, R;\, \theta, \theta_{\mathcal{R}})) \\
&= \mathbb{E}(\mathbb{I}(\mathcal{R} = 1) \Big\{ \prod_{R_i \in \mathcal{R}} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1} \Big\}^{-1} M(\tilde{X}; \theta)) \\
&= \mathbb{E}(\frac{\mathbb{E}(\mathbb{I}(s_1 = 1) \mid X, R \backslash s_1)}{\pi_{s_1}(\mathrm{pa}_{\mathcal{G}}(s_1); \theta_{s_1})|_{\mathcal{S}_{s_1}^r = 1}} \mathbb{I}(\mathcal{R} \backslash s_1 = 1) \Big\{ \prod_{R_i \in \mathcal{R} \backslash s_1} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1} \Big\}^{-1} M(\tilde{X}; \theta)) \\
&= \mathbb{E}(\frac{\mathbb{E}(\mathbb{I}(s_1 = 1) \mid \mathrm{pa}_{\mathcal{G}}(s_1); \theta_{s_1})}{\pi_{s_1}(\mathrm{pa}_{\mathcal{G}}(s_1); \theta_{s_1})|_{\mathcal{S}_{s_1}^r = 1}} \overset{1}{\mathbb{I}}(\mathcal{R} \backslash s_1 = 1) \Big\{ \prod_{R_i \in \mathcal{R} \backslash s_1} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r = 1} \Big\}^{-1} M(\tilde{X}; \theta)) \\
&\vdots \\
&= \mathbb{E}(M(\tilde{X}; \theta)) = 0,
\end{aligned}$$

where $\mathcal{R}_{s_1}^r$ denotes the indicator-induced selection set of s_1. The set $\mathcal{R}$ includes $\mathcal{S}_{s_1}^r$ by construction. Consequently, in the third equality, the numerator of the ratio functional, $\mathbb{E}(\mathbb{I}(s_1 = 1) \mid \mathrm{pa}_{\mathcal{G}}(s_1); \theta_{s_1})$, is evaluated at $\mathcal{S}_{s_1}^r$ and therefore cancels with the denominator. The omitted steps apply the same argument to the indicators s_2 to s_l. ■

Table 8: Key definitions used in the tree-based identification algorithm, illustrated using the mDAG in Fig. E.1(b).

R_k	**Propensity scores**	$\mathcal{S}_k^x$	$\mathcal{R}_k^p$	$\mathcal{C}_{k,k}^{\text{dir}}$	$\text{ch}_{\mathbb{T}_k}(R_k)$	$\tilde{\mathcal{S}}_k$	$\mathcal{S}_k^r$	$\mathcal{S}_k$
R_1	$p(R_1\|R_2, R_3)$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
R_2	$p(R_2\|X_1, X_3)$	$\{R_1, R_3\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$	$\{R_1, R_3\}$	$\emptyset$	$\{R_1, R_3\}$
R_3	$p(R_3\|X_1, X_2)$	$\{R_1, R_2\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$	$\{R_1, R_2\}$	$\emptyset$	$\{R_1, R_2\}$

Table 9: Key definitions used in the tree-based identification algorithm, illustrated using the mDAG in Fig. E.1(e).

R_k	**Propensity scores**	$\mathcal{S}_k^x$	$\mathcal{R}_k^p$	$\mathcal{C}_{k,k}^{\text{dir}}$	$\text{ch}_{\mathbb{T}_k}(R_k)$	$\tilde{\mathcal{S}}_k$	$\mathcal{S}_k^r$	$\mathcal{S}_k$
R_1	$p(R_1\|X_4, R_2)$	$\{R_4\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{R_4\}$	$\emptyset$	$\{R_4\}$
R_2	$p(R_2\|X_5, R_3)$	$\{R_5\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{R_5\}$	$\emptyset$	$\{R_5\}$
R_3	$p(R_3\|R_4, R_5)$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
R_4	$p(R_4\|X_1, R_5)$	$\{R_1\}$	$\{R_1\}$	$\{R_1\}$	$\{R_2, R_3\}$	$\{R_1, R_5\}$	$\{R_5\}$	$\{R_1, R_5\}$
R_4	$(R_4 \to R_2$ pruned)	$\{R_1\}$	$\{R_1\}$	$\{R_1\}$	$\{R_3\}$	$\{R_1, R_5\}$	$\emptyset$	$\{R_1\}$
R_5	$p(R_5\|X_2)$	$\{R_2\}$	$\{R_2\}$	$\{R_2\}$	$\{R_1, R_3, R_4\}$	$\{R_1, R_2, R_4\}$	$\emptyset$	$\{R_2\}$

