CAUSAL AND COUNTERFACTUAL VIEWS OF MISSING DATA MODELS

Razieh Nabi[1, *], Rohit Bhattacharya[2], Ilya Shpitser[3], and James M. Robins[4]

[1]Department of Biostatistics and Bioinformatics, Emory University

[2]Department of Computer Science, Williams College

[3]Department of Computer Science, Johns Hopkins University

[4]T.H. Chan School of Public Health, Harvard University

[*]Corresponding author: Razieh Nabi, razieh.nabi@emory.edu

Abstract: It is often said that the fundamental problem of causal inference is a missing data problem – the comparison of responses to two hypothetical treatment assignments is made difficult because for every experimental unit only one potential response is observed. In this paper, we consider the implications of the converse view: that missing data problems are a form of causal inference. We make explicit how the missing data problem of recovering the complete data law from the observed law can be viewed as identification of a joint distribution over counterfactual variables corresponding to values had we (possibly contrary to fact) been able to observe them. Drawing analogies with causal inference, we show how identification assumptions in missing data can be encoded in terms of graphical models defined over counterfactual and observed variables. We review recent results in missing data identification from this viewpoint. In doing so, we note interesting similarities and differences between missing data and causal identification theories.

Key words and phrases: Causal inference, causal graphs, missing data DAG models, missing not at random, non-ignorable missingness.

1. Introduction

Missing data is a common challenge in the analysis of survey, experimental, and observational data, both for the purpose of prediction and for drawing causal conclusions. Complete-case analysis is a popular and simple approach to handling missing data, but it is justified only when data entries are missing-completely-at-random (MCAR) (Rubin, 1976). When data entries are missing in a way that only depends on observed data values, the data are said to be missing-at-random (MAR) (Rubin, 1976). Under MAR assumptions, it is possible to identify target parameters of the underlying data distribution without the need for further parametric assumptions. Moreover, we can estimate parameters identified under MAR via likelihood-based methods such as expectation maximization (Dempster et al., 1977; Horton and Laird, 1999; Little and Rubin, 2002), multiple imputation (Rubin, 1987; Schafer, 1999), inverse probability weighting (Robins et al., 1994; Li et al., 2013), or semiparametric methods that exploit information about mechanisms that determine missingness and are closely related to methods for estimating causal parameters (Robins et al., 1995; Scharfstein et al., 1999; Robins and Rotnitzky, 2001; Tsiatis, 2006; Tchetgen, 2009).

However, it is often the case that missingness status depends on the underlying values that are themselves censored. This type of missingness is known as missing-not-at-random (MNAR) (Rubin, 1976). Without any assumptions, parameters of interest in an MNAR model cannot be uniquely identified from the observed data distribution. A common approach to MNAR problems is to impose sufficient parametric or semiparametric restrictions on the underlying data distribution and missingness selection model, such that they yield identification (Wu and Carroll, 1988; Little and Rubin, 2002; Wang et al., 2014; Miao et al., 2016; Miao and Tchetgen Tchetgen, 2016; Sun et al., 2018). Other approaches to handling MNAR mechanisms include conduct-

ing sensitivity analysis and obtaining nonparametric bounds (Rotnitzky et al., 1998; Robins et al., 2000; Scharfstein and Irizarry, 2003; Vansteelandt et al., 2007; Mattei et al., 2014; Moreno-Betancur and Chavance, 2016; Scharfstein et al., 2021; Duarte et al., 2021).

A successful approach to identification proceeds by imposing a set of independence restrictions among variables and missingness indicators in the model that are sufficient to express parameters of interest as functions of the observed data distribution. This approach is similar to nonparametric identification theory developed in causal inference. Causal identification theory takes advantage of independence restrictions among counterfactual and observed variables in a causal model encoded via directed acyclic graphs (DAGs); this approach has led to sound and complete algorithms for identification of a wide set of causal parameters as functions of the observed data (Tian and Pearl, 2002; Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Shpitser and Sherman, 2018; Richardson et al., 2017; Bhattacharya et al., 2020). Completeness here means that failure of the algorithm on a particular parameter input implies that the parameter is, in fact, not identified given the set of restrictions enocded by the proposed model. These algorithms generalize many existing results regarding special cases, such as identification by covariate adjustment that relies on the stable unit treatment value assumption and conditional ignorability (Rubin, 1976), or the g-computation algorithm that relies on sequential ignorability (Robins, 1986).

Directed acyclic graphs have also been adapted to encode independence restrictions in missing data models. Using such representations, many scenarios have been described where it is possible to recover target parameters as functions of the observed data distribution, including ones where it was previously believed that nonparametric identification is impossible (Glymour, 2006; Daniel et al., 2012; Martel García, 2013;

Mohan et al., 2013; Thoemmes and Rose, 2014; Tian, 2015; Shpitser, 2016a; Bhattacharya et al., 2019; Saadati and Tian, 2019; Nabi et al., 2020; Mohan and Pearl, 2021; Scharfstein et al., 2021; Nabi and Bhattacharya, 2022). This line of work has shed light on several classes of MNAR models that still permit identification of the target parameter without relying on any parametric assumptions on the full data distribution. In addition to providing concise representations of statistical models by means of factorizations, graphs also illustrate the causal mechanisms responsible for missingness and provide a natural interpretation of such mechanisms in applied settings.

It has been noted by many authors that causal inference and missing data are analogous in terminology, theory of identification, and statistical inference. Causal inference has often been phrased as a missing data problem since responses to some treatment interventions are not observed (Robins, 1986; Ding and Li, 2018) and missing data is viewed as a form of causal inference where interventions on missingness indicators can be carried out (Robins, 1986; Shpitser et al., 2015; Bhattacharya et al., 2019). At the same time, not much discussion has been devoted to important differences between these frameworks. In this paper, we discuss identifiability of models with MNAR mechanisms and examine new developments in graphical missing data models. We show how identification theory may be understood by viewing missing data models counterfactually, by analogy with causal models, and discuss the shortcomings of generalizing causal identification theories to missing data models.

The paper is organized as follows. We first provide an overview of statistical and causal models of DAGs in Section 2. In Section 3, we redefine missing data models using the described causal terminology. We formally define missing data DAG models in Section 4. In Section 5, we discuss several unique techniques for nonparametric identification of complete data distributions in graphical models of missing data. We

conclude the paper in Section 6 by providing a discussion on whether ideas explored in missing data DAG models joined with rank preservation assumptions can be used for identification arguments in causal inference. We also consider missing data DAGs that allow for the presence of unmeasured confounders in the Appendix.

2. Causal Models

To motivate causal and counterfactual views of missing data, we first provide a description of causal models, and in particular, ones that can be represented via directed acyclic graphs (DAGs). Causal models are often phrased in terms of counterfactual responses to interventions. Random variables of the form $Y^{(a)}$ are used to denote the response of an outcome Y when a treatment A is intervened on and set, possibly contrary to fact, to value a. The observed (factual) outcomes Y are typically defined as coarsened versions of counterfactual outcomes via the *consistency* property. For example, for a binary treatment A with values 0 and 1, the observed outcome is obtained via the following coarsening mechanism $Y := Y^{(a=1)} \times A + Y^{(a=0)} \times (1 - A)$. That is, the observed outcome Y gives us an imperfect view into the underlying counterfactuals $Y^{(a=1)}$ and $Y^{(a=0)}$: for individuals that received treatment $A = 1$, the observed outcome corresponds to the counterfactual response $Y^{(a=1)}$; for those that received $A = 0$, we gain information regarding the counterfactual response $Y^{(a=0)}$. In general, we see only one of the (potentially several for non-binary treatments) counterfactual responses for each individual. This complicates the task of computing causal parameters, which are phrased in terms of contrasts between different counterfactual responses. This forms the basis of the observation that the fundamental problem of causal inference is a missing data problem.

Progress is made by defining restrictions on a causal model, which are sets of joint distributions defined over the factual and counterfactual random variables. Consider

the task of identifying the average causal effect (ACE) $\mathbb{E}[Y^{(a=1)} - Y^{(a=0)}]$. Given a set of baseline covariates L we may restrict ourselves to the set of distributions satisfying an independence assumption that $A \perp\!\!\!\perp Y^{(a)} \mid L$ for every value a, and positivity of the distribution of A conditioned on L, denoted by $p_{A|L}$. Under the assumptions of this causal model, the ACE is identified via the adjustment formula: $\mathbb{E}[\mathbb{E}[Y|A = 1, L] - \mathbb{E}[Y|A = 0, L]]$. Assumptions in a causal model can often be encoded in a more intuitive fashion via DAGs. We formally introduce the statistical and causal models of a DAG below.

2.1 Statistical DAG Models

Let $V = (V_1, \ldots, V_k)^T$ be a vector of K random variables with finite support and probability density p_V. We will abbreviate the joint probability $p_V(V = v)$ as simply $p(v)$. Restrictions on the observed distribution p_V can be encoded via a DAG as follows. Define a DAG $\mathcal{G}(V)$ consisting of a set of nodes V associated with each random variable $V_i \in V$, and a set of directed edges that form connections between these variables with the restriction that these edges do not form a directed cycle. We will sometimes abbreviate $\mathcal{G}(V)$ as simply $\mathcal{G}$ if the set of vertices V is assumed or obvious. For a given DAG $\mathcal{G}$, the statistical model $\mathcal{M}^{\mathcal{G}}$ is the set of distributions that factorize as $p(v) = \prod_{v_i \in v} p\left(v_i \mid \text{pa}_{\mathcal{G}}(v_i)\right)$, where $\text{pa}_{\mathcal{G}}(v_i)$ is the set of values of variables corresponding to the parents of V_i, $\text{pa}_{\mathcal{G}}(V_i)$, i.e., the set of vertices in $\mathcal{G}$ with directed edges into V_i. Distributions in $\mathcal{M}^{\mathcal{G}}$ are said to be Markov relative to $\mathcal{G}$.

We use the following notation for standard genealogical relations in DAGs. We denote the *children* of a vertex V_i in $\mathcal{G}$ – the set of all vertices that have V_i as a parent – as $\text{ch}_{\mathcal{G}}(V_i)$. We denote the *descendants* of V_i – the set of all vertices V_j such that there exists a directed path from V_i to V_j – as $\text{de}_{\mathcal{G}}(V_i)$. By convention, $\text{de}_{\mathcal{G}}(V_i)$ is defined to include V_i itself.

When $\mathcal{G}$ is *complete* – all vertices are pairwise connected via a directed edge – $\mathcal{M}^{\mathcal{G}}$ imposes no restrictions on p_V. When $\mathcal{G}$ is not complete, it is informative to compare the DAG factorization of p_V to the chain rule factorization to understand how missing edges entail restrictions on the observed distribution. Consider any valid topological ordering $\prec_{\mathcal{G}}$ of the variables – an ordering satisfying the property that whenever $V_i \prec_{\mathcal{G}} V_j$, V_i is not a descendant of V_j in $\mathcal{G}$. Then for every variable $V_i \in V$, define $\text{past}_{\mathcal{G}}(V_i)$ to be the set of vertices earlier than V_i under $\prec_{\mathcal{G}}$ (we suppress explicit reference to $\prec_{\mathcal{G}}$ to avoid notational clutter.) Under any variable ordering $\prec_{\mathcal{G}}$, we have the following equality between the chain factorization and DAG factorization of $p(v)$: $p(v) = \prod_{v_i \in v} p(v_i \mid \text{past}_{\mathcal{G}}(v_i)) = \prod_{v_i \in v} p(v_i \mid \text{pa}_{\mathcal{G}}(v_i))$.

Whenever $\text{pa}_{\mathcal{G}}(V_i) \subset \text{past}_{\mathcal{G}}(V_i)$ (corresponding to missing edges in $\mathcal{G}$) the above equality implies that $(V_i \perp\!\!\!\perp \text{past}_{\mathcal{G}}(V_i) \setminus \text{pa}_{\mathcal{G}}(V_i) \mid \text{pa}_{\mathcal{G}}(V_i))$. That is, given a DAG $\mathcal{G}$, restrictions in p_V are characterized by the following ordered local Markov property: each variable V_i is independent of its non-parental past given its parents. All restrictions entailed by a DAG $\mathcal{G}$ are easily read via the d-separation criterion (Pearl, 2009).

2.2 Causal DAG Models

In addition to a statistical DAG model $\mathcal{M}^{\mathcal{G}}$ representing restrictions on the factual variables V, it is possible to define a causal DAG model associated with $\mathcal{G}$. Causal DAG models can be generalized from statistical DAG models by equipping them with a special subset $A^{\dagger} \subseteq V$ referred to as treatment or action variables, with non-action variables $V \setminus A^{\dagger}$ denoted as Y.

Although a number of different such causal DAG models have been proposed (Robins, 1986; Pearl, 2009) they all satisfy the properties below.

(i) **counterfactual existence.** For each factual variable V_i represented on the

DAG, there exists a set of counterfactual random variables $V_i^{(a^\dagger)}$ representing the behavior of V_i when, possibly contrary to fact, $A^\dagger$ is set to $a^\dagger$ (for any such value) by external manipulation or intervention.

(ii) **causal ordering.** There exists a causal total ordering $\prec$ which must be topological with respect to $\mathcal{G}$. We index elements in any set of variables $X \subseteq V$ via subscript indices that are consistent with this ordering. As an example, X_i and X_j are both elements of $X \subseteq V$, and if $i < j$, then $X_i \prec X_j$. Given any variable X_i in $X \subseteq V$, define $\overline{X}_i$ to be the set consisting of X_i and all elements in X earlier than X_i in the ordering $\prec$. Similarly, define $\underline{X}_i$ to be the set consisting of X_i and all elements in X later than X_i in the ordering $\prec$. We will extend this notation in the natural way to values as well, e.g. $\overline{x}_i$ are values of $\overline{X}_i$. Finally, for every $V_i \in V$, define $\text{past}_\prec(V_i)$ to be all variables in V earlier than V_i in the ordering.

