Identification In Missing Data Models Represented By Directed Acyclic Graphs

Rohit Bhattacharya[†*], **Razieh Nabi**[†*], **Ilya Shpitser**[†], **James M. Robins**[‡]
[†] Department of Computer Science, Johns Hopkins University, Baltimore, MD
[‡] Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA
[*] Equal contribution
(*rbhattacharya*@, *rnabi*@, *ilyas*@cs.)jhu.edu, *robins*@hsph.harvard.edu

Abstract

Missing data is a pervasive problem in data analyses, resulting in datasets that contain censored realizations of a target distribution. Many approaches to inference on the target distribution using censored observed data, rely on missing data models represented as a factorization with respect to a directed acyclic graph. In this paper we consider the identifiability of the target distribution within this class of models, and show that the most general identification strategies proposed so far retain a significant gap in that they fail to identify a wide class of identifiable distributions. To address this gap, we propose a new algorithm that significantly generalizes the types of manipulations used in the ID algorithm [14, 16], developed in the context of causal inference, in order to obtain identification.

1 INTRODUCTION

Missing data is ubiquitous in applied data analyses resulting in target distributions that are systematically censored by a missingness process. A common modeling approach assumes data entries are censored in a way that does not depend on the underlying missing data, known as the missing completely at random (MCAR) model, or only depends on observed values in the data, known as the missing at random (MAR) model. These simple models are insufficient however, in problems where missingness status may depend on underlying values that are themselves censored. This type of missingness is known as missing not at random (MNAR) [9, 10, 17].

While the underlying target distribution is often not identified from observed data under MNAR, there exist identified MNAR models. These include the permutation model [9], the discrete choice model [15], the no self-censoring model [11, 12], the block-sequential MAR model [18], and others. Restrictions defining many, but not all, of these models may be represented by a factorization of the full data law (consisting of both the target distribution and the missingness process) with respect to a directed acyclic graph (DAG).

The problem of identification of the target distribution from the observed distribution in missing data DAG models bears many similarities to the problem of identification of interventional distributions from the observed distribution in causal DAG models with hidden variables. This observation prompted recent work [3, 4, 13] on adapting identification methods from causal inference to identifying target distributions in missing data models.

In this paper we show that the most general currently known methods for identification in missing data DAG models retain a significant gap, in the sense that they fail to identify the target distribution in many models where it is identified. We show that methods used to obtain a complete characterization of identification of interventional distributions, via the ID algorithm [14, 16], or their simple generalizations [3, 4, 13], are insufficient on their own for obtaining a similar characterization for missing data problems. We describe, via a set of examples, that in order to be complete, an identification algorithm for missing data must recursively simplify the problem by removing *sets* of variables, rather than single variables, and these must be removed according to a *partial order*, rather than a total order. Furthermore, the algorithm must be able to handle subproblems where selection bias or hidden variables, or both, are present even if these complications are missing in the original problem. We develop a new general algorithm that exploits these observations and significantly narrows the identifiability gap in existing methods. Finally, we show that in certain classes of missing data DAG models, our algorithm takes on a particularly simple formulation to identify the target distribution.

Our paper is organized as follows. In section 2, we introduce the necessary preliminaries from the graphical causal inference literature. In section 3 we introduce missing data models represented by DAGs. In section 4, we illustrate, via examples, that existing identification strategies based on simple generalizations of causal inference methods are not sufficient for identification in general, and describe generalizations needed for identification in these examples. In section 5, we give a general identification algorithm which incorporates techniques needed to obtain identification in the examples we describe. Section 6 contains our conclusions. We defer longer proofs to the supplement in the interests of space.

2 PRELIMINARIES

Many techniques useful for identification in missing data contexts were first derived in causal inference. Causal inference is concerned with expressing counterfactual distributions, obtained after the intervention operation, from the observed data distribution, using constraints embedded in a causal model, often represented by a DAG.

A DAG is a graph $\mathcal{G}$ with a vertex set $\mathbf{V}$ connected by directed edges such that there are no directed cycles in the graph. A statistical model of a DAG $\mathcal{G}$ is the set of distributions $p(\mathbf{V})$ such that $p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V|\,\mathrm{pa}_{\mathcal{G}}(V))$, where $\mathrm{pa}_{\mathcal{G}}(V)$ are the set of parents of V in $\mathcal{G}$. Causal models of a DAG are also sets of distributions, but on counterfactual random variables. Given $Y \in \mathbf{V}$ and $\mathbf{A} \subseteq \mathbf{V} \setminus \{Y\}$, a counterfactual variable, or potential outcome, written as $Y(\mathbf{a})$, represents the value of Y in a hypothetical situation where $\mathbf{A}$ were set to values $\mathbf{a}$ by an *intervention operation* [6]. Given a set $\mathbf{Y}$, define $\mathbf{Y}(\mathbf{a}) \equiv \{\mathbf{Y}\}(\mathbf{a}) \equiv \{Y(\mathbf{a}) \mid Y \in \mathbf{Y}\}$. The distribution $p(\mathbf{Y}(\mathbf{a}))$ is sometimes written as $p(\mathbf{Y}|\mathrm{do}(\mathbf{a}))$ [6].

A causal parameter is said to be *identified* in a causal model if it is a function of the observed data distribution $p(\mathbf{V})$. Otherwise the parameter is said to be *non-identified*. In all causal models of a DAG $\mathcal{G}$ that are typically used, all interventional distributions $p(\{\mathbf{V} \setminus \mathbf{A}\}(\mathbf{a}))$ are identified by the *g-formula* [8]:

$$p(\{\mathbf{V} \setminus \mathbf{A}\}(\mathbf{a})) = \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V|\,\mathrm{pa}_{\mathcal{G}}(V))\big|_{\mathbf{A}=\mathbf{a}}. \quad (1)$$

If a causal model contains hidden variables, only data on the observed marginal distribution is available. In this case, not every interventional distribution is identified, and identification theory becomes more complex. A general algorithm for identification of causal effects in this setting was given in [16], and proven complete in [14, 1]. Here, we describe a simple reformulation of this algorithm as a truncated nested factorization analogous to the g-formula, phrased in terms of kernels and mixed graphs recursively defined via a fixing operator [7]. As we will see, many of the techniques developed for identification in the presence of hidden variables will need to be employed (and generalized) for missing data, even if no variables are completely hidden.

We describe acyclic directed mixed graphs (ADMGs) obtained from a hidden variable DAG by a latent projection operation in section 2.1, and a nested factorization associated with these ADMGs in section 2.2. This factorization is formulated in terms of conditional ADMGs and kernels (described in section 2.2.1), via the fixing operator (described in section 2.2.2). The truncated nested factorization that yields all identifiable functions for interventional distributions is described in section 2.3.

As a prelude to the rest of the paper, we introduce the following notation for some standard genealogic sets of a graph $\mathcal{G}$ with a set of vertices $\mathbf{V}$: parents $\mathrm{pa}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | U \to V\}$, children $\mathrm{ch}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | V \to U\}$, descendants $\mathrm{de}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | V \to \cdots \to U\}$, ancestors $\mathrm{an}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | U \to \cdots \to V\}$, and non-descendants $\mathrm{nd}_{\mathcal{G}}(V) \equiv \mathbf{V} \setminus \mathrm{de}_{\mathcal{G}}(V)$. A district $\mathbf{D}$ is defined as the maximal set of vertices that are pairwise connected by a bidirected path (a path containing only $\leftrightarrow$ edges). We denote the district of V as $\mathrm{dis}_{\mathcal{G}}(V)$, and the set of all districts in $\mathcal{G}$ as $\mathcal{D}(\mathcal{G})$. By convention, for any V, $\mathrm{dis}_{\mathcal{G}}(V) \cap \mathrm{de}(V) \cap \mathrm{an}_{\mathcal{G}}(V) = \{V\}$. Finally, the Markov blanket $\mathrm{mb}_{\mathcal{G}}(V) \equiv \mathrm{dis}_{\mathcal{G}}(V) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathcal{G}}(V))$ is defined as the set that gives rise to the following independence relation through m-separation: $V \perp\!\!\!\perp \mathrm{nd}_{\mathcal{G}}(V) \setminus \mathrm{mb}_{\mathcal{G}}(V) | \,\mathrm{mb}_{\mathcal{G}}(V)$ [7]. The above definitions apply disjunctively to sets of variables $\mathbf{S} \subset \mathbf{V}$; e.g. $\mathrm{pa}_{\mathcal{G}}(\mathbf{S}) = \cup_{S \in \mathbf{S}} \mathrm{pa}_{\mathcal{G}}(S)$.