D Simulation details

D.1 Key identification concepts

The mDAG in Figure E.1(a) represents a MAR model, for which the target law is easily identified. For the remaining three mDAGs, we summarize the key definitions underlying the identification procedure in Appendix Tables 8, 9, and 10.

D.2 Data generating processes for simulation study

D.2.1 Task 1: mean estimation

($\mathcal{G}_1$ shown in Appendix Figure E.1(a))

$$X_1 \sim \mathcal{N}(0,1),\ X_2 \sim \mathcal{N}(1 - X_1, 1),\ X_3 \sim \mathcal{N}(1 - 2X_2 + 3X_1, 1),$$

$$R_2 \sim \text{Binomial}(\text{expit}(2 + X_1)),\ R_3 \sim \text{Binomial}(\text{expit}(1 + 0.5X_1)).$$

Table 10: Key definitions used in the tree-based identification algorithm, illustrated using the mDAG in Fig. E.1(g).

R_k	**Propensity scores**	$\mathcal{S}_k^x$	$\mathcal{R}_k^p$	$\mathcal{C}_{k,k}^{\text{dir}}$	$\text{ch}_{\mathbb{T}_k}(R_k)$	$\tilde{\mathcal{S}}_k$	$\mathcal{S}_k^r$	$\mathcal{S}_k$
R_1	$p(R_1\|R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10})$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
R_2	$p(R_2\|X_1, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10})$	$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$	$\{R_1\}$	$\emptyset$	$\{R_1\}$
R_3	$p(R_3\|X_1, X_2, R_4, R_5, R_6, R_7, R_8, R_9, R_{10})$	$\{R_1, R_2\}$	$\{R_1, R_2\}$	$\emptyset$	$\{R_1, R_2\}$	$\{R_1, R_2\}$	$\emptyset$	$\{R_1, R_2\}$
R_4	$p(R_4\|X_2, X_3, R_5, R_6, R_7, R_8, R_9, R_{10})$	$\{R_2, R_3\}$	$\{R_2, R_3\}$	$\emptyset$	$\{R_1, R_2, R_3\}$	$\{R_1, R_2, R_3\}$	$\emptyset$	$\{R_2, R_3\}$
$\vdots$								
R_{10}	$p(R_{10}\|X_8, X_9)$	$\{R_8, R_9\}$	$\{R_8, R_9\}$	$\emptyset$	$\{R_1, \cdots, R_9\}$	$\{R_1, \cdots, R_9\}$	$\emptyset$	$\{R_8, R_9\}$

($\mathcal{G}_2$ shown in Appendix Figure E.1(c))

$$X_1 \sim \mathcal{N}(1, 1),\ \ X_2 \sim \mathcal{N}(3 - 0.6|X_1|, 1),$$

$$X_3 \sim \mathcal{N}(2 - X_2^2 + 4X_2 + 2X_1X_2, 1.5),$$

$$R_1 \sim \text{Binomial}(\text{expit}(1 + R_2 + R_3)),$$

$$R_2 \sim \text{Binomial}(\text{expit}(-0.5X_1 + 0.15X_3)),$$

$$R_3 \sim \text{Binomial}(\text{expit}(3 + 0.5X_1 - X_2)).$$

($\mathcal{G}_3$ shown in Appendix Figure E.1(e))