(iii) **no backwards causation.** For every $V_i^{(a^\dagger)}$, we have $V_i^{(a^\dagger)} \equiv V_i^{(\overline{a}_j^\dagger, \underline{a}_{j+1}^\dagger)} = V_i^{(\overline{a}_j^\dagger, \underline{a}_{j+1}^{\dagger\prime})} \equiv V_i^{(\overline{a}_j^\dagger)}$, where $\underline{a}_{j+1}^{\dagger\prime}$ and $\underline{a}_{j+1}^{\dagger\prime}$ correspond to distinct manipulations of variables in $\underline{A}_{j+1}^\dagger$, and $A_j^\dagger$ is the $\prec$-largest element of $A^\dagger$ that occurs prior to V_i according to $\prec$. That is, if $V_i = A_k$, $A_j^\dagger = A_{k-1}$, and if $V_i = Y_k$, $A_j^\dagger = A_k$.

(iv) **recursive substitution.** Given any subset $A \subset A^\dagger$, and any set of values a of A, we recursively define $V_i^{(a)}$ to be equal to $V_i^{\left(a, \{A_j(a) : A_j \in A^\dagger \setminus A\}\right)}$. In words, we define $V_i^{(a)}$ in terms of the set of $V_i^{(a^\dagger)}$ for all values $a^\dagger$ as the response V_i had A been manipulated to values a and each element A_j in $A^\dagger \setminus A$ is manipulated to whatever value it would have had under the intervention that sets A to a. Note that this definition is recursive since $A_j^{(a)}$ must also in general be defined. Properties (ii) and (iii) ensure the induction of the definition is always well-defined in the sense that every counterfactual $V_i^{(a)}$ is definable in terms of counterfactuals assumed to exist by (i).

(v) **consistency.** Every counterfactual variable $V_i^{(a)}$ is linked to the corresponding factual variable V_i by the consistency property, which states that $V_i = \sum_a \mathbb{I}(A = a) \times V_i^{(a)} = \sum_{\overline{a}_j} \mathbb{I}(\overline{A}_j = \overline{a}_j) \times V_i^{(\overline{a}_j)}$; equivalently, $V_i = V_i^{(a)}$ if $A = a$. Thus, in any equation or probability expression where v_i and $v_i^{(a)}$ appear and $A = a$ holds, we implicitly assume $v_i = v_i^{(a)}$. This allows us to assert the equality of $p(v_i^{(a)} \mid a) = p(v_i \mid a)$ as always true.

(vi) **positivity.** For every $A_i^\dagger \in A^\dagger$, $p(a_i^\dagger \mid \text{past}(A_i^\dagger)) > 0$ for all values of variables in $\text{past}_\prec(A_i^\dagger)$.

(vii) **sequential ignorability.** For every treatment $A_k \in A \subseteq A^\dagger$ and the $\prec$-earliest variable Y_k that occurs after A_k in the order $\prec$, we assume $\underline{Y}_k^{(a)} \perp\!\!\!\perp A_k \mid \text{Past}_\prec(A_k)$, for values of $\text{Past}_\prec(A_k)$ that are consistent with the treatment assignment a.

(viii) **Markov property.** The joint $p(v)$ is Markov relative to $\mathcal{G}$.

Assumptions (i),(ii), and (iv) allows us to meaningfully discuss counterfactuals $Y^{(a)}$. Furthermore, under the above causal model, the distribution $p(y^{(a)})$ – the joint distribution over counterfactuals obtained by setting a subset A of the action variables to a set of values a – can be expressed as a function of the joint distribution over factual variables as follows. Assumptions (v) and (iv), and our notational convention together imply $p(y^{(a)}) \times p(a \mid y^{(a)}) = p(y^{(a)}, a) = p(y, a)$, so

$$p(y^{(a)}) = \frac{p(y, a)}{p(a \mid y^{(a)})}. \qquad \textit{(counterfactual g-formula)} \tag{2.1}$$

We refer to (2.1) as the counterfactual g-formula, since counterfactual variables appear in the denominator.

It follows that $p(y^{(a)})$ is identified from the distribution of the factuals V if and only if $p(a \mid y^{(a)})$ is itself so identified. The additional assumptions (v),(vii),(vi), and

(viii) imply $p(a \mid y^{(a)})$ is indeed identified as follows:

$$p(a \mid y^{(a)}) = \prod_{a_k \in a} p\big(a_k | \overline{a}_{k-1}, \overline{y}_{k-1}^{(a)}, \underline{y}_k^{(a)}\big) = \prod_{a_k \in a} p\big(a_k | \text{past}_{\mathcal{G}}(a_k), \underline{y}_k^{(a)}\big)$$
$$= \prod_{a_k \in a} p\big(a_k | \text{past}_{\mathcal{G}}(a_k)\big) = \prod_{a_k \in a} p\big(a_k | \, \text{pa}_{\mathcal{G}}(a_k)\big)$$

The first equality follows from the chain rule of probability: for each conditional $p(a_k \mid .)$ we split $y^{(a)}$ into the sets $\overline{y}_{k-1}^{(a)}$ and $\underline{y}_k^{(a)}$. The second equality follows from the consistency assumption (v): $\overline{y}_{k-1}^{(a)} = \overline{y}_{k-1}^{(\overline{a}_{k-1})} = \overline{y}_{k-1}$ and noting $\text{past}_{\mathcal{G}}(a_k) = \big(\overline{y}_{k-1}, \overline{a}_{k-1}\big)$. The third equality follows from (vii). The fourth from (viii). Finally the positivity assumption (vi) implies the RHS of the last equality is a unique function of the distribution of V. We conclude that $p(y^{(a)}) = \frac{p(y,a)}{p(a|y^{(a)})}$is identified by the g-formula $p_{\mathcal{G}}(y \parallel a)$ (Lauritzen, 1996) equal to

$$\frac{\prod\limits_{v_k \in y \cup a} p(v_k \mid \text{past}_{\mathcal{G}}(v_k))}{\prod\limits_{a_k \in a} p\big(a_k | \, \text{past}_{\mathcal{G}}(a_k)\big)} = \frac{\prod\limits_{v_k \in y \cup a} p(v_k \mid \text{pa}_{\mathcal{G}}(v_k))}{\prod\limits_{a_k \in a} p\big(a_k | \, \text{pa}_{\mathcal{G}}(a_k)\big)} = \prod_{y_k \in y} p\left(y_k \mid \text{pa}_{\mathcal{G}}(y_k)\right)\Big|_{A=a} \tag{2.2}$$

and $p(.)|_{A=a}$ is taken to mean that any free variables within a probability expression $p(.)$ that intersect the set A are to be evaluated at corresponding values of a. The final equality follows from cancellation of terms. Note by the notational convention in (ii), it follows that $p_{\mathcal{G}}(y^{(a)} \parallel a) = p_{\mathcal{G}}(y \parallel a)$

The g-formula $p_{\mathcal{G}}(y \parallel a)$ in (2.2) can be viewed as a truncated factorization in the sense that the terms $p(a_k \mid \text{pa}_{\mathcal{G}}(a_k))$ that occur in the Markov factorization of the density $p(y, a)$ are no longer present in the g-formula factorization representing the intervention distribution. As we will see in Section 4, the g-formula is also closely connected to identification in missing data problems.

The truncated factorization $p_{\mathcal{G}}(y \parallel a)$ is still a distribution as it provides a mapping from values a of A to normalized densities over variables in Y. We call objects of this type *kernels*. A kernel acts in most respects like a conditional distribution. In

particular, given a kernel $p(v \parallel w)$ and a subset $Z \subseteq V$, conditioning and marginalization are defined in the usual way as

$$p(z \parallel w) \coloneqq \sum_{v \setminus z} p(v \parallel w) \quad \text{and} \quad p(v \setminus z \mid z \parallel w) \coloneqq \frac{p(v \parallel w)}{p(z \parallel w)}. \tag{2.3}$$

Property (vii) above is the critical identifying assumption: In words, for each k, there exists a factual past $(\overline{A}_{k-1}, \overline{Y}_{k-1}) \subset V$ such that treatment A_k is as if randomly assigned (and hence independent of future counterfactuals). In the language of epidemiologists, conditional on $(\overline{A}_{k-1}, \overline{Y}_{k-1})$, the causal effect of A_k on $\underline{Y}_k$ is unconfounded.

In most observational studies, some of the variables that need to be included in $\overline{Y}_{k-1}$ (and thus in V) to make A_k unconfounded may be unknown to the investigators and/or known but not measured for financial or logistical reasons. As a consequence $p(y^{(a)})$ will obviously not be identified from the the factual distribution $p\,(o, a)$ of the observed variables (O, A), where $O \subset Y$. Moreover, even the counterfactual distribution over observed outcomes,

$$\begin{aligned} p(o^{(a)}) &= \sum_{y^{(a)} \setminus o^{(a)}} p(y^{(a)}) = \sum_{y \setminus o} p_{\mathcal{G}}(y \parallel a) \equiv p_{\mathcal{G}}(o \parallel a) \\ &= \sum_{y \setminus o} \frac{\prod_{v_k} p(v_k \mid \text{past}_{\mathcal{G}}(v_k))}{\prod_{a_k \in a} p\big(a_k \mid \text{past}_{\mathcal{G}}(a_k)\big)} = \sum_{y \setminus o} \frac{\prod_{v_k} p(v_k \mid \text{pa}_{\mathcal{G}}(v_k))}{\prod_{a_k \in a} p\big(a_k \mid \text{pa}_{\mathcal{G}}(a_k)\big)} \end{aligned}$$

may not be identified from $p\,(o, a)$. This is because, in general, property (vii) (sequential ignorability) does not imply

$$\underline{O}_k^{(a)} \perp\!\!\!\perp A_k \mid \text{past-obv}_{\mathcal{G}}(A_k) \text{ for all } k,$$

where $\text{past-obv}_{\mathcal{G}}(A_k) \equiv (\overline{A}_{k-1}, \overline{O}_{k-1})$ are the elements in $\text{past}_{\mathcal{G}}(A_k)$ that are observed.

As an example, consider the DAG in Fig. 1(a). The statistical model $\mathcal{M}^{\mathcal{G}}$ of this DAG is the set of distributions $p(v) = p(u_1, u_2, r_1, r_2, l_1, l_2)$ that factorize as:

(a) (b)

Figure 1: (a) A DAG where U_1 and U_2 may be unmeasured; (b) A conditional DAG illustrating interventions on R_1 and R_2.

$p(u_1)\times p(u_2 \mid u_1)\times p(r_1 \mid u_2)\times p(l_1 \mid u_1, r_1)\times p(r_2 \mid r_1, l_1)\times p(l_2 \mid u_2, r_2)$. Suppose we choose the treatment set $A^\dagger = \{R_1, R_2\}$ and the outcome set $Y := \{U_1, U_2, L_1, L_2\}$. Under the causal DAG model corresponding to the DAG in Fig. 1(a), the counterfactual distribution $p(u_1^{(r_1,r_2)}, u_2^{(r_1,r_2)}, l_1^{(r_1,r_2)}, l_2^{(r_1,r_2)})$ is identified via the g-formula as:

$$p_{\mathcal{G}}(y \parallel r_1, r_2) = p(u_1, u_2, r_1, l_1, r_2, l_2)/\{p(r_2 \mid r_1, l_1) \times p(r_1 \mid u_2)\}. \tag{2.4}$$

Similar to how the DAG factorization is the factorized representation of a factual density Markov relative to a given causal DAG, the g-formula $p_{\mathcal{G}}(y \parallel a)$, is the factorized representation of a counterfactual density Markov relative to a truncated version of the causal DAG, a.k.a conditional causal DAG, in which the edges pointing into the treatment variables A that are intervened upon are removed. The conditional causal DAG corresponding to (2.4) is shown in Fig. 1(b).

Suppose now U_1, U_2 are unobserved so $O = (L_1, L_2)$. The distribution of interest $p\left(o^{(r_1,r_2)}\right) \equiv p(l_1^{(r_1,r_2)}, l_2^{(r_1,r_2)}) := p_{\mathcal{G}}(l_1, l_2 \parallel r_1, r_2) = \sum_{u_1,u_2} p_{\mathcal{G}}(l_1, l_2 \parallel r_1, r_2)$ is not a function of the observed data distribution $\sum_{u_1,u_2} p(u_1, u_2, r_1, r_2, l_1, l_2)$ (Shpitser and Pearl, 2006). Intuitively, this is because the g-formula $p(l_1, l_2 \parallel r_1, r_2)$ (moving forward we drop the subscript $\mathcal{G}$, when it is clear from the context what graph we are referring to) contains the term $p(r_1 \mid u_2)$ in the denominator, and there is no information on

U_2 in the observed data.

As we will see, in missing data problems every variable L_k is either always observed, or (if it is missing) corresponds to a single counterfactual version $L_k^{(1)}$ – the value of L_k had we in fact been able to observe it. This allows additional identification theory to be developed specifically for missing data problems that have no analogue in causal inference identification theory. We now formalize missing data as a causal inference problem before discussing identification in graphical models of missing data.

3. Missing Data Models As Causal Models

Inferring a parameter of interest in the presence of missing data often involves posing a statistical model that encodes a set of assumptions on the missingness mechanisms. Let $Z = (Z_1, \ldots, Z_K)^T$ be a vector of K random variables with finite support and probability density p_Z. Given a sample of the random vector Z, let $R = (R_1, \ldots, R_K)^T$ be the associated vector of binary missingness indicators with $R_k = 1$ if Z_k is observed and $R_k = 0$ if Z_k is missing. Denote the conditional distribution of R given Z by $p_{R|Z}$ and the joint distribution of Z and R by $p_{(R,Z)}$, and assume p_Z, $p_{R|Z}$, and $p_{(R,Z)}$ are contained in models $\mathcal{M}_Z$, $\mathcal{M}_{R|Z}$, and $\mathcal{M} = \mathcal{M}_Z \otimes \mathcal{M}_{R|Z}$, respectively. $\mathcal{M}$ is referred to as a selection model in the missing data literature. The observed data is often denoted by $O = (R, Z_{\text{obs}})$, where Z_{obs} is the subvector of Z corresponding to the subvector of R with components 1, i.e., $Z_{\text{obs}} \coloneqq \{Z_k \in Z \text{ s.t. } R_k = 1\}$. If $\mathcal{M}$ imposes no restrictions on the observed data distribution p_O, then it is called a *nonparametric saturated* model (Robins, 1997).