2.1 LATENT PROJECTION ADMGS

Given a DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, where $\mathbf{V}$ are observed and $\mathbf{H}$ are hidden variables, a latent projection $\mathcal{G}(\mathbf{V})$ is the following ADMG with a vertex set $\mathbf{V}$. An edge $A \to B$ exists in $\mathcal{G}(\mathbf{V})$ if there exists a directed path from A to B in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ with all intermediate vertices in $\mathbf{H}$. Similarly, an edge $A \leftrightarrow B$ exists in $\mathcal{G}(\mathbf{V})$ if there exists a path without consecutive edges $\to \circ \leftarrow$ from A to B with the first edge on the path of the form $A \leftarrow$ and the last edge on the path of the form $\to B$, and all intermediate vertices on the path in $\mathbf{H}$. Latent projections define an infinite class of hidden variable DAGs that share identification theory. Thus, identification algorithms are typically defined on latent projections for simplicity.

2.2 NESTED FACTORIZATION

The nested factorization of $p(\mathbf{V})$ with respect to an ADMG $\mathcal{G}(\mathbf{V})$ is defined on *kernel* objects derived from $p(\mathbf{V})$ and *conditional ADMGs* derived from $\mathcal{G}(\mathbf{V})$. The

derivations are via a fixing operation, which can be causally interpreted as a single application of the g-formula on a single variable (to either a graph or a kernel) to obtain another graph or another kernel.

2.2.1 Conditional Graphs And Kernels

A conditional acyclic directed mixed graph (CADMG) $\mathcal{G}(\mathbf{V}, \mathbf{W})$ is an ADMG in which the nodes are partitioned into $\mathbf{W}$, representing *fixed variables*, and $\mathbf{V}$, representing *random variables*. Only outgoing directed edges may be adjacent to variables in $\mathbf{W}$.

A *kernel* $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ is a mapping from values in $\mathbf{W}$ to normalized densities over $\mathbf{V}$ [2]. In other words, kernels act like conditional distributions in the sense that $\sum_{\mathbf{v}\in\mathbf{V}} q_{\mathbf{V}}(\mathbf{v}|\mathbf{w}) = 1, \forall \mathbf{w} \in \mathbf{W}$. Conditioning and marginalization in kernels are defined in the usual way. For $\mathbf{A} \subseteq \mathbf{V}$, we define $q(\mathbf{A}|\mathbf{W}) \equiv \sum_{\mathbf{V}\setminus\mathbf{A}} q(\mathbf{V}|\mathbf{W})$ and $q(\mathbf{V} \setminus \mathbf{A}|\mathbf{A}, \mathbf{W}) \equiv q(\mathbf{V}|\mathbf{W})/q(\mathbf{A}|\mathbf{W})$.

2.2.2 Fixability And Fixing

A variable $V \in \mathbf{V}$ in a CADMG $\mathcal{G}$ is *fixable* if $\mathrm{de}_{\mathcal{G}}(V) \cap \mathrm{dis}_{\mathcal{G}}(V) = \{V\}$. In other words, V is fixable if paths $V \leftrightarrow \cdots \leftrightarrow U$ and $V \rightarrow \cdots \rightarrow U$ do not *both* exist in $\mathcal{G}$ for any $U \in \mathbf{V} \setminus \{V\}$. Given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$ and $V \in \mathbf{V}$ fixable in $\mathcal{G}$, the fixing operator $\phi_V(\mathcal{G})$ yields a new CADMG $\mathcal{G}'(\mathbf{V} \setminus \{V\}, \mathbf{W} \cup \{V\})$, where all edges with arrowheads into V are removed, and all other edges in $\mathcal{G}$ are kept. Similarly, given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, a kernel $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$, and $V \in \mathbf{V}$ fixable in $\mathcal{G}$, the fixing operator $\phi_V(q_{\mathbf{V}}; \mathcal{G})$ yields a new kernel $q'_{\mathbf{V}\setminus\{V\}}(\mathbf{V}\setminus\{V\}|\mathbf{W}\cup\{V\}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(V|\,\mathrm{nd}_{\mathcal{G}}(V),\mathbf{W})}$. Fixing is a probabilistic operation in which we divide a kernel by a conditional kernel. In some cases this operates as a conditioning operation, in other cases as a marginalization operation, and in yet other cases, as neither, depending on the structure of the kernel being divided.

For a set $\mathbf{S} \subseteq \mathbf{V}$ in a CADMG $\mathcal{G}$, if all vertices in $\mathbf{S}$ can be ordered into a sequence $\sigma_{\mathbf{S}} = \langle S_1, S_2, \ldots \rangle$ such that S_1 is fixable in $\mathcal{G}$, S_2 in $\phi_{S_1}(\mathcal{G})$, etc., $\mathbf{S}$ is said to be *fixable* in $\mathcal{G}$, $\mathbf{V} \setminus \mathbf{S}$ is said to be *reachable* in $\mathcal{G}$, and $\sigma_{\mathbf{S}}$ is said to be valid. A reachable set $\mathbf{C}$ is said to be *intrinsic* if $\mathcal{G}_{\mathbf{C}}$ has a single district, where $\mathcal{G}_{\mathbf{C}}$ is the induced subgraph where we keep all vertices in $\mathbf{C}$ and edges whose endpoints are in $\mathbf{C}$. We will define $\phi_{\sigma_{\mathbf{S}}}(\mathcal{G})$ and $\phi_{\sigma_{\mathbf{S}}}(q_{\mathbf{V}}; \mathcal{G})$ via the usual function composition to yield operators that fix all elements in $\mathbf{S}$ in the order given by $\sigma_{\mathbf{S}}$.

The distribution $p(\mathbf{V})$ is said to obey the nested factorization for an ADMG $\mathcal{G}$ if there exists a set of kernels $\left\{q_{\mathbf{C}}\left(\mathbf{C} \mid \mathrm{pa}_{\mathcal{G}}(\mathbf{C})\right) \mid \mathbf{C} \text{ is intrinsic in } \mathcal{G}\right\}$ such that for every fixable $\mathbf{S}$, and any valid $\sigma_{\mathbf{S}}$, $\phi_{\sigma_{\mathbf{S}}}(p(\mathbf{V}); \mathcal{G}) = \prod_{\mathbf{D}\in\mathcal{D}(\phi_{\sigma_{\mathbf{S}}}(\mathcal{G}))} q_{\mathbf{D}}(\mathbf{D}|\,\mathrm{pa}_{\mathcal{G}_{\mathbf{S}}}(\mathbf{D}))$. All valid fixing sequences for $\mathbf{S}$ yield the same CADMG $\mathcal{G}(\mathbf{V} \setminus \mathbf{S}, \mathbf{S})$, and if $p(\mathbf{V})$ obeys the nested factorization for $\mathcal{G}$, all valid fixing sequences for $\mathbf{S}$ yield the same kernel. As a result, for any valid sequence σ for $\mathbf{S}$, we will redefine the operator ϕ_σ, for both graphs and kernels, to be $\phi_{\mathbf{S}}$. In addition, it can be shown that the above kernel set is characterized as: $\left\{q_{\mathbf{C}}\left(\mathbf{C} \mid \mathrm{pa}_{\mathcal{G}}(\mathbf{C})\right) \mid \mathbf{C} \text{ is intrinsic in } \mathcal{G}\right\} = \left\{\phi_{\mathbf{V}\setminus\mathbf{C}}(p(\mathbf{V}); \mathcal{G}) \mid \mathbf{C} \text{ is intrinsic in } \mathcal{G}\right\}$ [7]. Thus, we can re-express the above nested factorization as stating that for any fixable set $\mathbf{S}$, we have $\phi_{\mathbf{S}}(p(\mathbf{V}); \mathcal{G}) = \prod_{\mathbf{D}\in\mathcal{D}(\phi_{\mathbf{S}}(\mathcal{G}))} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G})$.

An important result in [7] states that if $p(\mathbf{V} \cup \mathbf{H})$ obeys the factorization for a DAG $\mathcal{G}$ with vertex set $\mathbf{V} \cup \mathbf{H}$, then $p(\mathbf{V})$ obeys the nested factorization for the latent projection ADMG $\mathcal{G}(\mathbf{V})$.

2.3 IDENTIFICATION AS A TRUNCATED NESTED FACTORIZATION

For any disjoint subsets $\mathbf{Y}, \mathbf{A}$ of $\mathbf{V}$ in a latent projection $\mathcal{G}(\mathbf{V})$ representing a causal DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, define $\mathbf{Y}^* \equiv \mathrm{an}_{\mathcal{G}(\mathbf{V})_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y})$. Then $p(\mathbf{Y}(\mathbf{a}))$ is identified from $p(\mathbf{V})$ in $\mathcal{G}$ *if and only if* every set $\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*})$ is intrinsic. If identification holds, we have:

$$p(\mathbf{Y}(\mathbf{a})) = \sum_{\mathbf{Y}^*\setminus\mathbf{Y}} \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\mathbf{A}=\mathbf{a}}.$$

In other words, $p(\mathbf{Y}(\mathbf{a}))$ is identified if and only if it can be expressed as a factorization, where every piece corresponds to a kernel associated with a set intrinsic in $\mathcal{G}(\mathbf{V})$. Moreover, no term in this factorization contains elements of $\mathbf{A}$ as random variables, just as was the case in (1). The above provides a concise formulation of the ID algorithm [16, 14] in terms of the nested Markov model which contains the causal model of the observed distribution.