$$X_1 \sim \mathcal{N}(1, 1),$$

$$X_2 \sim \mathcal{N}(3 - 0.6|X_1|, 1),$$

$$X_3 \sim \mathcal{N}(2 - X_2^2 + 4X_2 + 2X_1X_2, 1.5),$$

$$X_5 \sim \mathcal{N}(2, 1),$$

$$X_4 \sim \mathcal{N}(5X_5^3 - 5|X_3|X_5, 1),$$

$$R_1 \sim \text{Binomial}(\text{expit}(1.2 + 0.01X_4 + 1.5R_2)),$$

$$R_2 \sim \text{Binomial}(\text{expit}(-4 + X_5 + R_3)),$$

$$R_3 \sim \text{Binomial}(\text{expit}(-0.8 + 2R_4 + 1.8R_5)),$$

$$R_4 \sim \text{Binomial}(\text{expit}(0.3 + 1.5X_1 + 2R_5)),$$

$$R_5 \sim \text{Binomial}(\text{expit}(0.8 + 1.5X_2)).$$

($\mathcal{G}_4$ shown in Appendix Figure E.1(g))

$X_1 \sim \mathcal{N}(1, 1),\ X_2 \sim \mathcal{N}(3 - 0.6|X_1|, 1),\ X_3 \sim \mathcal{N}(2 - X_2^2 + 4X_2 + 2X_1X_2, 1.5),$

$X_4 \sim \mathcal{N}(1 + X_3 - 0.5X_2, 1),\ X_5 \sim \mathcal{N}(1 + 0.9X_4 - 0.4X_3, 1),\ X_6 \sim \mathcal{N}(1 + 0.8X_5 - 0.3X_4, 1),$

$X_7 \sim \mathcal{N}(1 + 0.7X_6 - 0.2X_5, 1),\ X_8 \sim \mathcal{N}(1 + 0.7X_7 - 0.2X_6, 1),$

$X_9 \sim \mathcal{N}(1 + 0.7X_8 - 0.2X_7, 1),\ X_{10} \sim \mathcal{N}(1 + 0.7X_9 - 0.2X_8, 1),$

$R_1 \sim \text{Binomial}(\text{expit}(-0.1 + 0.1R_2 - 0.1R_3 + 0.1R_4 - 0.1R_5 + 0.1R_6 - 0.1R_7 + 0.1R_8 - 0.1R_9 + 0.1R_{10})),$

$R_2 \sim \text{Binomial}(\text{expit}(0.1 + 0.3X_1 - 0.1R_3 + 0.1R_4 - 0.1R_5 + 0.1R_6 - 0.1R_7 + 0.1R_8 - 0.1R_9 + 0.1R_{10})),$

$R_3 \sim \text{Binomial}(\text{expit}(0.1 - 0.3X_2 + 0.3X_1 - 0.1R_4 + 0.1R_5 - 0.1R_6 + 0.1R_7 - 0.1R_8 + 0.1R_9 - 0.1R_{10})),$

$R_4 \sim \text{Binomial}(\text{expit}(10 - X_3 + 0.2X_2 - 0.1R_5 + 0.1R_6 - 0.1R_7 + 0.1R_8 - 0.1R_9 + 0.1R_{10})),$

$R_5 \sim \text{Binomial}(\text{expit}(10 - X_4 + 0.2X_3 - 0.1R_6 + 0.1R_7 - 0.1R_8 + 0.1R_9 - 0.1R_{10})),$

$R_6 \sim \text{Binomial}(\text{expit}(2 - X_5 + X_4 - 0.1R_7 + 0.1R_8 - 0.1R_9 + 0.1R_{10})),$

$R_7 \sim \text{Binomial}(\text{expit}(2 - X_6 + X_5 - 0.1R_8 + 0.1R_9 - 0.1R_{10})),$

$R_8 \sim \text{Binomial}(\text{expit}(2 - 2X_7 + 2X_6 - 0.1R_9 + 0.1R_{10})),$

$R_9 \sim \text{Binomial}(\text{expit}(2 - 2X_8 + 2X_7 - 0.1R_{10})),$

$R_{10} \sim \text{Binomial}(\text{expit}(2 - X_9 + X_8)).$

D.2.2 Task 2: parametric regression

For the current task, the DGP from Subsection D.2.1 is modified by setting the coefficients on X_1 to zero in the conditional distribution of X_3 given X_1 and X_2. Specifically, we have the following.