We can redefine the above using the terminology of causal models described in Section 2. We may view each missingness indicator $R_k \in R$ as a treatment variable that can be intervened on. Each $Z_k \in Z$ can then be interpreted counterfac-

tually – a random variable had we, possibly contrary to fact, intervened and set the corresponding missingness indicator R_k to 1. By analogy with causal models, from here onward, we refer to Z_k as $L_k^{(r_k=1)}$ to highlight the counterfactual nature of the variable. This notation explicitly encodes, in counterfactual language, the assumption implicit in classical missing data models, that the value of Z_k remains the same regardless of whether any other $Z_j \in Z$ is observed or missing (or equivalently R_j is 1 or 0). This assumption is closely related to the "lack of interference" assumption in causal inference. We collect all these counterfactual variables into a vector $L^{(r=1)} \coloneqq (L_1^{(r_1=1)}, \ldots, L_K^{(r_K=1)})^T$, and simplify the notation for $L_k^{(r_k=1)}, L^{(r=1)}$ via $L_k^{(1)}, L^{(1)}$.

The link between Z and Z_{obs} can be viewed as the link between the counterfactual variables $L^{(1)}$ that are of substantive interest, treatment variables R, and factual variables L (a.k.a. proxies) that we observe. Specifically, for any $L_k^{(1)} \in L^{(1)}$, its corresponding proxy $L_k \in L$ is *deterministically* defined as a function of $L_k^{(1)}$ and R_k as follows: $L_k = L_k^{(1)}$ if $R_k = 1$ and $L_k =$ "?" if $R_k = 0$. This link is closely related to the *consistency* assumption in causal inference, described above.

The state space of any $L_k \in L$ is equal to the state space of the corresponding $L_k^{(1)}$ in $L^{(1)}$ joined with the special value "?". Hence, we denote generic realizations of $L_k \in L$ and $L_k^{(1)} \in L^{(1)}$ as l_k and $l_k^{(1)}$, respectively. We denote realizations of $L^{(1)}$ and L by $l^{(1)}$ and l, respectively. By consistency, it is always true that $p(L_k^{(1)} \mid R_k)|_{R_k=1} = p(L_k \mid R_k)|_{R_k=1}$, however $p(l_k^{(1)} \mid r_k)|_{r_k=1} = p(l_k \mid r_k)|_{r_k=1}$ is true if and only if $l_k^{(1)} = l_k$. Our convention is that in any equation or probability expression where $l_k^{(1)}$ and l_k appear and $r_k = 1$, $l_k^{(1)} = l_k$. Thus $p(l_k^{(1)} \mid r_k)|_{r_k=1} = p(l_k \mid r_k)|_{r_k=1}$ becomes always true. In this way, we can evaluate probability expressions at realizations rather than at the corresponding random variables while still imposing consistency.

We redefine Z_{obs} via $L = L^{(R)} = L^{(r)}|_{R=r}$. With this new notation, the observed

data then changes from $O = (R, Z_{\text{obs}})$ to $O = (R, L)$. Note, however, that the state space of L is formed by augmenting the state space of Z with a special value "?". A common goal in missing data problems is to determine whether the joint distribution of the complete data $L^{(1)} \coloneqq Z$, that is $p(l^{(1)}) \coloneqq p(z) \in \mathcal{M}_{L^{(1)}} \coloneqq \mathcal{M}_Z$, is identified from the observed data $O = (R, L) \coloneqq (R, Z_{\text{obs}})$, in a model $\mathcal{M} = \mathcal{M}_{R|L^{(1)}} \otimes \mathcal{M}_{L^{(1)}}$ defined over the joint distribution of $\left(R, L^{(1)}\right) \coloneqq (R, Z)$. When discussing identification in missing data problems below, we will refer to $p(l^{(1)})$ as the *target law*, $p(r \mid l^{(1)})$ as the *missingness mechanism*, and $p(l^{(1)}, r)$ as the *full law*. These distributions may also be extended with a set of auxiliary variables W that are always observed.

The model $M = M_{R|L^{(1)}} \otimes M_{L^{(1)}}$ is said to be nonparametric just identified if: (i) $M_{L^{(1)}} = M^{\text{np}}_{L^{(1)}}$, where $M^{\text{np}}_{L^{(1)}}$ is the set of all distributions of $L^{(1)}$; (ii) the distribution of the observed data $O = (L, R)$ is unrestricted except for positivity and $L_k =$ "?" if and only if $R_k = 0$; (iii) $p\left(l^{(1)}\right)$ is identified from $p\left(L, R\right)$.

A *necessary and sufficient* condition for target law identification is that for all $p(r \mid l^{(1)}) \in \mathcal{M}_{R|L^{(1)}}$, $p(R = 1 \mid l^{(1)}) > 0$ wp1 and $p(R = 1 \mid l^{(1)})$ is identified from $p(o)$. This follows from the fact that $p\left(R = 1, l^{(1)}\right) = p\left(R = 1, l\right)$ by consistency (which is a function of $p(o)$) and an application of chain rule: $p(l^{(1)}) = p(l, R = 1)/p(R = 1 \mid l^{(1)})$. Note this expression is exactly the counterfactual g-formula (2.1) of the previous section with $a = r$ and $r = 1$.

A *necessary and sufficient* condition for full law identification is that for all $p(r \mid l^{(1)}) \in \mathcal{M}_{R|L^{(1)}}$, $p(R = 1 \mid l^{(1)}) > 0$ wp1 and $p(R = r \mid l^{(1)})$ is identified from $p(o)$, for any missingness pattern $r \in \{0, 1\}^K$. This follows from the chain rule of probability, $p(l^{(1)}, R = r) = \left\{p(l, R = 1)/p(R = 1 \mid l^{(1)})\right\} \times p(R = r \mid l^{(1)})$.

Unless explicitly stated otherwise, we assume $\mathcal{M}_{R|L^{(1)}}$ encompasses positive distributions, meaning that for every $p(R = r \mid l^{(1)}) \in \mathcal{M}_{R|L^{(1)}}$, $p(R = r \mid l^{(1)}) >$

0 wp1 for all $r \in \{0,1\}^K$. This assumption may be extended to pattern restrictions such as monotonicity, where $p(R = r \mid l^{(1)}) > 0$ for any pattern of values of R where, under some ordering of missingness indicators $R_1, \ldots, R_K$, if $R_k = 0$ then $R_{k+1} = 0$ wp1 for every $k \in \{1, \ldots, K-1\}$.

A general procedure for analyzing MNAR models proceeds by imposing a set of restrictions on the full data distribution (the target distribution and its missingness mechanism) that are sufficient to yield identification of the parameter of interest. In many models, the restrictions may be represented by a factorization of the full data law with respect to a DAG. Our objective is to identify the target law $p(l^{(1)})$ from factual data on variables (R, L) where the missing data model is represented via a DAG with potentially hidden variables. As we will see, restricting attention to such missing data models allows ideas inspired by causal identification theory to be brought to bear. We now describe how DAGs are used to encode independence restrictions of a missing data model.

4. Missing Data DAG Models

Unlike a causal DAG, a missing data DAG (or m-DAG for short), includes counterfactual variables on the graph. The only well-defined counterfactual variables in missing data problems are ones representing variables L had they been observed, possibly contrary to fact. The superscript in $L^{(1)}$ thus corresponds to an intervention setting indicators R to 1.

An m-DAG $\mathcal{G}_m$ consists of a set of vertices V associated with variables in $\{L^{(1)}, R, L\}$. In addition to acyclicity, an m-DAG restricts the presence of certain edges in the following ways:

(a) $\text{pa}_{\mathcal{G}_m}(L_k) = \{L_k^{(1)}, R_k\}$. That is, each proxy variable $L_k \in L$ has only two parents in $\mathcal{G}_m$ to represent the deterministic function that defines L_k in terms of $L_k^{(1)}$ and

R_k. These edges are drawn in gray in all m-DAGs in this manuscript in order to distinguish deterministic relations from probabilistic ones.

(b) $L^{(1)} \cap \{\mathrm{de}_{\mathcal{G}_m}(R) \cup \mathrm{de}_{\mathcal{G}_m}(L)\} = \emptyset$. That is, variables in R and L cannot have directed paths to variables in $L^{(1)}$. In the presence of fully observed auxiliary variables W, we also assume $W \cap \{\mathrm{de}_{\mathcal{G}_m}(R) \cup \mathrm{de}_{\mathcal{G}_m}(L)\} = \emptyset$.

A special case of missing data DAG models where $\mathrm{ch}_{\mathcal{G}_m}(L_i) = \emptyset$ for every L_i was considered in (Mohan et al., 2013).

Restriction (a) is imposed by definition. Since every L_k is a deterministic function of $L_k^{(1)}$ and R_k, only those two variables can serve as causes of L_k, and thus as parents of L_k in the graph. Restriction (b) is imposed to ensure observed treatment variables cannot influence counterfactual variables, which correspond to outcomes after treatments have been intervened on. This restriction is a consequence of the missing data version of consistency, which implies observed variables are caused by their corresponding counterfactuals, and not vice versa.

The above restrictions imply that $\mathrm{ch}_{\mathcal{G}_m}(L_k) \subseteq \{R_i \in R \mid i \neq k\}$ and $\mathrm{ch}_{\mathcal{G}_m}(R_k) \subseteq \{L_k\} \cup \{R_i \in R \mid i \neq k\}$.

The missing data model of an m-DAG $\mathcal{G}_m$, denoted by $\mathcal{M}^{\mathcal{G}_m}$, consists of a set of joint distributions $p(l, r, l^{(1)})$ that factorize with respect to an m-DAG $\mathcal{G}_m$ with vertex set $\{L, R, L^{(1)}\}$ as follows

$$\prod_{v_i \in l \cup r \cup l^{(1)}} p(v_i \mid \mathrm{pa}_{\mathcal{G}_m}(v_i)) = \prod_{l_k \in l} p(l_k \mid l_k^{(1)}, r_k) \times \prod_{v_i \in r \cup l^{(1)}} p(v_i \mid \mathrm{pa}_{\mathcal{G}_m}(v_i)). \tag{4.5}$$

The terms $p(l_k \mid l_k^{(1)}, r_k)$ are deterministically defined: $p(L_k = l_k \mid L_k^{(1)} = l_k^{(1)}, R_k = 1) = 1$ and $p(L_k = \text{“?”} \mid L_k^{(1)} = l_k^{(1)}, R_k = 0) = p(L_k = \text{“?”} \mid R_k = 0) = 1$ for any $L_k \in L$. The joint distribution $p(l, r, l^{(1)})$ further satisfies the positivity assumption stated as follows: $p(R_k = r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k)) > 0$ wp1 for all $R_k \in R$. This assumption excludes monotone missingness.

We can view missing data models associated with m-DAGs as a special case of a causal model of a DAG (with hidden variables) described in the previous section where Y is taken to be L, and $A^\dagger$ is taken to be R, with a set of additional restrictions. Specifically, every variable L_k that is potentially missing corresponds to *one* non-trivial counterfactual $L_k^{(1)}$, with the other counterfactual $L_k^{(0)}$ trivially defined as "?." In addition, every treatment variable R_k in R can only affect exactly one outcome variable, namely L_k.

Since missing data models associated with m-DAGs are a special case of causal models associated with DAGs, the following results immediately follows.

Proposition 1. *Under the missingness model of an m-DAG* $\mathcal{G}_m$

$$p(l^{(1)}) = \prod_{v_k \in l^{(1)}} p(v_k \mid \text{pa}_{\mathcal{G}_m}(v_k)) = \left. \frac{p(l,r)}{\prod_{r_k \in r} p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))} \right|_{r=1}. \tag{4.6}$$

The second equality in (4.6) is the missing data DAG equivalent of the counterfactual g-formula in (2.1) for causal inference problems. In missing data DAG models, any counterfactual variable $L_k^{(1)}$ is allowed to have elements of R as children. This means that the g-formula in (4.6) does not necessarily lead to identification of $p(l^{(1)})$ in terms of the observed data distribution $p(l,r)$. This is because $\text{pa}_{\mathcal{G}_m}(r_k)$ in $p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ may involve values in $l^{(1)}$ which are not always observed and cannot be immediately dropped via independence assumptions. In the next section, we illustrate via a number of examples how identification may nevertheless be accomplished in some missing data DAG models representing MNAR mechanisms.

Similar to Rubin's hierarchy of missingness mechanisms, it is possible to set up a hierarchy for missing data DAG models that define the complexity of identification techniques required. The missing data DAG model for a graph $\mathcal{G}_m$ with vertices $\{L^{(1)}, R, L\}$ is said to be

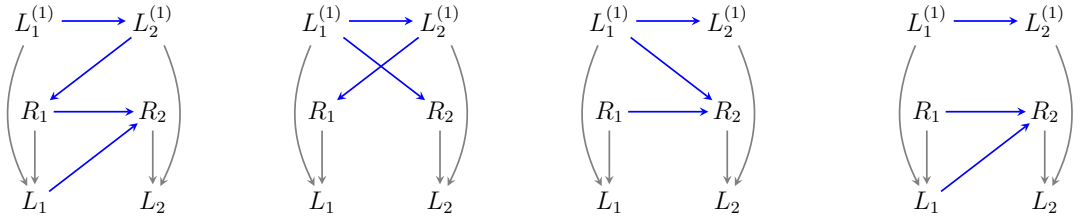
- Missing Completely At Random (MCAR) if $\prod_{r_k \in r} p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ is not a

function of variables in $L^{(1)}$ and L. Graphically speaking there are no edges that point to variables in R.