If $\mathbf{Y} = \{Y\}$, and $\mathbf{A} = \{\mathrm{pa}_{\mathcal{G}}(Y)\}$, then the above truncated factorization has a simpler form:

$$p(Y(\mathbf{a})) = \phi_{\mathbf{V}\setminus\{Y\}}(p(\mathbf{V}); \mathcal{G})|_{\mathbf{A}=\mathbf{a}}.$$

In words, to identify the interventional distribution of Y where all parents (direct causes) $\mathbf{A}$ of Y are set to values $\mathbf{a}$, we must find a total ordering on variables other than Y ($\mathbf{V} \setminus \{Y\}$) that forms a valid fixing sequence. If such an ordering exists, the identifying functional is found from $p(\mathbf{V})$ by applying the fixing operator to each variable in succession, in accordance with this ordering. Fig. 1 shows the identification of the functional $p(Y(a))$ following a total ordering of fixing M, B, A.

Before generalizing these tools to the identification of missing data models, we first introduce the representation of these models using DAGs.

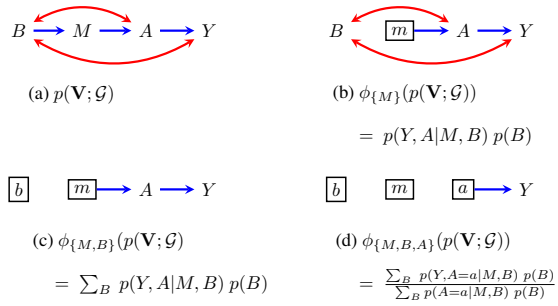


Figure 1: Identification of $p(Y(a))$ by following a total order of valid fixing operations.

3 MISSING DATA MODELS OF A DAG

Missing data models are sets of full data laws (distributions) $p(\mathbf{X^{(1)}}, \mathbf{O}, \mathbf{R})$ composed of the target laws $p(\mathbf{X^{(1)}}, \mathbf{O})$, and the nuisance laws $p(\mathbf{R}|\mathbf{X^{(1)}}, \mathbf{O})$ defining the missingness processes. The target law is over a set $\mathbf{X^{(1)}} \equiv \{X_1^{(1)}, \ldots, X_k^{(1)}\}$ of random variables that are potentially missing, and a set $\mathbf{O} \equiv \{O_1, \ldots, O_m\}$ of random variables that are always observed. The nuisance law defines the behavior of missingness indicators $\mathbf{R} \equiv \{R_1, \ldots, R_k\}$ given values of missing and observed variables. Each missing variable $X_i^{(1)} \in \mathbf{X^{(1)}}$ has a corresponding observed proxy variable X_i, defined as $X_i \equiv X_i^{(1)}$ if $R_i = 1$, and defined as $X_i \equiv$ "?" if $R_i = 0$ (this is the missing data analogue of the consistency property in causal inference). As a result, the observed data law in missing data problems is $p(\mathbf{R}, \mathbf{O}, \mathbf{X})$, while some function of the target law $p(\mathbf{X^{(1)}}, \mathbf{O})$, as its name implies, is the target of inference. The goal in missing data problems is to estimate the latter from the former. By chain rule of probability,

$$p(\mathbf{X^{(1)}}, \mathbf{O}) = \frac{p(\mathbf{X}, \mathbf{O}, \mathbf{R} = \mathbf{1})}{p(\mathbf{R} = \mathbf{1}|\mathbf{X^{(1)}}, \mathbf{O})}. \quad (2)$$

In other words, $p(\mathbf{X^{(1)}}, \mathbf{O})$ is identified from the observed data law $p(\mathbf{R}, \mathbf{O}, \mathbf{X})$ if and only if $p(\mathbf{R} = \mathbf{1}|\mathbf{X^{(1)}}, \mathbf{O})$ is. In general, $p(\mathbf{X^{(1)}})$ is not identified from the observed data law, unless sufficient restrictions are placed on the full data law defining the missing data model.

Many popular missing data models may be represented as a factorization of the full data law with respect to a DAG [4]. These include the permutation model, the monotone MAR model, the block sequential MAR model, and certain submodels of the no self-censoring model [9, 12, 18].

Given a set of full data laws $p(\mathbf{X^{(1)}}, \mathbf{O}, \mathbf{R})$, a DAG $\mathcal{G}$ with the following properties may be used to represent a missing data model: $\mathcal{G}$ has a vertex set $\mathbf{X^{(1)}}, \mathbf{O}, \mathbf{R}, \mathbf{X}$; for each $X_i \in \mathbf{X}$, $\mathrm{pa}_{\mathcal{G}}(X_i) = \{R_i, X_i^{(1)}\}$; for each $R_i \in \mathbf{R}$, $\mathrm{de}_{\mathcal{G}}(R_i) \cap (\mathbf{X}^{(1)} \cup \mathbf{O}) = \emptyset$. Given a DAG $\mathcal{G}$ with the above properties, a missing data model associated with $\mathcal{G}$ is the set of distributions $p(\mathbf{X^{(1)}}, \mathbf{O}, \mathbf{R})$ that can be written as

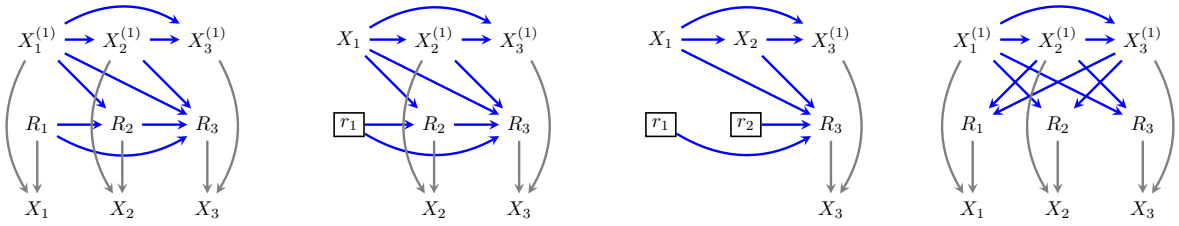
$$\prod_{X_i \in \mathbf{X}} p(X_i|R_i, X_i^{(1)}) \prod_{V \in \mathbf{X}^{(1)} \cup \mathbf{O} \cup \mathbf{R}} p(V|\,\mathrm{pa}_{\mathcal{G}}(V)), \quad (3)$$

where the set of factors of the form $p(X_i|R_i, X_i^{(1)})$ are deterministic to remain consistent with the definition of X_i. Note that by standard results on DAG models, conditional independences in $p(\mathbf{X^{(1)}}, \mathbf{O}, \mathbf{R})$ may be read off from $\mathcal{G}$ by the d-separation criterion [5].

4 EXAMPLES OF IDENTIFIED MODELS

In this section, we describe a set of examples of missing data models that factorize as in (3) for different DAGs, where the target law is identified. We start with simpler examples where sequential fixing techniques from causal inference suffice to obtain identification, then move on to describe more complex examples where existing algorithms in the literature suffice, and finally proceed to examples where no published method known to us obtains identification, illustrating an identifiability gap in existing methods. In these examples, we show how identification may be obtained by appropriately generalizing existing techniques. In these discussions, we concentrate on obtaining identification of the nuisance law $p(\mathbf{R}|\mathbf{X^{(1)}}, \mathbf{O})$ evaluated at $\mathbf{R} = \mathbf{1}$, as this suffices to identify the target law $p(\mathbf{X^{(1)}}, \mathbf{O})$ by (2). In the course of describing these examples, we will obtain intermediate graphs and kernels. In these graphs, lower case letters (e.g. v) indicates the variable V is evaluated at v (for $R_i, r_i = 1$). A square vertex indicates V had been fixed. Drawing the vertex normally with lower case indicates V was conditioned on (creating selection bias in the subproblem). For brevity, we use $\mathbf{1}_{R_i}$ to denote $\{R_i = 1\}$.