$$X_3 \sim \mathcal{N}(1 - 2X_2 + 0X_1, 1), \quad (\mathcal{G}_1 \text{ shown in Appendix Figure E.1(a)})$$

$$X_3 \sim \mathcal{N}(2 - X_2^2 + 0X_1X_2, 1.5), \quad (\mathcal{G}_2 - \mathcal{G}_4 \text{ shown in Appendix Figure E.1(c,e,g)})$$

D.2.3 Task 3: causal effect estimation

For the current task, the DGP from Subsection D.2.1 is modified by generating X_2 as a binary variable, with details specified below.

$$X_2 \sim \text{Binomial}(\text{expit}(1 - X_1)), \quad (\mathcal{G}_1 \text{ shown in Appendix Figure E.1(a)})$$

$$X_2 \sim \text{Binomial}(\text{expit}(3 - 0.6|X_1|)), \quad (\mathcal{G}_2 - \mathcal{G}_4 \text{ shown in Appendix Figure E.1(c,e,g)}).$$

D.3 Details of the estimating procedure

D.3.1 Task 1: mean estimation.

Estimation under the proposed method begins with estimating the propensity scores, with the fitted propensity score models collected in $\mathcal{P}$. The mean of X_3 is then computed as the empirical inverse-propensity-weighted mean of X_3^*, as shown in Equation (43), with θ_i retrieved from $\mathcal{P}$:

$$P_n\Big(\frac{\mathbb{I}(\mathcal{R} = 1)}{\prod_{R_i \in \mathcal{R}} \pi_i(\text{pa}_{\mathcal{G}}(R_i); \hat{\theta}_i)|_{\mathcal{S}_i^r = 1}} X_3\Big). \tag{43}$$

For $\mathcal{G}_1$ through $\mathcal{G}_4$, the corresponding sets $\mathcal{R}$ are $\{R_3\}$, $\{R_1, R_2, R_3\}$, $\{R_3\}$, and $\{R_1, R_2, R_3\}$, respectively.

D.3.2 Task 2: parametric regression.

For each mDAG, regression coefficients are estimated under a correctly specified model. For $\mathcal{G}_1$, we consider $\mathbb{E}(X_3 \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. For $\mathcal{G}_2$ through $\mathcal{G}_4$, we consider $\mathbb{E}(X_3 \mid X_1, X_2) = \beta_0 + \beta_1 X_1 X_2 + \beta_2 X_2 + \beta_3 X_2^2$. The regression coefficients can be estimated either via estimating equations or via weighted regressions, with inverse propensity weights defined by the set $\mathcal{R}$. To be consistent with our implementation with the other three missing-data methods, where we performed regressions using complete or imputed data, we adopt the latter approach for the propsed method. The weighted regressions are fitted using observations with $\mathcal{R} = 1$ and weight $\{\prod_{R_i \in \mathcal{R}} \pi_i(\text{pa}_{\mathcal{G}}(R_i); \theta_i) \big|_{\mathcal{S}_i^r = 1}\}^{-1}$. For $\mathcal{G}_1$ through $\mathcal{G}_4$, the corresponding sets $\mathcal{R}$ are $\{R_2, R_3\}$, $\{R_1, R_2, R_3\}$, $\{R_1, R_2, R_3, R_4, R_5\}$, and $\{R_1, R_2, R_3\}$, respectively.

On the unbiasedness of complete-case analysis for mDAGs $\mathcal{G}_3$ and $\mathcal{G}_4$. For $\mathcal{G}_3$, the d-separation criterion implies that $X_3 \perp X_1 \mid X_2, R_1, R_2, R_3$. As a result, complete-case analysis yields unbiased estimation of the regression coefficients.