- Missing At Random (MAR) if $\prod_{r_k \in r} p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ is not a function of variables in $L^{(1)}$. Graphically speaking there are no edges from variables in $L^{(1)}$ to variables in R. Note that this model is a submodel of the MAR model, described below, with the additional restriction that the full law factorizes with respect to an m-DAG.
- Missing Not At Random (MNAR) otherwise.

A MAR missingness mechanism, according to Rubin's definition, requires that for any given missingness pattern, the missingness is independent of the missing values given the observed values. These types of restrictions are with respect to missingness patterns $R = r$ (a total of 2^K distinct patterns for models with K missing variables). This missingness mechanism cannot be represented via graphs, and several authors have noted the difficulty in interpretation of MAR models in practice (Schafer and Graham, 2002; McKnight et al., 2007; Graham, 2012; Tian, 2015). The above graphical hierarchy provides a more intuitive description of the above types of missingness mechanisms, in the sense that any missing data model associated with an m-DAG comes equipped with a ready description of a data generating process, where variables are generated sequentially according to a total order consistent with the m-DAG.

Further, while there exist MNAR models whose restrictions cannot be represented graphically, for instance the complete-case missing value model (Little, 1993), or the discrete choice model (Tchetgen et al., 2018), the restrictions posed in several popular MNAR models, such as the permutation model (Robins, 1997), the block-sequential MAR model (Zhou et al., 2010), and those in Daniel et al. (2012); Thoemmes and Rose (2013); Martel García (2013); Mohan et al. (2013); Shpitser (2016a); Saadati



(a) Permutation (b) Block-parallel (c) Block-sequential (d) MAR example

Figure 2: Examples of missing data DAG models.

and Tian (2019); Bhattacharya et al. (2019) and Nabi et al. (2020) correspond to DAGs. Models described in (Shpitser, 2016b; Sadinle and Reiter, 2017; Malinsky et al., 2021) correspond to graphical models that generalize DAG models, and are instead associated with chain graphs (Lauritzen, 1996).

We now describe some missing data models, focusing on those with restrictions that can be defined via m-DAGs, in more detail.

Permutation model. Given an ordering (permutation) indexed by $k \in \{1, \ldots, K\}$ on variables in $L^{(1)}$, Robins (1997) defined a *permutation missingness* model by the restrictions that R_k is independent of the current and past $L^{(1)}$ given the observed past and future $L^{(1)}$ variables. We denote this model by $\mathcal{M}^{\text{per}}$. Formally, $\mathcal{M}^{\text{per}}$ is defined by the following conditional independence restrictions: $R_k \perp\!\!\!\perp \overline{L}_k^{(1)} \mid \overline{R}_{k-1}, \overline{L}_{k-1}, \underline{L}_{k+1}^{(1)}$ for all $k \in \{1, \ldots, K\}$. The graphical representation of $\mathcal{M}^{\text{per}}$ for two time points defined by these set of independencies is shown in Fig. 2(a). The local Markov property for this DAG model yields the following set of independence restrictions: $R_1 \perp\!\!\!\perp L_1^{(1)} \mid L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)}, L_1^{(1)} \mid L_1, R_1$.

Block-parallel MNAR model. The block-parallel MNAR model, denoted $\mathcal{M}^{\text{b-par}}$, was introduced in Mohan et al. (2013). It is defined by the following conditional independence restrictions: $R_k \perp\!\!\!\perp \{L_k^{(1)}, R_{-k}\} \mid L_{-k}^{(1)}$, for all $k \in \{1, \ldots, K\}$,

where $V_{-k} \coloneqq V \setminus V_k$. The reason for our use of this name should become apparent in Section 5.2 when we discuss the identification of this missing data model. The graphical representation of $\mathcal{M}^{\text{b-par}}$ for two time points, defined by this set of independencies, is shown in Fig. 2(b). The local Markov property for this DAG model implies: $R_1 \perp\!\!\!\perp L_1^{(1)}, R_2 \mid L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)}, R_1 \mid L_1^{(1)}$.

Block-sequential MNAR model. The block-sequential MNAR model, denoted by $\mathcal{M}^{\text{b-seq}}$, was introduced in Zhou et al. (2010), under the name "block-conditional MAR." We use our alternative name to make it clear that this model is not MAR but MNAR. It is defined by the following conditional independence restrictions: $R_{k+1} \perp\!\!\!\perp \underline{L}_{k+1}^{(1)} \mid \overline{R}_k, \overline{L}_k^{(1)}, \quad \forall k \in \{1, \ldots, K\}$. The graphical representation of $\mathcal{M}^{\text{b-seq}}$ for two time points, defined by this set of independencies, is shown in Fig. 2(c). The local Markov property for this DAG model implies: $R_1 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, L_1^{(1)}$.

An example of a MAR model. The model introduced in Rubin (1976) is defined via the following conditional independence restrictions:

$$\mathbb{I}(R = r) \perp\!\!\!\perp \left\{ L_i^{(1)};\ r_i = 0 \right\} \ \Big|\ \left\{ L_j^{(1)};\ r_j = 1 \right\} \quad \text{for all } r \in \{0,1\}^K,$$

where $\mathbb{I}(.)$ is the indicator function. This MAR model cannot be represented by a DAG $\mathcal{G}_m$ with vertices $V = \{L^{(1)}, R, L\}$. This follows because the 2^K variables $\mathbb{I}(R = r)$ are not vertices on the graph. However, we can have more intuitive submodels of the MAR model that can be represented graphically. The missing data DAG for two time points in Fig. 2(d) is one example. The local Markov property for this DAG model implies: $R_1 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, L_1$.

An interesting observation is that under the monotonicity assumption, the MAR model in Fig. 2(d) and the block-sequential MNAR model in Fig. 2(c) are identical. The monotonicity assumption in the block-sequential MNAR model imposes restrictions on the univariate conditionals of each R_k given their parents on the graph:

evaluating any of the parental missingness indicators at zero deterministically defines the conditional density of $p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$. This is because if $i \prec_{\mathcal{G}_m} k$, and $R_i = 0$ then it must be the case that $R_k = 0$, otherwise the monotonicity assumption is violated. The only non-deterministic evaluation of the univariate conditionals occurs when $R_i = 1, \forall R_i \in \text{pa}_{\mathcal{G}_m}(R_k)$. Thus, by consistency, we can replace the pair $(R_i = 1, L_i^{(1)}) \in \text{pa}_{\mathcal{G}_m}(R_k)$ with $(R_i = 1, L_i)$. The equivalent graphical operation is replacing the set of edges $\{L_i^{(1)} \rightarrow R_j \leftarrow R_i\}$ with the set $\{L_i \rightarrow R_j \leftarrow R_i\}$. This renders the DAG in Fig. 2(c) and the one in Fig. 2(d) equivalent. The monotonicity assumption is a restriction on patterns, and cannot be directly represented in an m-DAG, unless it is augmented with additional edge markings denoting deterministic relationships among R variables that define monotonicity.

5. Identification in Missing Data DAG Models

Given a missing data DAG model, our objective is to determine whether the target law $p(l^{(1)})$, or a fixed function $h(l^{(1)})$, can be identified as a function of the observed data law $p\,(r, l)$, and if so, find the identifying functional. To aid subsequent developments we will reformulate the identification problem in missing data models using the language common in causal inference. In this view, given an underlying variable $L_i^{(1)}$, the corresponding proxy variable L_i is viewed as an observed version of "the outcome," and the corresponding missing indicator R_i as the observed version of "the treatment." In this view, the distribution $p(l^{(1)})$ may be viewed instead as $p(l \parallel r = 1)$ – the distribution where all missingness indicators in R are intervened on and set to 1. By consistency in missing data, we have the following equalities $p(l^{(1)}) = p(l \parallel r = 1) = p(l^{(1)} \parallel r = 1) = p(l^{(1)}, l \parallel r = 1)$.

Recall from Proposition (1) that identification of the target law $p(l^{(1)}) \coloneqq p(l \parallel r = 1)$ is equivalent to identification of the missingness mechanism evaluated at $R = 1$, i.e.,

the probability of observing the complete cases missing data pattern: $\prod_{R_i \in R} p(R_i = 1 \mid \text{pa}_{\mathcal{G}_m}(r_i))|_{R=1}$. If each conditional factor $p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))$ (evaluated at $R = 1$) is identified in this product, then the complete cases missing data pattern, and consequently the target law would be identified. We refer to the conditional factor $p(R_k = 1 \mid \text{pa}_{\mathcal{G}_m}(r_k))$ as the *propensity score* for R_k. Via a series of examples, we explore different identification strategies for identifying the distribution of the missingness mechanism $p(r \mid \text{pa}_{\mathcal{G}_m}(r))$ evaluated at the complete cases pattern values: $R = 1$.

A Single Variable Interventional Reformulation of the Counterfactual G-formula. Given a joint distribution $p(l^{(1)}, r, l)$ that factorizes according to a DAG $\mathcal{G}_m$, the intervention on $R_k \in R$ is defined via a truncated factorization where the joint distribution is divided by the identified propensity score of R_k. That is:

$$p(l^{(1)} \setminus l_k^{(1)}, r \setminus r_k, l \parallel r_k = 1) := \left. \frac{p(l^{(1)} \setminus l_k^{(1)}, r, l)}{p(r_k \mid \text{pa}_{\mathcal{G}_m}(r_k))} \right|_{R_k=1}. \tag{5.7}$$

When we intervene on R_k and set it to 1, it becomes redundant to include both $L_k^{(1)}$ and L_k in the joint distribution as they represent the same random variable when $R_k = 1$; hence we drop $L_k^{(1)}$ from both sides of the equation.

The propensity score of R_k is identified if we can replace each $L_j^{(1)} \in \text{pa}_{\mathcal{G}_m}(R_k)$ with $\{L_j, R_j = 1\}$. Such replacements are sometimes justified due to conditional independence restrictions in the full data distribution. For instance in Fig. 2(b), the propensity of R_1 is identified because we have $p(r_1 \mid l_2^{(1)}) = p(r_1 \mid l_2^{(1)}, R_2 = 1) = p(r_1 \mid l_2, R_2 = 1)$, since $R_1 \perp\!\!\!\perp R_2 \mid L_2^{(1)}$. When missingness indicators are connected, we may lose some of these convenient independence constraints. For instance in Fig. 2(a), R_1 and R_2 are no longer conditionally independent; thus identification of $p(r_1 \mid l_2^{(1)})$ is not as straightforward as it is in Fig. 2(b). However, due to an *invariance* property of the propensity scores, we can sometimes succeed in identification by exploring

interventional distributions where a subset of observed variables are intervened on and consequently certain edges are removed. We formalize this property in the following lemma, which is analogous to the invariance property in causal inference.

Lemma 1 (Invariance property)**.** *Given the propensity score for $R_k \in R$, the conditioning set $\mathrm{pa}_{\mathcal{G}_m}(R_k)$ captures the direct causes of R_k and hence remains invariant under any set of interventions that disrupts other parts of the joint factorization. Given $R^* \subseteq R \setminus R_k$, we have*

$$p\left(r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k)\right) = p\left(r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k) \parallel r^* = 1\right). \tag{5.8}$$

Using this property, we now explore various strategies for target law identification in MNAR models. In Section 5.1, we use the permutation model as an example to illustrate how missingness mechanisms can sometimes be identified via a *sequence* of interventions on missingness indicators. This is analogous to causal inference techniques used in longitudinal studies to sequentially identify the effect of multiple treatments. An intervention in missing data, unlike interventions in causal inference, can sometimes induce *selection bias* due to conditioning on a subset of missingness indicators that have not yet been intervened on. Introduction of selection during intervention operations may make identification by means of sequential applications of the truncated factorization impossible. In Section 5.2, we illustrate the selection issue of sequential interventions on the block-parallel model, and show how *parallel* (simultaneous) interventions can obtain identification even if selection is present. In some missing data models, such as the block-sequential model, we can identify the missingness probability of the complete cases pattern via either sequential or parallel application of interventions. In Sections 5.3 through 5.5, we go over examples where a combination of sequential and parallel interventions are needed to identify each propensity score. In Section 5.6, we unify ideas explored in this section to yield

a general identification procedure for arbitrary missing data DAGs by exploring all possible *partial orders* of interventions defined on the observed variables on the graph. We also discuss some graphical structures in missing data DAG models that impede nonparametric identification of the model.

5.1 Sequential Interventions

Consider the permutation model $\mathcal{M}^{\text{per}}$ with two time points, redrawn in Fig. 3(a). Our objective is to identify $p(l_1^{(1)}, l_2^{(1)}) \coloneqq p(l_1, l_2 \parallel r_1 = 1, r_2 = 1)$ as a function of observed data. This can be done in two steps: we first intervene on R_2, then we intervene on R_1.

Step 1. Intervene on R_2 to get $p(. \parallel r_2 = 1)$. The propensity score of R_2, $p(r_2 \mid r_1, l_1)$, is a function of observed data and so the corresponding post intervention distribution is immediately identified. Intervening on R_2 and setting it to 1 yields the following kernel:

$$p(l_1^{(1)}, l_2, r_1, l_1 \parallel r_2 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid r_1, l_1)}. \tag{5.9}$$

By consistency in missing data, $p(l_2 \parallel r_2 = 1)$, $p(l_2^{(1)} \parallel r_2 = 1)$, and $p(l_2^{(1)}, l_2 \parallel r_2 = 1)$ all represent the same object. This kernel factorizes with respect to the (conditional) m-DAG shown in Fig. 3(b), obtained from Fig. 3(a) by removing edges $R_1 \to R_2$ and $L_1 \to R_2$, and denoting the vertex R_2 as a square showing $r_2 = 1$, indicating the intervention setting R_2 to 1.

In general, the graphical analogue of the intervention on R_k entails removing all edges with arrowheads into R_k in the corresponding missing data graph. We denote vertices corresponding to variables that have been intervened on with rectangles. Further, the pair $(L_k^{(1)}, L_k)$ is treated as a single variable on the graph after intervening and setting $R_k = 1$ due to the deterministic relation of L_k with $L_k^{(1)}$ and R_k. We keep

the proxy variable on the graph but in gray with dashed edges.