We first consider the block-sequential MAR model [18], shown in Fig. 2 for three variables. The target law is identified by applying the (valid) fixing sequence $\langle R_1, R_2, R_3\rangle$ via the operator ϕ to $\mathcal{G}$ and $p(\mathbf{R}, \mathbf{X})$. We proceed as follows. $p(R_1|\,\mathrm{pa}_{\mathcal{G}}(R_1)) = p(R_1|\,\mathrm{nd}_{\mathcal{G}}(R_1)) = p(R_1)$ is identified immediately. Applying the fixing operator ϕ_{R_1} yields the graph $\mathcal{G}_1 \equiv \phi_{R_1}(\mathcal{G})$ shown in Fig. 2(b), and corresponding kernel $q_1(X_1^{(1)}, X_2, X_3, R_2, R_3|\mathbf{1}_{R_1}) \equiv p(X_1, X_2, X_3, R_2, R_3, \mathbf{1}_{R_1})/p(\mathbf{1}_{R_1})$ where $X_1^{(1)}$ is now observed. Thus, in the new subproblem represented by $\mathcal{G}_1$ and q_1, $p(R_2|\,\mathrm{pa}_{\mathcal{G}}(R_2))|_{\mathbf{R}=\mathbf{1}} = q_1(R_2|X_1^{(1)}, \mathbf{1}_{R_1})$ is identified. Applying the fixing operator ϕ_{R_2} to $\mathcal{G}_1$ and q_1 yields $\mathcal{G}_2 \equiv \phi_{R_2}(\mathcal{G}_1)$ shown



(a) $\mathcal{G}$ (b) $\mathcal{G}_1 \equiv \phi_{R_1}(\mathcal{G})$ (c) $\mathcal{G}_2 \equiv \phi_{R_2}(\mathcal{G}_1)$ (d)

Figure 2: (a), (b), (c) are intermediate graphs obtained in identification of a block-sequential model by fixing $\{R_1, R_2, R_3\}$ in sequence. (d) is an MNAR model that is identifiable by fixing all Rs in parallel.

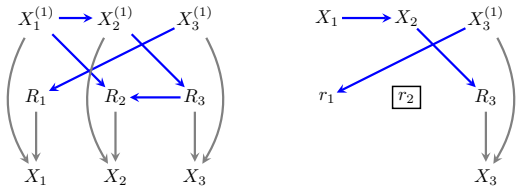
in Fig. 2(c), and $q_2(X_1^{(1)}, X_2^{(1)}, X_3, R_3|\mathbf{1}_{R_1,R_2}) = q_1(X_1^{(1)}, X_2, X_3, R_2, R_3|\mathbf{1}_{R_1})/q_1(R_2|X_1^{(1)}, \mathbf{1}_{R_1})$. Finally, in the new subproblem represented by $\mathcal{G}_2$ and q_2, $p(R_3|\,\mathrm{pa}_{\mathcal{G}}(R_3))|_{\mathbf{R}=\mathbf{1}} = q_2(R_3|X_1^{(1)}, X_2^{(1)}, \mathbf{1}_{R_1,R_2})$ is identified. Applying the fixing operator ϕ_{R_3} to $\mathcal{G}_2$ and q_2 yields $q_3(X_1^{(1)}, X_2^{(1)}, X_3^{(1)}|\mathbf{1}_{R_1,R_2,R_3}) = p(X_1^{(1)}, X_2^{(1)}, X_3^{(1)})$. The identifying functional for the target law only involves monotone cases (cases where $R_i = 0$ implies $R_{i+1} = 0$) just as would be the case under the monotone MAR model, although this model does not assume monotonicity and is not MAR. In this simple example, identification may be achieved purely by causal inference methods, by treating variables in $\mathbf{R}$ as treatments, and finding a valid fixing sequence on them. In this example, each R_i in the sequence is fixable given that the previous variables are fixable, since all parents of each R_i become observed at the time it is fixed.

Following a total order to fix is not always sufficient to identify the target law, as noted in [4, 3, 13]. Consider the model represented by DAG in Fig. 2(d). For any R_i in this model, say R_1, we have, by d-separation, that $p(R_1|\,\mathrm{pa}_{\mathcal{G}}(R_1)) = p(R_1|X_2^{(1)}, X_3^{(1)}, \mathbf{1}_{R_2,R_3})$, which is identified. However, if we were to fix R_1 in $p(\mathbf{X}, \mathbf{R})$, we would obtain a kernel $q_1(X_1^{(1)}, X_2, X_3, \mathbf{1}_{R_2,R_3}|\mathbf{1}_{R_1})$ where selection bias on R_2 and R_3 is introduced. The fact that q_1 is not available at all levels of R_2 and R_3 prevents us from sequentially obtaining $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$, for $R_i = R_2, R_3$, due to our inability to sum out those variables from q_1.

The model in Fig. 2(d) allows identification of the target law in another way, however. This follows from the fact that $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$ is identified for each R_i by exploiting conditional independences in $p(\mathbf{X}, \mathbf{R})$ displayed by Fig. 2(d). Since $p(\mathbf{R}|\mathbf{X}^{(\mathbf{1})}) = \prod_{i=1}^{3} p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$, the nuisance law is identified, which means the target law is also identified, as long as we fix R_1, R_2, R_3 *in parallel* (as in (2)) rather than sequentially. In other words, the model is identified, but no total order on fixing operations suffices for identification. A general algorithm that aimed to fix indicators in $\mathbf{R}$ in parallel, while potentially exploiting causal inference fixing operations to identify each $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$ was proposed in [13]. Our subsequent examples show that this algorithm is insufficient to obtain identification of the target law in general, and thus is incomplete.

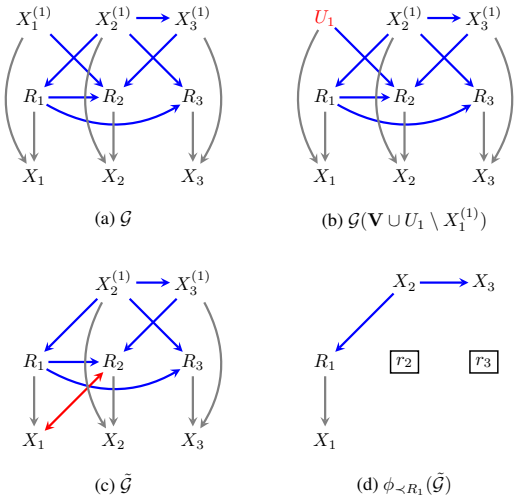
Consider the DAG in Fig. 3. Since R_2 is a child of R_3 and $X_2^{(1)}$ is a parent of R_3, we cannot obtain $p(R_3|\,\mathrm{pa}_{\mathcal{G}}(R_3)) = p(R_3|X_2^{(1)})$ by d-separation in any kernel (including the original distribution) where R_2 is not fixed. Thus, any total order on fixing operations of elements in $\mathbf{R}$ must start with R_1 or R_2. Fixing either of these variables entails dividing $p(\mathbf{X}, \mathbf{R})$ by some factor $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$, which is identified as either $p(R_1|X_3^{(1)}, \mathbf{1}_{R_3})$ or $p(R_2|X_1^{(1)}, \mathbf{1}_{R_1})$. This division entails inducing selection bias on the subsequent kernel q_1 for a variable not yet fixed (either R_3 or R_1). Thus, no total order on fixing operations works to identify the target law in this model. At the same time, attempting to fix all R variables in parallel would fail as well, since we cannot identify $p(R_3|X_2^{(1)})$ either in the original distribution or any kernel obtained by standard causal inference operations described in [13]. In particular, in any such kernel or distribution R_3 remains dependent on R_2 given $X_2^{(1)}$.

However, the target law in this model is identified by following a *partial order* $\prec$ of fixing operations. In this partial order, R_1 is incompatble with R_2, and $R_2 \prec R_3$. This results in an identification strategy where we fix each variable *only* given that variables earlier than it in the *partial* order are fixed. That is, distributions $p(R_1|X_3^{(1)}) = p(R_1|X_3, \mathbf{1}_{R_3})$ and $p(R_2|X_1^{(1)}, R_3) = p(R_2|X_1, \mathbf{1}_{R_1}, R_3)$ are obtained directly in the original distribution without fixing anything. The distribution $p(R_3|\,\mathrm{pa}_{\mathcal{G}}(R_3))$, on the other hand, is obtained in the kernel $q_1(X_1, X_2^{(1)}, X_3, \mathbf{1}_{R_1}, R_3|\mathbf{1}_{R_2}) = p(\mathbf{X}, \mathbf{R})/p(R_2|X_1, \mathbf{1}_{R_1}, R_3)$ after R_2 (the variable earlier than R_3 in the partial order) is fixed. The graph cor-



(a) $\mathcal{G}$ (b) $\phi_{R_2}(\mathcal{G})$

Figure 3: (a) A DAG where Rs are fixed according to a partial order. (b) The CADMG obtained by fixing R_2.