For $\mathcal{G}_4$, valid regression analysis is conducted under the post-intervention distribution $p_{\mathbb{T}_4}$, obtained by interventions $R_1 = R_2 = R_3 = 1$. Since the regression involves only X_1 through X_3, we further marginalize $p_{\mathbb{T}_4}$ over the irrelevant variables X_4 through X_{10} and R_4 through R_{10}. We now show step by step how $p_{\mathbb{T}_4}$ is derived. First, intervening on R_1 yields $p_{\mathbb{T}_2}$.

$$p_{\mathbb{T}_2}(X, R) \propto \mathbb{I}(R_1 = 1)\ \pi_1^{-1} p(X_1^*, X_2, \cdots, X_{10}, R).$$

Under $p_{\mathbb{T}_2}$, π_2 is identified and can therefore be divided out from $p_{\mathbb{T}_2}$ to obtain $p_{\mathbb{T}_3}$:

$$p_{\mathbb{T}_3}(X, R) \propto \mathbb{I}(R_1 = 1, R_2 = 1)\ \pi_1^{-1}\ \pi_2^{-1}\ p(X_1^*, X_2^*, \cdots, X_{10}, R).$$

A similar argument applies to R_3, yielding $p_{\mathbb{T}_4}$, under which valid regression analysis is performed:

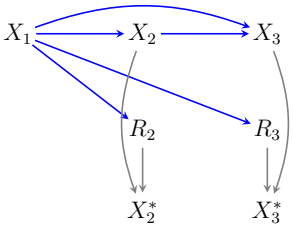
$$p_{\mathbb{T}_4}(X, R) \propto \mathbb{I}(R_1 = 1, R_2 = 1, R_3 = 1)\ \pi_1^{-1}\ \pi_2^{-1}\ \pi_3^{-1}\ p(X_1^*, X_2^*, X_3^* \cdots, X_{10}, R).$$

The distribution $p_{\mathbb{T}_4}$, marginalized over irrelevant variables, is proportional to the complete-data distribution $\mathbb{I}(R_1 = 1, R_2 = 1, R_3 = 1)\, p(X_1^*, X_2^*, X_3^*)$. As a result, for this mDAG, complete-case analysis yields unbiased estimates of the regression coefficients.

The same argument fails for $\mathcal{G}_2$ because π_3 is not identified in the post-intervention distribution obtained by intervening on R_1 and R_2.

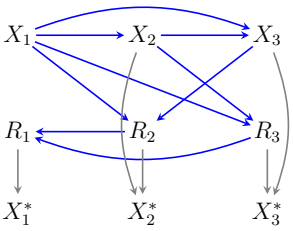
D.3.3 Task 3: causal effect estimation.

With $\mathcal{G}_1$, we illustrate the estimation of $\mathbb{E}(X_3^{x_2})$ for $x_2 \in \{0, 1\}$. The same procedure applies to the other mDAGs. We first fit a weighted regression as described in Subsection D.3.2. We then form predictions $\hat{\mathbb{E}}(X_3 \mid X_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 x_2$ on observations with $\mathcal{R} = 1$. The target parameter $\mathbb{E}(X_3^{x_2})$ is estimated as the empirical inverse-propensity-weighted mean of $\hat{\mathbb{E}}(X_3 \mid X_1, x_2)$, using the same weights as those used to fit the regression.

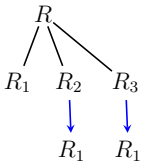


(a) $\mathcal{G}_1$

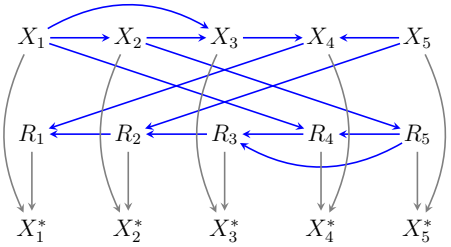
(b) $\mathcal{F}_1$ corresponding to $\mathcal{G}_1$