Step 2. Intervene on R_1 after intervening on R_2 to get $p(. \parallel r_1 = 1, r_2 = 1)$. In the second step, we want to intervene on R_1 in the kernel $p(. \parallel r_2 = 1)$ which is Markov relative to the DAG in Fig. 3(b). In order to perform this intervention, we need to show that $p(r_1 \mid l_2^{(1)} \parallel r_2 = 1)$ is a function of $p(r, l)$. Using consistency, we have $p(r_1 \mid l_2^{(1)} \parallel r_2 = 1) = p(r_1 \mid l_2 \parallel r_2 = 1)$, and using kernel probability rules (provided in 2.3), we have $p(r_1 \mid l_2 \parallel r_2 = 1) = p(l_2, r_1 \parallel r_2 = 1)/\sum_{r_1} p(l_2, r_1 \parallel r_2 = 1)$. The numerator here is simply a marginal of the kernel $p(. \parallel r_2 = 1)$ in (5.9) and is identified as follows:

$$\begin{aligned} p(l_2, r_1 \parallel r_2 = 1) &= \sum_{l_1, l_1^{(1)}} p(l_1^{(1)}, l_2, r_1, l_1 \parallel r_2 = 1) = \sum_{l_1, l_1^{(1)}} \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid r_1, l_1)} \\ &= \sum_{l_1} \frac{p(l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid r_1, l_1)} = \sum_{l_1} p(r_1, l_1) \times p(l_2 \mid r_1, l_1, r_2 = 1). \end{aligned}$$

Consequently, we are able to identify the propensity score of R_1 in the second step and proceed with the intervention on R_1, using the kernel in (5.9). This yields a new kernel that factorizes with respect to the DAG in Fig. 3(c) and corresponds to the target law $p(l_1^{(1)}, l_2^{(1)})$. Putting all the pieces together, the target law is identified as:

$$\begin{aligned} p(l_1, l_2 \parallel r = 1) &= \frac{p(l_1^{(1)}, l_2, r_1 = 1, l_1 \parallel r_2 = 1)}{p(r_1 = 1 \mid l_2 \parallel r_2 = 1)} \\ &= \frac{p(l_1, l_2, r_1 = 1, r_2 = 1)}{p(r_2 = 1 \mid r_1 = 1, l_1) \times p(r_1 = 1 \mid l_2 \parallel r_2 = 1)} \\ &= \frac{p(l_1, l_2, r_1 = 1, r_2 = 1)}{p(r_2 = 1 \mid r_1 = 1, l_1) \times \dfrac{\sum_{l_1} p(l_2 \mid r_2 = 1, r_1 = 1, l_1) \times p(r_1 = 1, l_1)}{\sum_{l_1, r_1} p(l_2 \mid r_2 = 1, r_1, l_1) \times p(r_1, l_1)}}. \end{aligned} \tag{5.10}$$

The identified forms of the two propensity scores in this example are quite different. The propensity score of R_2, $p(r_2 = 1 \mid r_1, l_1)$, corresponds to the conditional factor in the full factorization of the joint. The propensity score of R_1, $p(r_1 = 1 \mid l_2 \parallel r_2 = 1)$, is a complex function of $p(r, l)$ and corresponds to a conditional distribution in a

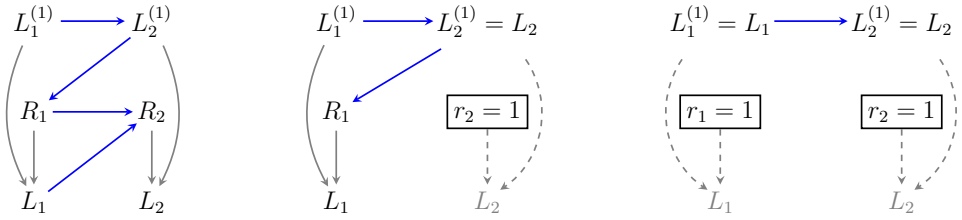


Figure 3: An illustration of the operation of a **sequential** identification algorithm. (a) Permutation model; (b) Intervention on R_2; (c) Intervention on R_1 after R_2.

hypothetical world where $L_2^{(1)}$ is rendered observed and equal to L_2 via an intervention that sets R_2 to 1.

The fact that R_1 had a counterfactual parent $L_2^{(1)}$ in Fig. 3(a), but $R_1 \not\perp\!\!\!\perp R_2 \mid \text{pa}_{\mathcal{G}_m}(R_1)$ ensured that the propensity score $p(r_1 \mid \text{pa}_{\mathcal{G}_m}(r_1))$ required to intervene on R_1 is not immediately identifiable. This induced a strict *total ordering* by which R_1 and R_2 had to be intervened on. A total order is a special case of a *strict partial order* which is defined as an irreflexible, anti-symmetric, and transitive binary relation. In our simple example above, to get identification for the target law we needed to follow the order $\{I_{r_2} < I_{r_1}\}$, where I_{r_k} denotes an intervention on R_k. This sequential procedure generalizes to an arbitrary number of time points in the permutation model, where investigators take repeated measurements of some outcome over K time points. The total order for identifying the target law $p(l_1^{(1)}, \ldots, l_K^{(1)})$ is then given by a reverse topological ordering on the missingness indicators $\{I_{r_K} < \cdots < I_{r_1}\}$, i.e., we begin by intervening on the missingness indicator R_K corresponding to the final time point and proceed backwards (Robins, 1997).

Recall that in the causal model discussed earlier corresponding to Fig. 1, identification was not possible due to the presence of U_1 and U_2 instead of the counterfactuals

$L_1^{(1)}$ and $L_2^{(2)}$. Clearly the key for obtaining identification was the additional information provided by counterfactuals $L^{(1)}$, which are sometimes observed, rather than U variables, which are never observed. In analogy with missing data, one may consequently suggest to consider placing counterfactual versions of observed variables on the graph in causal inference (instead of the hidden variables), with the understanding that a pair of counterfactuals $L^{(1)}, L^{(0)}$ corresponding to L (for some treatment A) may potentially be easier to take into account for identification compared to hidden variables. This is because the counterfactual pair $L^{(1)}, L^{(0)}$ is partiallly observed, due to the consistency property, unlike a purely hidden variable U. However, this is also insufficient for obtaining identification. In Section 6, we discuss why identification strategies in missing data does not readily translate to causal inference.

5.2 Parallel Interventions

The sequential identification strategy in $\mathcal{M}^{\text{per}}$ coupled with consistency (i.e., $L_i^{(1)} = L_i$ if $R_i = 1$) resembles identification strategies employed in causal inference problems where the intervention distribution associated with a multivariate treatment $R = (R_1, R_2)^T$ is obtained sequentially by intervening on treatment variables one at a time. We now discuss an example that shows sequential strategies as in causal inference are insufficient and that there exists a class of missing data models where missingness indicators must be intervened on "in parallel" to identify the target law.

The simplest example of a model in this class is the block-parallel MNAR model $\mathcal{M}^{\text{b-par}}$ with two time points, redrawn in Fig. 4(a). Identification of $p(l_1^{(1)}, l_2^{(1)}) \coloneqq p(l_1, l_2 \parallel r_1 = 1, r_2 = 1)$ can be obtained in a single step as follows:

$$\left.\frac{p(l_1, l_2, r_1, r_2)}{p(r_1 \mid l_2^{(1)}) \times p(r_2 \mid l_1^{(1)})}\right|_{r=1} = \frac{p(l_1, l_2, r_1 = 1, r_2 = 1)}{p(r_1 = 1 \,|\, l_2, r_2 = 1) \times p(r_2 = 1 \,|\, l_1, r_1 = 1)}. \tag{5.11}$$

The first equality holds by (4.6) and the second equality holds due to the indepen-

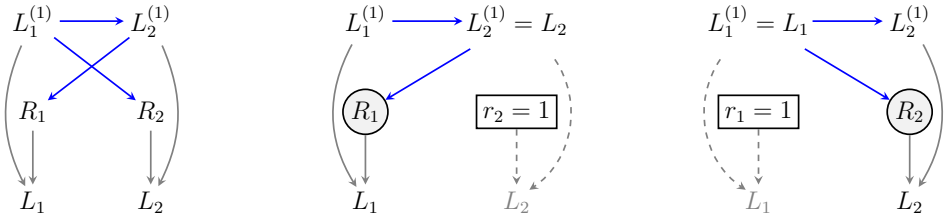


Figure 4: An illustration of the operation of a **parallel** identification algorithm. (a) Block-parallel; (b) Intervention on R_2 and selection on R_1; (c) Intervention on R_1 and selection on R_2.

dence restrictions implied by $\mathcal{M}^{\text{per}}$ which are: $R_1 \perp\!\!\!\perp R_2 \mid L_2^{(1)}$ and $R_1 \perp\!\!\!\perp R_2 \mid L_1^{(1)}$. Given these independencies, both propensity scores in the denominator are easily identified as functions of the observed data.

We now illustrate why the sequential approach to identification fails in this example and that the simultaneous evaluation of the two propensity scores is necessary to yield target law identification. Assume we proceed with our two-step sequential procedure by first intervening on $R_2 = 1$. This results in the following kernel Markov relative to the DAG in Fig. 4(b)

$$p(l_1^{(1)}, l_2, r_1, l_1 \parallel r_2 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2)}{p(r_2 \mid l_1^{(1)})}\bigg|_{r_2=1} = \frac{p(l_1^{(1)}, l_1, l_2, r_1, r_2 = 1)}{p(r_2 = 1 \mid l_1^{(1)}, r_1)}. \tag{5.12}$$

The intervention on R_1 in the second step of the sequential procedure requires dividing the above kernel $p(. \parallel r_2 = 1)$ by the propensity score of R_1, $p(r_1 \mid l_2^{(1)}) = p(r_1 \mid l_2 \parallel r_2 = 1)$. Using Bayes rule, this propensity can be obtained from $p(r_1, l_2 \parallel r_2 = 1)$. However, we can only identify $p(r_1 = 1, l_2 \parallel r_2 = 1)$ from $p(. \parallel r_2 = 1)$. This is because $l_1^{(1)}$ appears in both the numerator and denominator, and when $R_1 = 0$, we cannot observe $l_1^{(1)}$. Hence, the kernel in (5.12) is identified only when evaluated at

$R_1 = 1$. The fact that this kernel is not available at all levels of R_1 prevents us from sequentially obtaining $p(r_1 \mid \mathrm{pa}_{\mathcal{G}_m}(r_1)) = p(r_1 \mid l_2 \parallel r_2 = 1) = p(l_2, r_1 \parallel r_2 = 1)/p(l_2 \parallel r_2 = 1)$ (where the first equality follows from invariance and the second from rules of kernel probability) due to our inability to sum out R_1 from $p(l_2, r_1 \parallel r_2 = 1)$ to obtain $p(l_2 \parallel r_2 = 1)$.

The above is an example of what we term *selection* bias. In this case, division by $p(r_2 \mid l_1^{(1)})$ in (5.12) induces *selection* on R_1, a variable that is not yet intervened on, in the kernel $p(. \parallel r_2 = 1)$. We draw a gray circle around the vertices on the graph that are selected on (or conditioned on) to distinguish it from interventions. Attempting a sequential identification procedure starting with R_1 would similarly induce selection bias on R_2 in the kernel $p(. \parallel r_1 = 1)$, as shown in Fig. 4(c). Since it is not possible to obtain identification by performing the intervention operation in a sequence, no total ordering on missingness indicators can be imposed here. Instead, interventions on R_1 and R_2 are *incomparable*, and thus form a partial rather than a total ordering, which is simply denoted via $\{I_{r_1}, I_{r_2}\}$. Specifically, this notation denotes a partial order relationship on a set of two elements corresponding to interventions on indicators R_1 and R_2, where these elements are incompatible according to the order.

In some missingness models, we might be able to identify the target law in multiple ways. For instance, in the block-sequential MNAR model, the target law can be identified via either a sequence of interventions on missingness indicators in R, or via a one-step parallel intervention on all variables in R. These two strategies result in two equivalent representations of the same identifying functional. These representations may, however, suggest different estimation strategies.

5.3 Sequential and Parallel Interventions

The previous two examples considered how identification of the target law in missing data problems entails evaluating the g-formula either sequentially or in parallel. These examples are special cases of a general identification algorithm for graphical missing data models, introduced in Shpitser et al. (2015). However, this algorithm is not complete in that it fails to account for the following example, and examples like it, where sequential and parallel applications of the g-formula must be combined in order to identify all of the propensity scores, and hence the target law.

Consider the missing data model shown in Fig. 5(a) on three variables. The target law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}) \coloneqq p(l_1, l_2, l_3 \parallel r_1 = 1, r_2 = 1, r_3 = 1)$ is equivalent to the following:

$$\left. \frac{p(l_1, l_2, l_3, r_1, r_2, r_3)}{p(r_1 \mid l_2^{(1)}, l_3^{(1)}) \times p(r_2 \mid l_3^{(1)}, r_1) \times p(r_3 \mid l_2^{(1)}, r_1)} \right|_{r=1}. \tag{5.13}$$

Neither of the approaches discussed in the previous two subsections would yield an identified missingness mechanism for the example of Fig. 5(a). First, the sequential application discussed in (5.1) fails since the selection induced on R_2 after intervening on R_3 impedes our next move in doing an intervention on R_2. This is because obtaining $p(r_2 \mid \mathrm{pa}_{\mathcal{G}_m}(r_2))$ from the kernel $p(. \parallel r_3 = 1)$ requires summing over R_2 (by Bayes rule), and that is not possible as R_2 is conditioned/forced to be one in the kernel. A similar problem persists if we intervene on R_2 first (R_3 is conditioned to be one in the kernel $p(. \parallel r_2 = 1)$). A sequential application starting with R_1 is not possible since we cannot immediately obtain $p(r_1 \mid \mathrm{pa}_{\mathcal{G}_m}(r_1))$. This is because R_1 has counterfactual parents $L_2^{(1)}$ and $L_3^{(1)}$, but is not conditionally independent of the corresponding missingness indicators R_2 and R_3 given its parents. The parallel application discussed in (5.2) fails for the same reason, i.e., the propensity score for R_1 is not immediately identified.