(a) $\mathcal{G}$ (b) $\mathcal{G}(\mathbf{V} \cup U_1 \setminus X_1^{(1)})$

(c) $\tilde{\mathcal{G}}$ (d) $\phi_{\prec R_1}(\tilde{\mathcal{G}})$

Figure 4: A DAG where selection bias on R_1 is avoidable by following a partial order fixing schedule on an ADMG induced by latent projecting out $X_1^{(1)}$.

responding to this kernel is shown in Fig. 3(b). Note that in this graph $X_2^{(1)}$ is observed, and there is selection bias on R_1. However, it easily follows by d-separation that R_3 is independent of R_1. It can thus be shown that $p(R_3|X_2^{(1)}) = q_1(R_3|X_2^{(1)}, \mathbf{1}_{R_2})$ even if q_1 is only available at value $R_1 = 1$. Since all $p(R_i|\operatorname{pa}_{\mathcal{G}}(R_i))$ are identified, so is the target law in this model, by (2).

Next, we consider the model in Fig. 4. Here, $p(R_2|X_1^{(1)}, X_3^{(1)}, R_1) = p(R_2|X_1, X_3, \mathbf{1}_{R_1,R_3})$ and $p(R_3|X_2^{(1)}, R_1) = p(R_3|X_2, \mathbf{1}_{R_2}, R_1)$ are identified immediately. However, $p(R_1|X_2^{(1)})$ poses a problem. In order to identify this distribution, we either require that R_1 is conditionally independent of R_2, possibly after some fixing operations, or we are able to render $X_2^{(1)}$ observable by fixing R_2 in some way. Neither seems to be possible in the problem as stated. In particular, fixing R_2 via dividing by $p(R_2|X_1^{(1)}, X_3^{(1)}, R_1)$ will necessarily induce selection bias on R_1, which will prevent identification of $p(R_1|X_2^{(1)})$ in the resulting kernel.

However, we can circumvent the difficulty by treating $X_1^{(1)}$ as an *unobserved variable* U_1, and attempting the problem in the resulting (hidden variable) DAG shown in Fig. 4(b), and its latent projection ADMG $\tilde{\mathcal{G}}$ shown in Fig. 4(c), where U_1 is "projected out." In the resulting problem, we can fix variables according to a partial order $\prec$ where R_2 and R_3 are incompatible, $R_2 \prec R_1$, and $R_3 \prec R_1$. Thus, we are able to fix R_2 and R_3 in parallel by dividing by $p(R_2|\operatorname{mb}_{\tilde{\mathcal{G}}}(R_2)) = p(R_2|X_1, R_1, X_3^{(1)}, \mathbf{1}_{R_3})$ and $p(R_3|R_1, X_2^{(1)}) = p(R_3|R_1, X_2, \mathbf{1}_{R_2})$, leading to a kernel $\tilde{q}_1(X_1, X_2^{(1)}, X_3^{(1)}, R_1|\mathbf{1}_{R_2,R_3})$, and the graph $\phi_{\prec R_1}(\tilde{\mathcal{G}})$ shown in Fig. 4(d), where notation $\phi_{\prec R_1}$ means "fix all necessary elements that occur earlier than R_1 in the partial order, in a way consistent with that partial order." In this example, this means fixing R_2 and R_3 in parallel. We will describe how fixing operates given general *fixing schedules* given by a partial order later in the paper. In the kernel $\tilde{q}_1$ the parent of R_1 is observed data, meaning that $p(R_1|X_2^{(1)})$ is identified as $\tilde{q}_1(R_1|X_2, \mathbf{1}_{R_2,R_3})$. This implies the target law is identified in this model.

In general, to identify $p(R_i|\operatorname{pa}_{\mathcal{G}}(R_i))$, we may need to use separate partial fixing orders on different sets of variables for different $R_i \in \mathbf{R}$. In addition, the fact that fixing introduces selection bias sometimes results in having to divide by a kernel where a *set* of variables are random, something that was never necessary in causal inference problems. In general, for a given R_i, the goal of a fixing schedule is to arrive at a kernel where an independence exists allowing us to identify $p(R_i|\operatorname{pa}_{\mathcal{G}}(R_i))$, even if some elements of $\operatorname{pa}_{\mathcal{G}}(R_i)$ are in $\mathbf{X^{(1)}}$ in the original problem. This fixing must be given by a partial order, and sometimes on sets of variables. In addition, some elements of $\mathbf{X^{(1)}}$ must be treated as hidden variables. These complications are necessary in general to avoid creating selection bias in subproblems, and ultimately to identify the nuisance law. The following example is a good illustration.

Consider the graph in Fig. 5(a). For R_1 and R_3, the fixing schedules are empty, and we immediately obtain their distributions as $p(R_1|X_2^{(1)}, X_4^{(1)}, R_2, R_3) = p(R_1|X_2, X_4, R_3, \mathbf{1}_{R_2,R_4})$ and $p(R_3|X_4^{(1)}, R_2) = p(R_3|X_4, \mathbf{1}_{R_4}, R_2)$. For R_2, the partial order is $R_3 \prec R_1$ in a graph where we treat $X_2^{(1)}$ as a hidden variable U_2. This yields $p(R_2|X_1^{(1)}, R_4) = q_2(R_2|X_1^{(1)}, R_4, \mathbf{1}_{R_1,R_3})$, where $q_2(X_1^{(1)}, X_2, X_3^{(1)}, X_4, R_2, \mathbf{1}_{R_4}|\mathbf{1}_{R_1,R_3})$ is equal to $\frac{q_1(X_1, X_2, X_3^{(1)}, X_4, R_1, R_2, \mathbf{1}_{R_4}|\mathbf{1}_{R_3})}{q_1(\mathbf{1}_{R_1}|X_2, X_3, X_4, R_2, \mathbf{1}_{R_3,R_4})}$, and

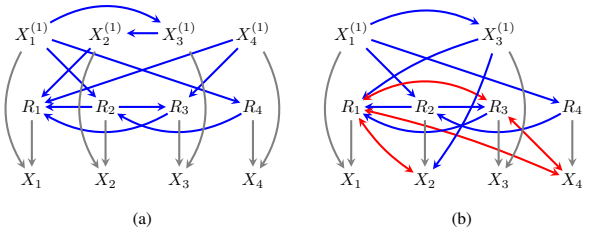


Figure 5: (a) A DAG where the fixing operator must be performed on a set of vertices. (b) A latent projection of a subproblem used for identification of $p(R_4|X_1^{(1)})$.

$$q_1(X_1, X_2, X_3^{(1)}, X_4, R_1, R_2, \mathbf{1}_{R_4}|\mathbf{1}_{R_3}) = \frac{p(\mathbf{X}, R_1, R_2, \mathbf{1}_{R_3,R_4})}{p(\mathbf{1}_{R_3}|R_2, X_4, \mathbf{1}_{R_4})}.$$

In order to obtain the propensity score for R_4 we must either render $X_1^{(1)}$ observable through fixing R_1 or perform valid fixing operations until we obtain a kernel in which R_4 is conditionally independent of R_1 given its parent $X_1^{(1)}$. However, there exists no partial order on elements of $\mathbf{R}$. All partial orders on elements in $\mathbf{R}$ induce selection bias on variables higher in the order, preventing the identification of the required distribution for R_4. For example, choosing a partial fixing order of $R_1 \prec R_3$, where we treat $X_2^{(1)}$ and $X_4^{(1)}$ as hidden variables results in selection bias on R_3 as soon as we fix R_1. Other partial orders fail similarly. However, the following approach is possible in the graph in which we treat $X_2^{(1)}$ and $X_4^{(1)}$ as hidden variables.

R_1 and R_3 lie in the same district in the resulting latent projection ADMG, shown in Fig. 5(b). Moreover, the set $\{R_1, R_3\}$ is closed under descendants in the district in Fig. 5(b). As a result, R_1 and R_3 can essentially be viewed as a single vertex from the point of view of fixing. Indeed we may choose a partial order $\{R_1, R_3\} \prec R_2$, where we fix R_1 and R_3 as a set. The fixing operation on the set is possible since $p(\mathbf{1}_{R_1,R_3}|\,\mathrm{mb}(R_1, R_3)) = p(\mathbf{1}_{R_1,R_3}|R_2, R_4, X_2, X_3^{(1)}, X_4)$ is a function of observed data law, $p(\mathbf{X}, \mathbf{R})$. Specifically, it is equal to $p(\mathbf{1}_{R_3}|R_2, R_4, X_2, X_4)p(\mathbf{1}_{R_1}|R_2, R_4, X_2, X_3, X_4, \mathbf{1}_{R_3})$, where the equality holds by d-separation $(R_3 \perp\!\!\!\perp X_3^{(1)}|R_2, R_4, X_2, X_4)$. We then obtain $p(R_4|X_1^{(1)}) = \frac{\sum_{X_3^{(1)},X_4} q_2(X_1^{(1)},X_3^{(1)},X_4,R_4|\mathbf{1}_{R_1,R_2,R_3})}{\sum_{X_3^{(1)},X_4,R_4} q_2(X_1^{(1)},X_3^{(1)},X_4,R_4|\mathbf{1}_{R_1,R_2,R_3})}$, where $q_2(.|\mathbf{1}_{\mathbf{R}\setminus R_4}) = \frac{q_1(X_1^{(1)},X_2,X_3^{(1)},X_4,R_2,R_4|\mathbf{1}_{R_1,R_3})}{q_1(R_2|X_1^{(1)},R_4,\mathbf{1}_{R_1,R_3})}$, and $q_1(.|\mathbf{1}_{R_1,R_3}) = \frac{p(\mathbf{X},R_2,R_4,\mathbf{1}_{R_1,R_3})}{p(\mathbf{1}_{R_1,R_3}|R_2,R_4,X_2,X_3^{(1)},X_4)}$.