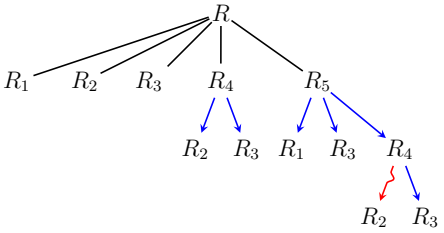
(c) $\mathcal{G}_2$



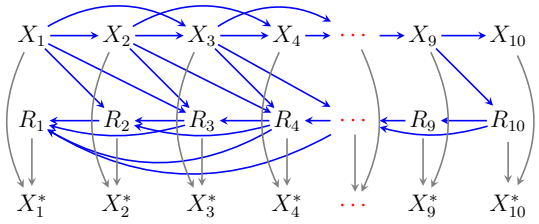
(d) $\mathcal{F}_2$ corresponding to $\mathcal{G}_2$



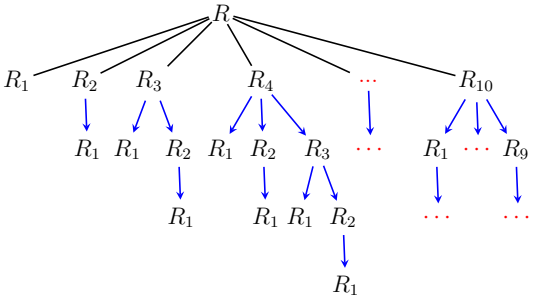
(e) $\mathcal{G}_3$



(f) $\mathcal{F}_3$ corresponding to $\mathcal{G}_3$

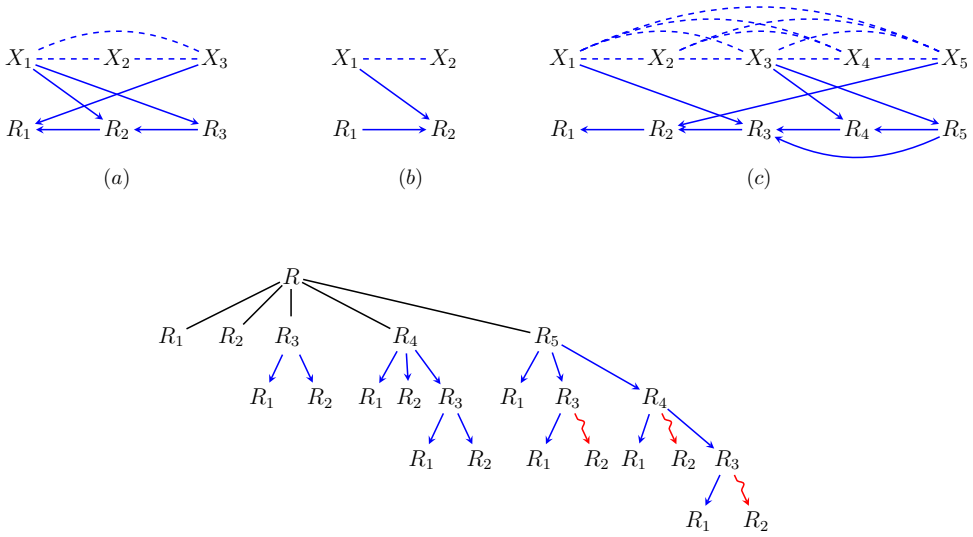


(g) $\mathcal{G}_4$



(h) $\mathcal{F}_4$ corresponding to $\mathcal{G}_4$

Figure E.1: (a), (c), (d), and (e) show the mDAGs used in the simulation study in Section 6. (a) corresponds to a MAR model, while (c), (e), and (g) correspond to MNAR models; the mDAG in (g) is a submodel of the permutation model in (Robins 1997). To save space, some variables and edges are omitted in (e). Specifically, each X_i receives edges from X_{i-1} and X_{i-2}, and each R_i receives edges from X_{i-1}, X_{i-2}, and R_j for all $j > i$. (b), (d), (f), and (h) are the corresponding visualizations of the constructed trees. In (h), the subtree corresponding to all R_j with $j < i$ is attached under R_i, and no pruning is performed.



(d) Constructed forest corresponding to the mDAG in (c)

Figure E.2: Additional figures referenced in the main manuscript.