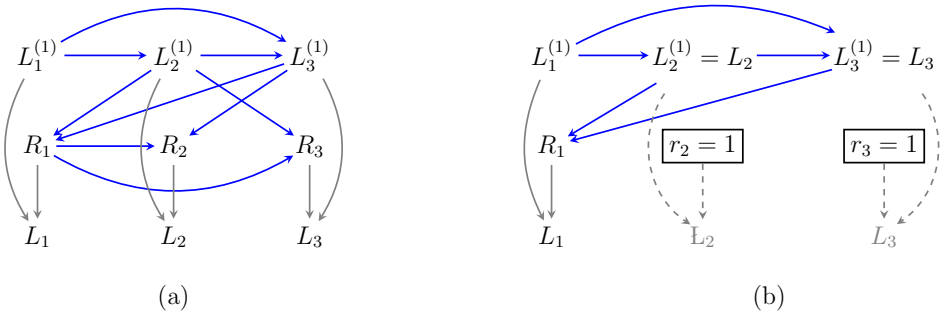


Figure 5: (a) An example m-DAG corresponding to a model where interventions must be applied both *sequentially and in parallel* to yield identification; (b) Graph derived from (a) representing the intermediate step of the identification algorithm where R_2 and R_3 are simultaneously set to 1.

We now show identification of this target distribution is still possible, but only by evaluating the g-formula in two sequential steps, with the first step consisting of a parallel evaluation of propensity scores for R_2 and R_3, and the second step consisting of the evaluation of propensity score for R_1 using the kernel distribution obtained from the first step.

Step 1. *Intervene on R_2 and R_3 in parallel to get $p(. \parallel r_2 = 1, r_3 = 1)$.* The factorization of join distribution in Fig. 5(a) implies $R_2 \perp\!\!\!\perp R_3 \mid L_3^{(1)}, R_1$ and $R_3 \perp\!\!\!\perp R_2 \mid L_2^{(1)}, R_1$. These independencies directly yield identified propensity scores for R_2 and R_3 as follows:

$$p(r_2 \mid \text{pa}_{\mathcal{G}_m}(r_2))|_{r=1} = p(r_2 \mid l_3^{(1)}, r_1)|_{r=1} = p(r_2 = 1 \mid l_3, r_1 = r_3 = 1), \tag{5.14}$$

$$p(r_3 \mid \text{pa}_{\mathcal{G}_m}(r_3))|_{r=1} = p(r_3 \mid l_2^{(1)}, r_1)|_{r=1} = p(r_3 = 1 \mid l_2, r_1 = r_2 = 1). \tag{5.15}$$

This immediately implies that R_2 and R_3 can be intervened on in parallel, similar to the block-parallel example in the previous subsection. This results in the following

kernel that factorizes with respect to the DAG in Fig. 5(b),

$$p(l_1^{(1)}, l_2, l_3, r_1, l_1 \parallel r_2 = r_3 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, l_3, r_1, r_2 = 1, r_3 = 1)}{p(r_2 = 1 | l_3, r_1, r_3 = 1) \times p(r_3 = 1 | l_2, r_1, r_2 = 1)}.$$

The same strategy cannot be used for expressing $p(r_1 \mid \mathrm{pa}_{\mathcal{G}}(r_1)) = p(r_1 \mid l_2^{(1)}, l_3^{(1)})$ as a function of the factual distribution since $R_1 \not\perp\!\!\!\perp \{R_2, R_3\} \mid L_2^{(1)}, L_3^{(1)}$ under this model. However, a second, more involved, step leads to identifying the propensity score of R_1 and thus the target law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)})$.

Step 2. *Intervene on R_1 after intervening on R_2, R_3 to get $p(. \parallel r_1 = 1, r_2 = 1, r_3 = 1)$.* In the second step, we want to intervene on R_1 using the above kernel $p(. \parallel r_2 = 1, r_3 = 1)$ Markov relative to the DAG in Fig. 5(b). In order to perform this intervention, we need to show that the propensity of R_1, $p(r_1 \mid l_2^{(1)}, l_2^{(1)}) = p(r_r \mid l_2, l_3 \parallel r_2 = 1, r_3 = 1)$ is a function of $p(r, l)$. We have $p(r_1 \mid l_2, l_3 \parallel r_2 = r_3 = 1) = p(l_2, l_3, r_1 \| r_2 = r_3 = 1) / \sum_{r_1} p(l_2, l_3, r_1 \| r_2 = r_3 = 1)$. The numerator here is identified by simply marginalizing out $l_1^{(1)}$ from the kernel $p(. \| r_2 = r_3 = 1)$. Consequently, the propensity score of R_3 is rendered identified and we can identify the target law in (5.13) via:

$$p(l_1, l_2, l_3 \parallel r = 1) = \frac{p(l_1^{(1)}, l_2, l_3, r_1, l_1 \parallel r_2 = 1, r_3 = 1)}{p(r_1 \mid l_2, l_3 \parallel r_2 = 1, r_3 = 1)}\Bigg|_{r_1 = 1}.$$

In conclusion, the parallel interventions on R_2 and R_3 do not induce selection bias on R_1. Therefore, we are able to proceed with the sequential application of the g-formula and obtain $p(r_1 \mid \mathrm{pa}_{\mathcal{G}_m}(r_1))$ from the kernel $p(. \parallel r_2 = 1, r_3 = 1)$. In other words, identification of the target law is only possible if R_1 was intervened on only after R_2 and R_3 were simultaneously intervened on. This identification procedure also induces a partial ordering for the interventions on the missingness indicators. The partial order of interventions in the above example can be written as $\{\{I_{r_2}, I_{r_3}\} < I_{r_1}\}$; that is interventions on R_2 and R_3 are incompatible, but both must occur prior to an intervention on R_1.

5.4 Identifying Propensity Scores with Different Partial Orders

In the examples discussed so far, all missingness indicators were intervened on according to a *partial order* defined on the set $\{R_i \in R\}$. A procedure for target law identification may then aim at exploring the space of all possible partial orders. This effectively entails trying out all possible combinations of parallel and sequential application of the g-formula. Bhattacharya et al. (2019) showed that such a procedure remains incomplete, meaning that it fails to recognize a class of missing data models where the target law is indeed identified. They took an alternative view of the target law identification problem, where each propensity score $p(r_i \mid \text{pa}_{\mathcal{G}}(r_i))$ may be identified separately, using a potentially distinct partial order of intervention operations. This entails considering subproblems where selection bias, hidden variables, or both, are introduced, even if these complications are absent in the original problem.

Consider the missing data DAG in Fig. 6(a). According to Proposition 1, the target law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}) := p(l_1, l_2, l_3 \parallel r_1 = 1, r_2 = 1, r_3 = 1)$ is equivalent to the following:

$$p(l_1, l_2, l_3 \parallel r = 1) = \frac{p(l_1, l_2, l_3, r_1, r_2, r_3)}{p(r_1 | l_2^{(1)}) \times p(r_2 | l_1^{(1)}, l_3^{(1)}, r_1) \times p(r_3 | l_2^{(1)}, r_1)} \Bigg|_{r=1}. \tag{5.16}$$

We can easily use the encoded independence restrictions, along with the consistency facts in missing data, to identify the propensity scores of R_2 and R_3 evaluated at $R = 1$. These restrictions are: $R_2 \perp\!\!\!\perp R_3 \mid R_1, L_1^{(1)}, L_3^{(1)}$ and $R_3 \perp\!\!\!\perp R_2 \mid R_1, L_2^{(1)}$. We have,

$$p(r_2 \mid \text{pa}_{\mathcal{G}_m}(r_2))|_{r=1} = p(r_2 = 1 \mid l_1, l_3, r_1 = 1, r_3 = 1), \tag{5.17}$$

$$p(r_3 \mid \text{pa}_{\mathcal{G}_m}(r_3))|_{r=1} = p(r_3 = 1 \mid l_2, r_1 = 1, r_2 = 1). \tag{5.18}$$

The propensity score of R_1 however, is not immediately identified since $R_1 \not\perp\!\!\!\perp R_2 \mid L_2^{(1)}$. A natural question to ask is whether this propensity score is identified from a

kernel distribution where R_2 and R_3 are simultaneously intervened on. If so, then we can follow a similar argument as in the previous example with the two-step sequential procedure. Performing parallel interventions on R_2 and R_3 yield the following kernel distribution,

$$p(l_1^{(1)}, l_2, l_3, r_1 \parallel r_2 = r_3 = 1) = \frac{p(l_1^{(1)}, l_1, l_2, l_3, r_1, r_2 = r_3 = 1)}{p(r_2 = 1 | l_1^{(1)}, l_3, r_1, r_3 = 1) \times p(r_3 = 1 | l_2, r_1, r_2 = 1)}.$$

Following the second step of the sequential procedure, we need to first identify the propensity score of R_1 via the above kernel, i.e., identifying $p(r_1 \mid l_2 \parallel r_2 = 1, r_3 = 1)$ which by Bayes rule can be obtained from $p(r_1, l_2 \parallel r_2 = 1, r_3 = 1)$. Unfortunately, the above kernel $p(. \parallel r_2 = 1, r_3 = 1)$ exhibits selection on R_1. Consequently, the two-stage sequential strategy in the previous example fails to identify the target law in this current example. Nevertheless, the target law is still identifiable using a more involved argument outlined below. Since the propensity scores of R_2 and R_3 are identified via (5.17) and (5.18), all we need to do to identify the target law is to find a way to identify the propensity score of R_1, $p(r_1 \mid l_2^{(1)})$.

Consider the marginal distribution $\sum_{l_1, l_1^{(1)}} p(l^{(1)}, r, l)$, where the full law factorizes according to the m-DAG in Fig. 6(a). The non-deterministic portion of this marginal distribution factorizes as: $p(l_2^{(1)}, l_3^{(1)}, r_1, r_2, r_3) = p(l_2^{(1)}) \times p(l_3^{(1)} | l_2^{(1)}) \times p(r_1 | l_2^{(1)}) \times p(r_2 | r_1, l_3^{(1)}) \times p(r_3 | l_2^{(1)}, r_1)$. The above factorization is Markov relative to the m-DAG in Fig. 6(b), where $L_1, L_1^{(1)}$ are projected out (or in other words, treated as hidden variables) from Fig. 6(a). Aside from absence of $p(l_1^{(1)})$ and $p(l_1 \mid r_1, l_1^{(1)})$ factors, the difference between the above factorization and the one from original m-DAG in Fig. 6(a) is the difference in propensity score of R_2, i.e., $p(r_2 \mid r_1, l_1^{(1)}, l_3^{(1)})$ vs. $p(r_2 \mid r_1, l_3^{(1)})$. We now illustrate that this change in the propensity score of R_2, which we call a *pseudo-propensity score* for R_2, can overcome the selection bias issue on R_1 that resulted from an intervention on R_2.

The pseudo-propensity score of R_2, denoted by $\widetilde{p}(r_2 \mid r_1, l_3^{(1)})$, is easily identified

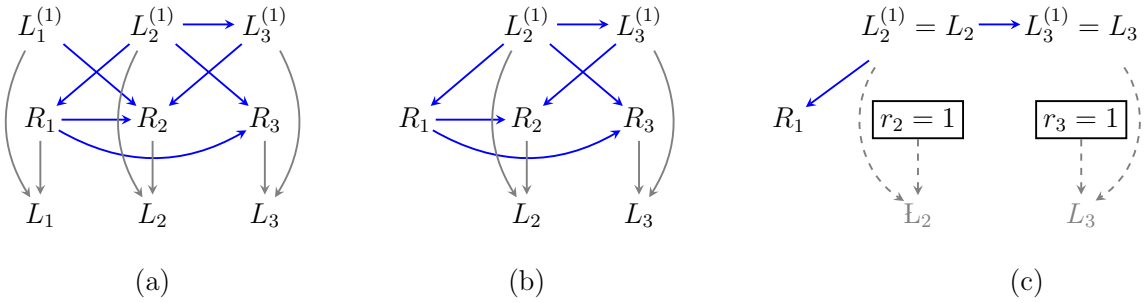


Figure 6: (a) An example m-DAG corresponding to a model where intervention on a single missingness indicator entails following a partial order of interventions; (b), (c) Selection bias on R_1 is avoidable by following a partial order intervention on a graph induced by projecting out $L_1^{(1)}, L_1$.

via $\widetilde{p}(r_2 \mid r_1, l_3, r_3 = 1)$ since $R_2 \perp\!\!\!\perp R_3 \mid R_1, L_3^{(1)}$ in Fig. 6(b). Now to identify the propensity score of R_1, we proceed as follows. Using the pseudo-propensity score of R_2 and the propensity score of R_3, we obtain the following kernel distribution:

$$\widetilde{p}(l_2, l_3, r_1 \parallel r_2 = 1, r_3 = 1) = \frac{p(l_2, l_3, r_1, r_2 = 1, r_3 = 1)}{\widetilde{p}(r_2 = 1 \mid r_1, l_3, r_3 = 1) \times p(r_3 = 1 \mid l_2, r_1, r_2 = 1)}.$$

The above kernel $\widetilde{p}(. \parallel r_2 = 1, r_3 = 1)$, which factorizes according to the m-DAG in Fig. 4(c), is a direct function of observed data law without inducing any selection bias on R_1. Consequently, we can identify the propensity score of R_1, $p(r_1 \mid l_2 \parallel r_2 = 1, r_3 = 1)$ using the kernel $\widetilde{p}(. \parallel r_2 = 1, r_3 = 1)$. Using Bayes rule, we have $p(r_1 \mid l_2 \parallel r_2 = 1, r_3 = 1) = \sum_{l_3} p(r_1, l_2, l_3 \parallel r_2 = 1, r_3 = 1) / \sum_{l_3, r_1} p(r_1, l_2, l_3 \parallel r_2 = 1, r_3 = 1)$. This concludes that the target law in (5.16) is identified, since we showed each term in the denominator is identified.