Our final example demonstrates that in order to identify the target law, we may potentially need to fix variables outside $\mathbf{R}$, including variables in $\mathbf{X}^{(\mathbf{1})}$ that become observed after fixing or conditioning on some elements of $\mathbf{R}$. Fig. 6(a) contains a generalization of the model considered in [13], where O_3 is fully observed. In this model, distributions for R_4 and R_1 are identified immediately, while identification of R_2 requires a partial order $R_4 \prec X_4^{(1)} \prec O_3 \prec R_1$ in the graph where we treat $X_1^{(1)}, X_2^{(1)}, X_4^{(1)}$ as latent variables (with the latent projection ADMG shown in Fig. 6(b)) until they are rendered observed by fixing the corresponding missingness indicators. To illustrate fixing operations according to this order, the intermediate graphs that arise are shown in Fig. 6(c),(d),(e),(f).

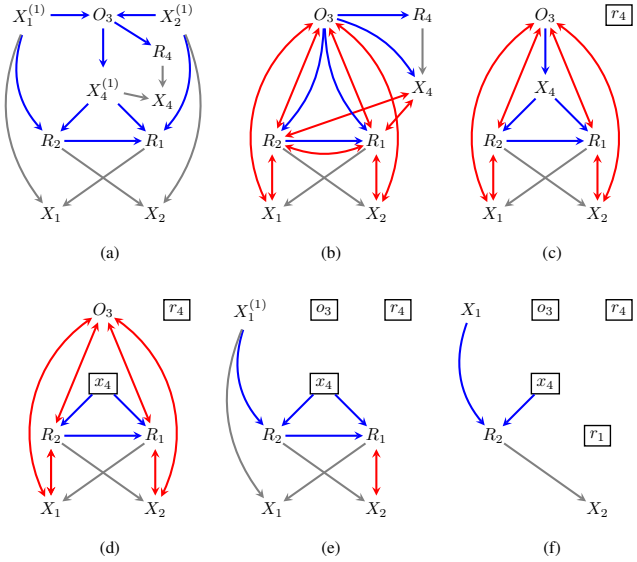


Figure 6: A DAG where variables besides Rs are required to be fixed.

5 A NEW IDENTIFICATION ALGORITHM

In order to identify the target law in examples discussed in the previous section, we had to consider situations where some variables were viewed as hidden, and marginalized out, and others were conditioned on, introducing selection bias. In addition, fixing operations were performed according to a partial, rather than a total, order as was the case in causal inference problems. Finally, we sometimes fixed sets of variables jointly, rather than individual variables. We now introduce relevant definitions that allow us to formulate a general identification algorithm that takes advantage of all these techniques.

Let $\mathbf{V}$ be a set of random variables (and corresponding vertices) consisting of observed variables $\mathbf{O}, \mathbf{R}, \mathbf{X}$, miss-

ing variables $\mathbf{X}^{(\mathbf{1})}$, and selected variables $\mathbf{S}$. Let $\mathbf{W}$ be a set of fixed observed variables. The following definitions apply to a latent projection $\mathcal{G}(\mathbf{V} \setminus \mathbf{X}_{\mathbf{U}}^{(\mathbf{1})}, \mathbf{W})$, for some $\mathbf{X}_{\mathbf{U}}^{(\mathbf{1})} \subseteq \mathbf{X}^{(\mathbf{1})}$, and a corresponding kernel $q(\mathbf{V} \setminus \mathbf{X}_{\mathbf{U}}^{(\mathbf{1})}|\mathbf{W}) \equiv \sum_{\mathbf{X}_{\mathbf{U}}^{(\mathbf{1})}} q(\mathbf{V}|\mathbf{W})$. Graph $\mathcal{G}$ can be viewed as a latent variable CADMG for q where $\mathbf{X}_{\mathbf{U}}^{(\mathbf{1})}$ are latent. Such CADMGs represent intermediate subproblems in our identification algorithm.

For $\mathbf{Z} \subseteq \mathbf{D}_{\mathbf{Z}} \in \mathcal{D}(\mathcal{G})$, let $\mathbf{R}_{\mathbf{Z}} = \{R_j | X_j^{(1)} \in \mathbf{Z} \cup \mathrm{mb}_{\mathcal{G}}(\mathbf{Z}), R_j \notin \mathbf{Z}\}$, and $\mathrm{mb}_{\mathcal{G}}(\mathbf{Z}) \equiv (\mathbf{D}_{\mathbf{Z}} \cup \mathrm{pa}_{\mathcal{G}}(\mathbf{D}_{\mathbf{Z}})) \setminus \mathbf{Z}$. We say $\mathbf{Z}$ is *fixable* in $\mathcal{G}(\mathbf{V} \setminus \mathbf{X}_{\mathbf{U}}^{(\mathbf{1})}, \mathbf{W})$ if

(i) $\mathrm{de}_{\mathcal{G}}(\mathbf{Z}) \cap \mathbf{D}_{\mathbf{Z}} \subseteq \mathbf{Z}$,

(ii) $\mathbf{S} \cap \mathbf{Z} = \emptyset$,

(iii) $\mathbf{Z} \perp\!\!\!\perp (\mathbf{S} \cup \mathbf{R}_{\mathbf{Z}}) \setminus \mathrm{mb}_{\mathcal{G}}(\mathbf{Z}) | \mathrm{mb}_{\mathcal{G}}(\mathbf{Z})$.

In words, these conditions apply to some $\mathbf{Z}$ that is a subset of its own district (which is trivial when the set $\mathbf{Z}$ is a singleton). The conditions, in the listed order, require that $\mathbf{Z}$ is closed under descendants in the district, should not contain any selected variables, and should be independent of both selected variables $\mathbf{S}$ and the missingness indicators $\mathbf{R}_{\mathbf{Z}}$ of the corresponding counterfactal parents given the Markov blanket of $\mathbf{Z}$, respectively. Consider the graph in Fig. 5(b) where $\mathbf{S} = \emptyset$ and let $\mathbf{Z} = \{R_1, R_3\}$. $\mathbf{Z}$ is fixable since $\mathbf{Z} \subseteq \mathbf{D}_{\mathbf{Z}} = \{R_1, R_3, X_2, X_4\}, \mathrm{de}_{\mathcal{G}}(\mathbf{Z}) = \{R_1, R_3, X_1, X_3\} \cap \mathbf{D}_{\mathbf{Z}} = \{R_1, R_3\}$ is closed, and both $\mathbf{S}$ and $\mathbf{R}_{\mathbf{Z}}$ are empty sets.

A set $\widetilde{\mathbf{Z}}$ spanning multiple elements in $\mathcal{D}(\mathcal{G})$ is said to be fixable if it can be partitioned into a set $\mathcal{Z}$ of elements $\mathbf{Z}$, such that each $\mathbf{Z}$ is a subset of a single district in $\mathcal{D}(\mathcal{G})$ and is fixable.

Given an ordering $\prec$ on vertices $\mathbf{V} \cup \mathbf{W}$ topological in $\mathcal{G}$ and $\widetilde{\mathbf{Z}}$ fixable in $\mathcal{G}$, define $\phi_{\widetilde{\mathbf{Z}}}(q; \mathcal{G})$ as

$$\frac{q(\mathbf{V} \setminus (\mathbf{X}_{\mathbf{U}}^{(\mathbf{1})} \cup \mathbf{R}_{\mathbf{Z}}), \mathbf{R}_{\mathbf{Z}} = \mathbf{1}|\mathbf{W})}{\prod_{\mathbf{Z} \in \mathcal{Z}} \prod_{Z \in \mathbf{Z}} q(Z | \mathrm{mb}_{\mathcal{G}}(Z; \mathrm{an}_{\mathcal{G}}(\mathbf{D}_{\mathbf{Z}}) \cap \{\preceq Z\})), \mathbf{R}_{\mathbf{Z}})|_{(\mathbf{R} \cap \mathbf{Z}) \cup \mathbf{R}_{\mathbf{Z}} = \mathbf{1}}}, \quad (4)$$

where $\mathrm{mb}_{\mathcal{G}}(V; \mathbf{S}) \equiv \mathrm{mb}_{\mathcal{G}_{\mathbf{S}}}(V)$ and $\{\preceq Z\}$ is the set of all elements earlier than Z in the order $\prec$ (this includes Z itself).

Given a set $\mathbf{Z} \subseteq \mathbf{R} \cup \mathbf{O} \cup \mathbf{X}^{(\mathbf{1})}$, and an equivalence relation $\sim$, let $\mathbf{Z}/_{\sim}$ be the partition of $\mathbf{Z}$ into equivalence classes according to $\sim$. Define a *fixing schedule* for $\mathbf{Z}/_{\sim}$ to be a partial order $\lhd$ on $\mathbf{Z}/_{\sim}$. For each $\mathbf{Z} \in \mathbf{Z}/_{\sim}$, define $\{\trianglelefteq \widetilde{\mathbf{Z}}\}$ to be the set of elements in $\mathbf{Z}/_{\sim}$ earlier than $\widetilde{\mathbf{Z}}$ in the order $\lhd$, and $\{\lhd \widetilde{\mathbf{Z}}\} \equiv \{\trianglelefteq \widetilde{\mathbf{Z}}\} \setminus \widetilde{\mathbf{Z}}$. Define $\trianglelefteq_{\widetilde{\mathbf{Z}}}$ and $\lhd_{\widetilde{\mathbf{Z}}}$ to be restrictions of $\lhd$ to $\{\trianglelefteq \widetilde{\mathbf{Z}}\}$ and $\{\lhd \widetilde{\mathbf{Z}}\}$, respectively. Both restrictions, $\trianglelefteq_{\widetilde{\mathbf{Z}}}$ and $\lhd_{\widetilde{\mathbf{Z}}}$, are also partial orders.

We inductively define a *valid* fixing schedule (a schedule where fixing operations can be successfully implemented), along with the fixing operator on valid schedules. The fixing operator will implement fixing as in (4) on $\widetilde{\mathbf{Z}}$ within an intermediate problem represented by a CADMG where some $\mathbf{X}_{\widetilde{\mathbf{Z}}}^{(\mathbf{1})} \subseteq \mathbf{X}^{(\mathbf{1})}$ will become observed after fixing $\widetilde{\mathbf{Z}}$, with $\mathbf{X}^{(\mathbf{1})} \setminus \mathbf{X}_{\widetilde{\mathbf{Z}}}^{(\mathbf{1})}$ treated as latent variables, and a kernel associated with this CADMG defined on the observed subset of variables. We also define $\mathbf{X}_{\{\trianglelefteq \widetilde{\mathbf{Z}}\}}^{(\mathbf{1})} \equiv \bigcup_{\mathbf{Z} \in \{\trianglelefteq \widetilde{\mathbf{Z}}\}} \mathbf{X}_{\mathbf{Z}}^{(\mathbf{1})}$.

We say $\lhd_{\widetilde{\mathbf{Z}}}$ is valid for $\{\lhd \widetilde{\mathbf{Z}}\}$ in $\mathcal{G}$ if for every $\lhd$-largest element $\widetilde{\mathbf{Y}}$ of $\{\lhd \widetilde{\mathbf{Z}}\}$, $\trianglelefteq_{\widetilde{\mathbf{Y}}}$ is valid for $\{\trianglelefteq \widetilde{\mathbf{Y}}\}$. If $\lhd_{\widetilde{\mathbf{Z}}}$ is valid for $\{\lhd \widetilde{\mathbf{Z}}\}$, we define $\phi_{\lhd_{\widetilde{\mathbf{Z}}}}(\mathcal{G})$ to be a new CADMG $\mathcal{G}(\mathbf{V} \setminus \bigcup_{\mathbf{Z} \in \{\lhd \widetilde{\mathbf{Z}}\}} \mathbf{Z}, \mathbf{W} \cup \bigcup_{\mathbf{Z} \in \{\lhd \widetilde{\mathbf{Z}}\}} \mathbf{Z})$ obtained from $\mathcal{G}(\mathbf{V}, \mathbf{W})$ by:

- Removing all edges with arrowheads into $\bigcup_{\mathbf{Z} \in \{\lhd \widetilde{\mathbf{Z}}\}} \mathbf{Z}$,
- Marking any $\{X_j^{(1)} | X_j^{(1)} \in \mathbf{Z} \cup \mathrm{mb}_{\phi_{\lhd_{\mathbf{Z}}}(\mathcal{G})}(\mathbf{Z}), \mathbf{Z} \in \{\lhd \widetilde{\mathbf{Z}}\}\}$ as observed,
- Marking any $\{\mathbf{R}_{\mathbf{Z}} \cap \mathbf{V} | \mathbf{Z} \in \{\lhd \widetilde{\mathbf{Z}}\}\} \setminus \bigcup_{\mathbf{Z} \in \{\lhd \widetilde{\mathbf{Z}}\}} \mathbf{Z}$ as selected to value 1, where $\mathbf{R}_{\mathbf{Z}}$ is defined with respect to $\phi_{\lhd_{\mathbf{Z}}}(\mathcal{G})$
- Treating elements of $\mathbf{X}^{(\mathbf{1})} \setminus \mathbf{X}_{\widetilde{\mathbf{Z}}}^{(\mathbf{1})}$ as hidden variables.

We say $\trianglelefteq_{\widetilde{\mathbf{Z}}}$ is valid for $\{\trianglelefteq \widetilde{\mathbf{Z}}\}$ if $\lhd_{\widetilde{\mathbf{Z}}}$ is valid for $\{\lhd \widetilde{\mathbf{Z}}\}$ and $\widetilde{\mathbf{Z}}$ is fixable in $\phi_{\lhd_{\widetilde{\mathbf{Z}}}}(\mathcal{G})$. If $\trianglelefteq_{\widetilde{\mathbf{Z}}}$ is valid, we define

$$\phi_{\trianglelefteq_{\widetilde{\mathbf{Z}}}}(q; \mathcal{G}) \equiv \phi_{\widetilde{\mathbf{Z}}} \left(\phi_{\lhd_{\widetilde{\mathbf{Z}}}}(q; \mathcal{G}); \phi_{\lhd_{\widetilde{\mathbf{Z}}}}(\mathcal{G}) \right), \quad (5)$$

where $\phi_{\lhd_{\widetilde{\mathbf{Z}}}}(q; \mathcal{G}) \equiv \frac{q(\mathbf{V}|\mathbf{W})}{\prod_{\widetilde{\mathbf{Y}} \in \{\lhd \widetilde{\mathbf{Z}}\}} q_{\widetilde{\mathbf{Y}}}}$, and $q_{\widetilde{\mathbf{Y}}}$ are defined inductively as the denominator of (4) for $\widetilde{\mathbf{Y}}$, $\phi_{\lhd_{\widetilde{\mathbf{Y}}}}(\mathcal{G})$ and $\phi_{\lhd_{\widetilde{\mathbf{Y}}}}(q; \mathcal{G})$.

We have the following claims.

Proposition 1. *Given a DAG $\mathcal{G}(\mathbf{X}^{(\mathbf{1})}, \mathbf{R}, \mathbf{O}, \mathbf{X})$, the distribution $p(R_i | \mathrm{pa}_{\mathcal{G}}(R_i))|_{\mathrm{pa}_{\mathcal{G}}(R_i) \cap \mathbf{R} = \mathbf{1}}$ is identifiable from $p(\mathbf{R}, \mathbf{O}, \mathbf{X})$ if there exists*

(i) $\mathbf{Z} \subseteq \mathbf{X}^{(\mathbf{1})} \cup \mathbf{R} \cup \mathbf{O}$,

(ii) *an equivalence relation $\sim$ on $\mathbf{Z}$ such that* $\{R_i\} \in \mathbf{Z}/_{\sim}$,

(iii) *a set of elements $\mathbf{X}_{\widetilde{\mathbf{Z}}}^{(\mathbf{1})}$ such that $\mathbf{X}_{\{\lhd \widetilde{\mathbf{Z}}\}}^{(\mathbf{1})} \subseteq \mathbf{X}_{\widetilde{\mathbf{Z}}}^{(\mathbf{1})} \subseteq \mathbf{X}^{(\mathbf{1})}$ for each $\widetilde{\mathbf{Z}} \in \mathbf{Z}/_{\sim}$,*

(iv) $\mathbf{X}^{(\mathbf{1})} \cap \mathrm{pa}_{\mathcal{G}}(R_i) \subseteq (\mathbf{Z} \setminus \{R_i\}) \cup \mathbf{X}_{\{R_i\}}^{(\mathbf{1})}$,

(v) and a valid fixing schedule $\lhd$ for ${}^{\mathbf{Z}}/_{\sim}$ in $\mathcal{G}$ such that for each $\widetilde{\mathbf{Z}} \in {}^{\mathbf{Z}}/_{\sim}$, $\widetilde{\mathbf{Z}} \lhd \{R_i\}$.