The idea of focusing on identifying each propensity score separately leads to new identification strategies. Specifically, in the above example, the propensity score of R_1 was identified only after treating $L_1^{(1)}$ as a hidden variable and marginalizing it out

from the original distribution. Bhattacharya et al. (2019) developed a general procedure based on these observations which has significantly narrowed the identifiability gap in graphical models of missing data. Their procedure generalizes the notion of finding a partial order of interventions on the set $\{R_i \in R\}$ to finding partial orders of interventions for each individual $R_k \in R$ by exploring subproblems where a set of variables are treated as hidden or unmeasured.

In the above example, the partial orders for R_2 and R_3 are trivial – the corresponding propensity scores are immediately identified from the observed distribution. We can summarize the identification procedure for obtaining the propensity score of R_1 via the following partial order executed in a graph where $\{L_1^{(1)}, L_1\}$ are treated as hidden variables, which we will denote as $\mathcal{G}_m(V \setminus \{L_1^{(1)}, L_1\})$: the partial order of interventions for R_1 can be summarized via $\big\{\{I_{r_2}, I_{r_3}\} < I_{r_1}\big\}$ in $\mathcal{G}_m(V \setminus \{L_1^{(1)}, L_1\})$. That is, intervention on R_1 must occur after interventions on R_2 and R_3 in a graph where $L_1^{(1)}$ and L_1 are marginalized out, and as mentioned earlier interventions on R_2 and R_3 are incompatible.

5.5 Intervention on Variables Outside of R

A feature common to all previous examples is that all propensity scores were obtained via partial orders of interventions defined only on the missingness indicators. This however, is also not always sufficient. In general the propensity score of R_k might be identified only by intervening on variables outside of R, including variables in $L^{(1)}$ that become observed after intervening or conditioning on relevant elements in R. As an example, consider the m-DAG in Fig 7(a) where L_3 is fully observed. This graph was considered in Bhattacharya et al. (2019) as a generalization of a model described in Shpitser et al. (2015). According to Proposition 1, we can write down the target

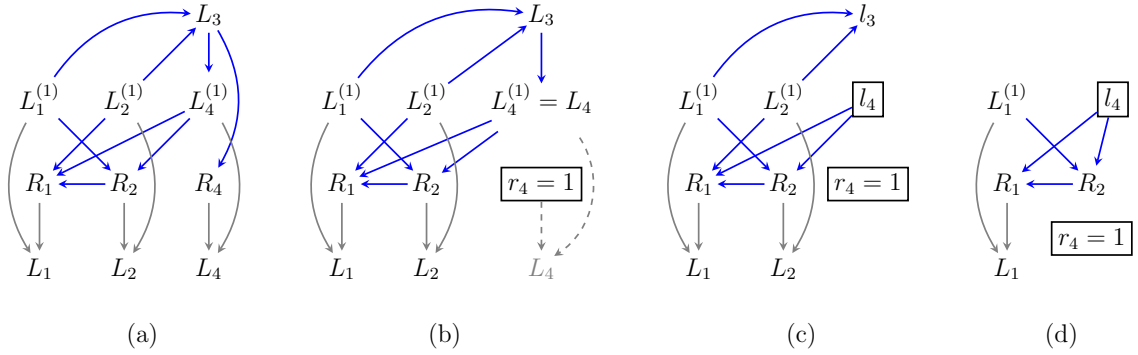


Figure 7: (a) An example m-DAG corresponding to a model where variables besides Rs are required to be intervened on in order to identify the propensity scores; (b) A conditional m-DAG where R_4 is intervened on; (c) A conditional m-DAG where L_4 is intervened on after an intervention on R_4; (d) A conditional m-DAG where $L_2^{(1)}$ and L_3 are marginalized out from the kernel $p(. \mid \text{do}(r_4 = 1, l_4))$.

law $p(l_1^{(1)}, l_2^{(1)}, l_4^{(1)}, l_3)$ as follows:

$$p(l_1, l_2, l_3, l_4 \parallel r = 1) = \frac{p(l_1, l_2, l_3, l_4, r_1, r_2, r_4)}{p(r_1 | l_2^{(1)}, l_4^{(1)}, r_2) \times p(r_2 | l_1^{(1)}, l_4^{(1)}) \times p(r_4 | l_3)}\Bigg|_{r=1}. \tag{5.19}$$

The propensity score of R_4, $p(r_4 \mid l_3)$, is a direct function of observed data. The propensity score of R_1, $p(r_1 \mid r_2, l_2^{(1)}, l_4^{(1)})$ evaluated at $R = 1$ is also a function of observed data via $p(r_1 = 1 \mid r_2 = 1, l_2, l_4, r_4 = 1)$ since $R_1 \perp\!\!\!\perp R_4 \mid R_2, L_2^{(1)}, L_4^{(1)}$. Thus, identification of the target law relies on whether the propensity score of R_2, $p(r_2 \mid l_1^{(1)}, l_4^{(1)})$ is identified or not, which is not immediately clear since $R_2 \not\perp\!\!\!\perp R_1 \mid L_1^{(1)}, L_4^{(1)}$. We now outline a procedure to identify this propensity score.

We first intervene on R_4, i.e., dropping the term $p(r_4 \mid l_3)$ from the original factorization, yielding the kernel below which is Markov relative to the graph in Fig. 7(b),

$$p(l_1^{(1)}, l_2^{(1)}, l_3, l_4, r_1, r_2, l_1, l_2 \parallel r_4 = 1) = \frac{p(l_1^{(1)}, l_2^{(1)}, l_3, l_4, l_1, l_2, r_1, r_2, r_4 = 1)}{p(r_4 = 1 \mid l_3)}.$$

After intervening on R_4, $L_4^{(1)}$ is fully observed. We then intervene on L_4, i.e., dropping the term $p(l_4 \mid l_3 \parallel r_4 = 1)$ from the above kernel and obtain the following kernel which is Markov relative to the graph in Fig. 7(c),

$$p(l_1^{(1)}, l_2^{(1)}, l_3, r_1, r_2, l_1, l_2 \parallel r_4 = 1, l_4) = \frac{p(l_1^{(1)}, l_2^{(1)}, l_3, l_4, l_1, l_2, r_1, r_2, r_4 = 1)}{p(r_4 = 1 \mid l_3) \times p(l_4 \mid l_3, r_4 = 1)}.$$

Note that in this example, marginalizing out L_3 is equivalent to intervening on L_3. We can thus safely marginalize out $L_2^{(1)}$, L_2, and L_3 from the above expression without changing the propensity score of R_2, yielding a new kernel that is Markov relative to the graph in Fig. 7(d),

$$p(l_1^{(1)}, r_1, r_2, l_1 \parallel r_4 = 1, l_4) = \sum_{l_3} \frac{p(l_1^{(1)}, l_3, l_4, r_1, r_2, r_4 = 1)}{p(r_4 = 1 \mid l_3) \times p(l_4 \mid l_3, r_4 = 1)}.$$

The propensity score of R_1 in the above kernel is different than the one in original factorization of the m-DAG in Fig. 7(a). We refer to this as the pseudo-propensity score of R_1 and denote it by $\widetilde{p}(r_1 \mid r_2 \parallel r_4, l_4)$. We now intervene on R_1 by dropping this term from the above kernel, yielding

$$p(l_1, r_2 \parallel r_1 = 1, r_4 = 1, l_4) = \frac{p(l_1, r_1 = 1, r_2 \parallel r_4 = 1, l_4)}{\widetilde{p}(r_1 = 1 \mid r_2 \parallel r_4 = 1, l_4)}.$$

The desired propensity score of R_2 (which remains invariant despite all previous operations as its direct causes are still present in Fig. 7(d)) is then identified in the above kernel as $p(l_1, r_2 \parallel r_1 = 1, r_4 = 1, l_4) / \sum_{r_2} p(l_1, r_2 \parallel r_1 = 1, r_4 = 1, l_4)$. Since all the propensity scores are identified, then the target law in (5.19) is identified as well.

In the above example, the partial orders for R_1 and R_4 are trivial – the corresponding propensity scores are immediately identified from the observed distribution. We can summarize the identification procedure for obtaining the propensity score of R_2 via the following partial order executed in a graph where R_4 then L_4 are intervened on and $\{L_2^{(1)}, L_2, L_3\}$ are treated as hidden variables, which we will denote

as $\mathcal{G}_m(V \setminus \{L_2^{(1)}, L_2, L_3\}, \{r_4, l_4\})$: the partial order of interventions for R_2 can be summarizes via $\left\{I_{r_4} < I_{l_4} < I_{r_1} < I_{r_2}\right\}$ in $\mathcal{G}_m(V \setminus \{L_2^{(1)}, L_2, L_3\}, \{r_4, l_4\})$. That is, intervention on R_2 must occur after interventions on R_4, L_4 and R_1 in a graph where $L_2^{(1)}, L_2$, and L_3 are marginalized out, and as mentioned earlier interventions on R_1 and R_4 are incompatible.

5.6 A Unifying Identification Procedure

Bhattacharya et al. (2019) proposed a procedure for target law identification that combines all the ideas discussed above. It proceed as follows. For each missingness indicator $R_k \in R$, it proceeds to identify its propensity score $p(r_k|\, \text{pa}_{\mathcal{G}_m}(r_k))$ evaluated at $R = 1$. It does so by checking if R_k is conditionally independent (given its parents) of the corresponding missingness indicators of its counterfactual parents. If this is the case, the propensity score is identified by a simple conditional independence argument (d-separation). Otherwise, the procedure checks if this condition holds in any intervention distribution where a subset of missingness indicators are intervened on, in either the original model or marginals of the model where the direct causes of R_k are still preserved. If the procedure succeeds in identifying the propensity score for each missingness indicator in this manner, then the target law is declared as being identified.

Necessary conditions for target law identification have been discussed in the literature. A well-known result states that if an underlying variable causes its own missingness status, known as a self-censoring mechanism or nonignorable mechanism, then the target law is provably not identified. This means we can construct two missing data models that differ in target law distributions but both map to the same observed data distribution. The graphical structure simply corresponds to existence of an edge of the form $L_k^{(1)} \to R_k$ in the m-DAG. Nabi and Bhat-

tacharya (2022) have also shown that the so-called "criss-cross" structure prevents target law identification. These structures involve a pair of variables $L_i^{(1)}, L_j^{(1)}$ that are directly connected as $L_i^{(1)} \to L_j^{(1)}$ or $L_i^{(1)} \leftarrow L_j^{(1)}$ and these edges exist simultaneously: $L_i^{(1)} \to R_j \leftarrow R_i \leftarrow L_j^{(1)}$. On the other hand, Nabi et al. (2020) have provided sufficient conditions under which a target law is identified. They show under the absence of self-censoring edges and so-called "colluder" structures, the target law is identified. A colluder is a special type of collider where there exists $R_i, R_j \in R$ such that $L_i^{(1)} \to R_j \leftarrow R_i$. Finding necessary and sufficient conditions (a sound and complete algorithm) for target law identification remains an open problem.

In many applied problems, some variables are not just missing but completely unobserved. We can use missing data DAG models with hidden variables to encode the presence of unmeasured confounders and generalize the aforementioned identification strategies to such settings. We discuss extensions to identification in the presence of both missing data and hidden variables in Appendix S2.

6. Discussion

We have shown how unique features of missing data models represented by DAGs allow identification in seemingly counterintuitive situations, by taking advantage of Markov restrictions linking missingness indicators and underlying counterfactual variables. For instance, the target law is identified in the bivariate permutation model (Fig. 3(a)), but not in the analogous hidden variable causal model (Fig. 1(a)). Similarly, the target law is identified in the bivariate block-parallel model, shown in Fig. 9(a), by (5.11), but not in the analogous causal model shown in Fig. 9(b).

We described how graphical missing data models are a special case of hidden variable graphical causal models, where hidden variables are replaced by counterfactual variables, which are sometimes observed. It is this partial observability which allowed

identification to be derived. Just as in missing data models, observed variables in causal models are derived from counterfactual variables by means of consistency, and thus may be viewed as causes (parents) of observed variables in a particular type of causal graphs. This leads to a natural question: would expressing causal models via graphs which make the relationship between observed and counterfactual variables explicit by placing counterfactuals as vertices on the graph yield new types of identification results?

Consider a modification of the graph in Fig. 1(a) example shown in Fig. 8(a), which is drawn in such a way that all counterfactual versions of observed variables all occur prior to every observed variable. In this graph, variables L_1 and L_2 each have *two* counterfactuals on the graph, with each counterfactual $L_i^{(r_i)}$ inheriting all incoming and outgoing edges from the corresponding missing data counterfactual $L_i^{(1)}$. In addition, $L_i^{(1)}, L_i^{(0)}$ are associated, as indicated by the red bidirected edge between them. Fig. 8(a) yields a causal model that simply asserts all independence statements in the graph.

Note that in this causal model, the intervention $r_2 = 1$ is identified from the observed data, since R_2 only has observed parents, namely L_1 and R_1. The graph where this intervention is performed is shown via the conditional causal DAG in Fig. 8(b).

Just as in missing data, intervention $r_2 = 1$ renders the counterfactual $L_2^{(1)}$ and L_2 the same, and thus renders the former as an observed variable. Since both $L_2^{(1)}$ and $L_2^{(0)}$ serve as parents of R_1, rendering $L_2^{(1)}$ as observed does not suffice to render the subsequent intervention on R_1 identified, with $L_2^{(0)}$ acting as an unobserved confounder preventing identification.