Moreover, $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))|_{\mathrm{pa}_{\mathcal{G}}(R_i)\cap\mathbf{R}=\mathbf{1}}$ is equal to $q_{\{R_i\}}$, defined inductively as the denominator of (4) for $\{R_i\}$, $\phi_{\lhd_{\{R_i\}}}(\mathcal{G})$ and $\phi_{\lhd_{\{R_i\}}}(p;\mathcal{G})$, and evaluated at $\mathrm{pa}_{\mathcal{G}}(R_i)\cap\mathbf{R}=\mathbf{1}$.

Proposition 1 implies that $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$ is identified if we can find a set of variables that can be fixed according to a partial order (possibly through set fixing) within subproblems where certain variables are hidden. At the end of the fixing schedule, we require that R_i itself is fixable given its Markov blanket in the original DAG. We encourage the reader to view the example provided in Appendix B, for a demonstration of valid fixing schedules that may be chosen by Proposition 1.

Corollary 1. *Given a DAG $\mathcal{G}(\mathbf{X}^{(\mathbf{1})},\mathbf{R},\mathbf{O},\mathbf{X})$, the target law $p(\mathbf{X}^{(\mathbf{1})},\mathbf{O})$ is identified if $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$ is identified via Proposition 1 for every $R_i\in\mathbf{R}$.*

Proof. Follows by Proposition 1 and (2). □

In addition, in special classes of models, the full law, rather than just the target law is identified.

Proposition 2. *Given a DAG $\mathcal{G}(\mathbf{X}^{(\mathbf{1})},\mathbf{R},\mathbf{O},\mathbf{X})$, the full law $p(\mathbf{R},\mathbf{X}^{(\mathbf{1})},\mathbf{O})$ is identifiable from $p(\mathbf{R},\mathbf{O},\mathbf{X})$ if for every $R_i\in\mathbf{R}$, all conditions in Proposition 1 (i-v) are met, and also for each $\widetilde{\mathbf{Z}}\in{}^{\mathbf{Z}}/_{\sim}$, $\mathbf{X}^{(\mathbf{1})}_{\widetilde{\mathbf{Z}}}$ does not contain any elements in $\{X_j^{(1)}|R_j\in\mathrm{pa}_{\mathcal{G}}(R_i)\}$. Moreover, $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$ is equal to $q_{\{R_i\}}$, defined inductively as the denominator of (4) for $\{R_i\}$, $\phi_{\lhd_{\{R_i\}}}(\mathcal{G})$ and $\phi_{\lhd_{\{R_i\}}}(p;\mathcal{G})$, and*

$$p(\mathbf{R},\mathbf{X}^{(\mathbf{1})},\mathbf{O})=\Big(\prod_{R_i\in\mathbf{R}}q_{R_i}\Big)\times\frac{p(\mathbf{R}=\mathbf{1},\mathbf{O},\mathbf{X})}{\Big(\prod_{R_i\in\mathbf{R}}q_{R_i}\Big)|_{\mathbf{R}=\mathbf{1}}}$$

Proof. Under conditions (i-v) in Proposition 1, we are guaranteed to identify the target law and obtain $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$ where some $R_j\in\mathrm{pa}_{\mathcal{G}}(R_i)$ may be evaluated at $R_j=1$. Under the additional restriction stated above, all $R_j\in\mathrm{pa}_{\mathcal{G}}(R_i)$ can be evaluated at all levels. □

Proposition 2 always fails if a special collider structure $X_j^{(1)}\to R_i\leftarrow R_j$, which we call *the colluder*, exists in $\mathcal{G}$. The following Lemma implies that colluders always imply the full law is not identified.

Lemma 1. *In a DAG $\mathcal{G}(\mathbf{X}^{(\mathbf{1})},\mathbf{R},\mathbf{O},\mathbf{X})$, if there exists $R_i,R_j\in\mathbf{R}$ such that $\{R_j,X_j^{(1)}\}\in\mathrm{pa}_{\mathcal{G}}(R_i)$, then $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))|_{R_j=0}$ is not identified. Hence, the full law $p(\mathbf{X}^{(\mathbf{1})},\mathbf{R})$ is not identified.*

Proof. Follows by providing two different full laws that agree on the observed law on a DAG with 2 counterfactual random variables (Appendix C). This result holds for an arbitrary DAG representing a missing data model that contains the colluder structure mentioned above. □

Propositions 1 and 2 do not address a computationally efficient search procedure for a valid fixing schedule $\lhd$ that permit identification of $p(R_i|\,\mathrm{pa}_{\mathcal{G}}(R_i))$ for a particular $R_i\in\mathbf{R}$. Nevertheless, the following Lemma shows how to easily obtain identification of the target law in a restricted class of missing data DAGs.

Lemma 2. *Consider a DAG $\mathcal{G}(\mathbf{X}^{(\mathbf{1})},\mathbf{R},\mathbf{O},\mathbf{X})$ such that for every $R_i\in\mathbf{R}$, $\{R_j|X_j^{(1)}\in\mathrm{pa}_{\mathcal{G}}(R_i)\}\cap\mathrm{an}_{\mathcal{G}}(R_i)=\emptyset$. Then for every $R_i\in\mathbf{R}$, a fixing schedule $\lhd$ for $\{\{R_j\}|R_j\in\mathcal{G}_{\mathbf{R}\cap\mathrm{de}_{\mathcal{G}}(R_i)}\}$ given by the partial order induced by the ancestrality relation on $\mathcal{G}_{\mathbf{R}\cap\mathrm{de}_{\mathcal{G}}(R_i)}$ is valid in $\mathcal{G}(\mathbf{X}^{(\mathbf{1})},\mathbf{R},\mathbf{O},\mathbf{X})$, by taking each $\mathbf{X}^{(\mathbf{1})}_{\widetilde{\mathbf{Z}}}=\bigcup_{\mathbf{Z}\in\{\trianglelefteq\widetilde{\mathbf{Z}}\}}\mathbf{X}^{(\mathbf{1})}_{\mathbf{Z}}$, for every $\widetilde{\mathbf{Z}}\in\{\trianglelefteq\{R_i\}\}$. Thus the target law is identified.*

6 DISCUSSION AND CONCLUSION

In this paper we addressed the significant gap present in identification theory for missing data models representable as DAGs. We showed, by examples, that straightforward application of identification machinery in causal inference with hidden variables do not suffice for identification in missing data, and discussed the generalizations required to make it suitable for this task. These generalizations included fixing (possibly sets of) variables on a partial order and avoiding selection bias by introducing hidden variables into the problem though they were not present in the initial problem statement. Proposition 1 gives a characterization of how to utilize these generalized procedures to obtain identification of the target law, while Proposition 2 gives a similar characterization for the full law. While neither of these propositions alluded to a computationally efficient algorithm to obtain identification in general, Lemma 2 provides such a procedure for a special class of missing data models where the partial order of fixing operations required for each R is easy to determine. Providing a computationally efficient search procedure for identification in all DAG models of missing data, and questions regarding the completeness of our proposed algorithm are left for future work.

Acknowledgements

This project is sponsored in part by the National Institutes of Health grant R01 AI127271-01 A1 and the Office of Naval Research grant N00014-18-1-2760.

References

[1] Yimin Huang and Marco Valtorta. Pearl's calculus of interventions is complete. In *Twenty Second Conference On Uncertainty in Artificial Intelligence*, 2006.

[2] Steffan L. Lauritzen. *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

[3] Karthika Mohan and Judea Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems*, pages 1520–1528. 2014.

[4] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems*, pages 1277–1285, 2013.

[5] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo, 1988.

[6] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[7] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv:1701.06686v2*, 2017. Working paper.

[8] James M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[9] James M. Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37, 1997.

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

[11] Mauricio Sadinle and Jerome P. Reiter. Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

[12] Ilya Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. In *Advances in Neural Information Processing Systems*, pages 3144–3152, 2016.

[13] Ilya Shpitser, Karthika Mohan, and Judea Pearl. Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pages 802–811. AUAI Press, 2015.

[14] Ilya Shpitser and Judea Pearl. 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, 2006.

[15] Eric J. Tchetgen Tchetgen, Linbo Wang, and BaoLuo Sun. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069–2088, 2018.

[16] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002.

[17] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag New York, 1st edition edition, 2006.

[18] Yan Zhou, Roderick J. A. Little, and John D. Kalbfleisch. Block-conditional missing at random models for missing data. *Statistical Science*, 25(4):517–532, 2010.