However, if an additional assumption of *rank preservation* is made, where $L_2^{(0)}$ is equal to $g(L_2^{(1)})$, for some bijection $g(.)$, then identification is recovered, since

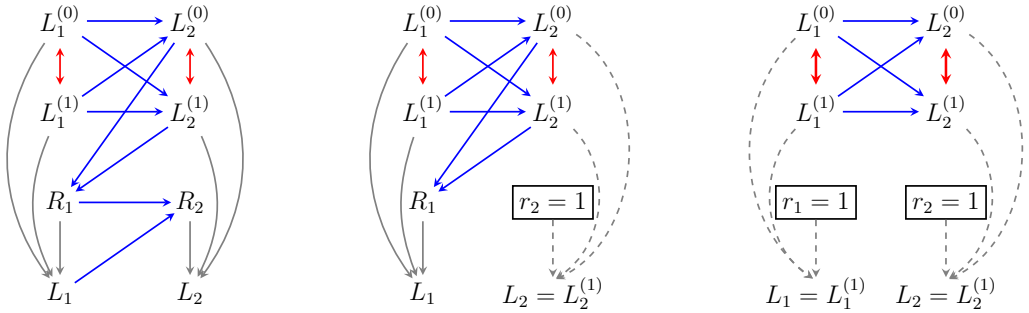


Figure 8: (a) A causal inference version of the bivariate permutation model with two counterfactual versions of L_1 and L_2 on the graph; (b) A world where an intervention $r_2 = 1$ is performed, yielding a situation where L_2 and $L_2^{(1)}$ coincide, i.e., $L_2 = L_2^{(1)}$; (c) A causal model where we impose a *rank preservation* relationship where a known bijective function $g(.)$ exists, such that $L_2^{(0)} = g(L_2^{(1)})$. In this model, the joint distribution $p(L_1^{(r_1)}, L_2^{(r_2)})$ is identified by sequentially intervening on r_2 then r_1.

$p(R_1 \mid L_2^{(1)}, L_1^{(0)}) = p(R_1 \mid L_2^{(1)}, g(L_2^{(1)})) = p(R_1 \mid L_2^{(1)})$ becomes a functional of the observed data once an intervention $r_2 = 1$ is performed. This is shown in Fig. 8 (c), with the bijective relationship between $L_2^{(1)}$ and $L_2^{(0)}$ shown as a deterministic relationship via a thicker bidirected edge. We discuss additional causal models in which rank preservation yields identification in the Appendix.

As a second example for illustrating why identification strategies in missing data models do not easily translate to causal models, consider the causal analogue of the bivariate block-parallel model shown in Fig. 9(b), with two binary treatments A_1, A_2 instead of missingness indicators, and two observed outcomes L_1, L_2, which implies two counterfactual versions of each observed outcome: $L_1^{(1)}, L_1^{(0)}$, and $L_2^{(1)}, L_2^{(0)}$ (we exclude the observed proxies L_1, L_2 from the figure for brevity). The model is defined

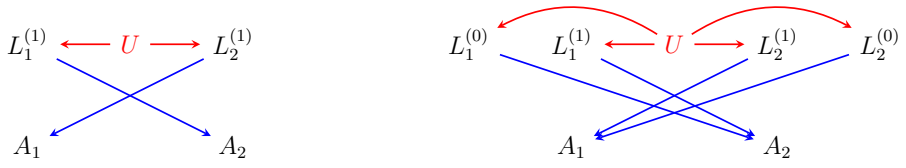


Figure 9: (a) The bivariate block parallel missing data model. (b) The causal model analogue of the model in (a) where identification of causal effects is not possible without further assumptions, but is possible with rank preservation.

by the following restrictions:

$$A_2 \perp\!\!\!\perp L_2^{(1)}, L_2^{(0)}, A_1 \mid L_1^{(1)}, L_1^{(0)} \quad \text{and} \quad A_1 \perp\!\!\!\perp L_1^{(1)}, L_1^{(0)}, A_2 \mid L_2^{(1)}, L_2^{(0)}. \tag{6.20}$$

These restrictions do not yield nonparametric identification since the propensity scores in the model depend on both versions of the counterfactuals, and are thus considered "cross-world" quantities:

$$p(a_2 \mid l_2^{(1)}, l_2^{(0)}, a_1, l_1^{(1)}, l_1^{(0)}) = p(a_2 \mid l_1^{(1)}, l_1^{(0)}),$$
$$p(a_1 \mid l_2^{(1)}, l_2^{(0)}, a_1, l_1^{(1)}, l_1^{(0)}) = p(a_1 \mid l_2^{(1)}, l_2^{(0)}).$$

However, we can recover an argument for identification of $p(l_1^{(i)}, l_2^{(j)})$ for any $(i,j) \in \{0,1\}^2$ via rank preservation, which states that for $k \in \{0,1\}$ there exist bijections $g_k(.)$ such that $L_1^{(1-i)} = g_1(L_1^{(i)})$ and $L_2^{(1-j)} = g_2(L_2^{(j)})$. Identification of $p(l_1^{(i)}, l_2^{(j)})$ then proceeds as follows:

$$\begin{aligned}
p(l_1^{(i)}, l_2^{(j)}) &= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{p(A_1 = i, A_2 = j \mid l_1^{(i)}, l_2^{(j)})} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{\sum_{l_1^{(1-i)}, l_2^{(1-j)}} p(A_1 = i, A_2 = j \mid l_1^{(i)}, l_2^{(j)}, l_1^{(1-i)}, l_2^{(1-j)}) \times p(l_1^{(1-i)}, l_2^{(1-j)} \mid l_1^{(i)}, l_2^{(j)})} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{\sum_{l_1^{(1-i)}, l_2^{(1-j)}} p(A_1 = i \mid l_2^{(j)}, l_2^{(1-j)}, A_2 = j) \times p(A_2 = j \mid l_1^{(i)}, l_1^{(1-i)}, A_1 = i) \times p(l_1^{(1-i)}, l_2^{(1-j)} \mid l_1^{(i)}, l_2^{(j)})} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{\sum_{l_1^{(1-i)}, l_2^{(1-j)}} p(A_1 = i | l_2^{(j)}, l_2^{(1-j)}, A_2 = j) \times p(A_2 = j | l_1^{(i)}, l_1^{(1-i)}, A_1 = i) \times \mathbb{I}(l_1^{(1-i)} = g_1(l_1^{(i)}), l_2^{(1-j)} = g_2(l_2^{(j)}))} \\
&= \frac{p(l_1^{(i)}, l_2^{(j)}, A_1 = i, A_2 = j)}{p(A_1 = i \mid l_2^{(j)}, A_2 = j) \times p(A_2 = j \mid l_1^{(i)}, A_1 = i)}
\end{aligned}$$

$$= \frac{p(l_1, l_2, A_1 = i, A_2 = j)}{p(A_1 = i \mid l_2, A_2 = j) \times p(A_2 = j \mid l_1, A_1 = i)}.$$

Here, the first and second equalities follow by rules of probability, the third by (6.20), the fourth and fifth by the rank preservation assumption, and the last by consistency. This derivation is structurally very similar to the derivation for the bivariate block-parallel model, except for the last two steps which explicitly rely on rank preservation. In Appendix S4 we show that this structural similarity is quite general, and a similar identification strategy can be defined for a K variable causal analogue of the block-parallel model endowed with rank preservation.

While much of the discussion in this paper has focused on how causal identification techniques can be applied or extended to missing data settings, the above example demonstrates missing data techniques can only be applied to causal settings given stronger assumptions, such as rank preservation, than those typically encoded in causal models. This raises interesting questions for future work on the suitability of these assumptions in missing data – e.g., an examination of the "no interference" assumption, which was implicit in all models discussed in this paper and violations of which may require different identification strategies.

Supplementary Materials

The supplementary materials contain discussions on identification of missing data DAG models using the odds ratio parameterization extension of identification techniques to m-DAG models with unmesaured confounders, results on identification of the full law, proofs, and some additional results on identification in causal models using rank preservation.

SUPPLEMENTARY MATERIALS

S1. Identification via Odds Ratio Parameterization

We mentioned in the main draft that Nabi et al. (2020) used an odds ratio parameterization to derive a sound and complete algorithm for full law identification in m-DAGs. In the following, we go over an example to show how the target law can be identified via such a parameterization.

Consider the m-DAG in Fig. 10(a). The non-deterministic portion of the full law factorizes as $p(l_1^{(1)}) \times p(l_2^{(1)} \mid l_1^{(1)}) \times p(l_3^{(1)} \mid l_1^{(1)}, l_2^{(1)}) \times p(r_1 \mid r_2, l_3^{(1)}) \times p(r_2 \mid r_3, l_1^{(1)}) \times p(r_3 \mid l_1^{(1)})$. The following conditional independence statements follow from this factorization: $R_1 \perp\!\!\!\perp \{L_1^{(1)}, L_2^{(1)}, R_3\} \mid R_2, L_3^{(1)}$, and $R_2 \perp\!\!\!\perp \{L_2^{(1)}, L_3^{(1)}\} \mid R_3, L_1^{(1)}$ and $R_3 \perp\!\!\!\perp \{L_2^{(1)}, L_3^{(1)}\} \mid L_1^{(1)}$.

According to Proposition 1, we have:

$$p(l_1, l_2, l_3 \parallel r = 1) = \left. \frac{p(l_1, l_2, l_3, r_1, r_2, r_3)}{p(r_1 | l_3^{(1)}, r_2) \times p(r_2 | l_1^{(1)}, r_3) \times p(r_3 | l_1^{(1)})} \right|_{r=1}. \tag{S1.1}$$

We can use the independencies encoded in the m-DAG and consistency in missing data models to identify the propensity score of R_1 as follows:

$$p(r_1 \mid \mathrm{pa}_{\mathcal{G}}(r_1))|_{r=1} = p(r_1 | l_3^{(1)}, r_2)|_{r=1} = p(r_1 = 1 | l_3, r_2 = 1, r_3 = 1). \tag{S1.2}$$

We cannot immediately obtain the propensity score of R_2, i.e., $p(r_2 \mid l_1^{(1)}, r_3)|_{r=1}$, since $R_2 \not\perp\!\!\!\perp R_1 \mid L_1^{(1)}, R_3$. This can still be identified using a total order where R_1 is intervened on before R_2.

Intervening on R_1 results in the following kernel that is Markov relative to the

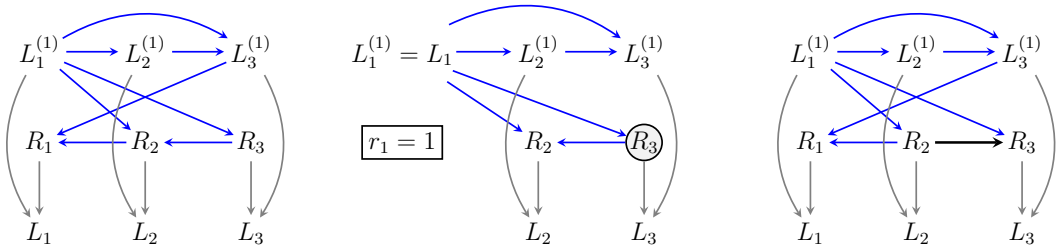


Figure 10: (a) Example of an m-DAG used to illustrate target law identification with odds ratio parameterization of the missingness selection model; (b) Graph derived from (a) representing an intervention on R_1 and the induced selection bias on R_3; (c) An m-DAG that is Markov equivalent to the m-DAG in (a).

graph in Fig. 10(b), with the induced selection bias on R_3.

$$p(l_2^{(1)}, l_3^{(1)}, l_1, l_2, l_3, r_2, r_3 \parallel r_1 = 1) = \left. \frac{p(l_1, l_2^{(1)}, l_3^{(1)}, l_2, l_3, r_1, r_2, r_3)}{p(r_1 \mid r_2, l_3^{(1)})} \right|_{r_1=1}$$

The propensity score of R_2 evaluated at $R_3 = 1$ is equivalent to $p(r_2 = 1 \mid r_3 = 1, l_1^{(1)} \parallel r_1 = 1)$. This is identified from the marginal kernel $p(l_1, l_3^{(1)}, r_2, r_3 = 1 \parallel r_1 = 1)$ which is equal to $p(l_1, l_3, r_1 = 1, r_2, r_3 = 1)/p(r_1 = 1 \mid r_2, l_3, r_3 = 1)$.

We now proceed to identify the propensity score of R_3, $p(r_3 \mid l_1^{(1)})|_{r=1}$, which is not immediately obvious since $R_3 \not\perp\!\!\!\perp R_1 \mid L_1^{(1)}$. Intervening on R_1 and setting it to 1 leads to a distribution where R_3 is necessarily selected on since the propensity score of R_1 is identified by restricting data to cases where $R_3 = 1$. Thus, we cannot identify the propensity score of R_3 in this post-intervention kernel distribution. A similar issue holds if we try to intervene on R_2 since identification of the propensity score of R_2 is obtained from a kernel distribution where we first intervene on R_1, which as mentioned introduces selection bias on R_3. It seems that we have exhausted all of our options based on the discussion of partial orders of identification. However, there is an alternative strategy that leads to identification of not just the target law,

but the full law as well.

Nabi et al. (2020) made the observation that the conditional density $p(r_3 \mid r_2, l_1^{(1)})$ is identified, since $R_3 \perp\!\!\!\perp R_1 \mid R_2, L_1^{(1)}$. From the preceding discussion, it is also clear that $p(r_2 \mid r_3 = 1, l_1^{(1)})$ is identified. Given that these conditional densities $p(r_2 \mid r_3 = 1, l_1^{(1)})$ and $p(r_3 \mid r_2, l_1^{(1)})$ are identified, they considered an odds ratio parameterization of the joint density $p(r_2, r_3 \mid \text{pa}_{\mathcal{G}}(r_2, r_3)) = p(r_2, r_3 \mid l_1^{(1)})$ as follows (Chen, 2007),

$$p(r_2, r_3 \mid l_1^{(1)}) = \frac{1}{Z} \times p(r_2 | r_3 = 1, l_1^{(1)}) \times p(r_3 | r_2 = 1, l_1^{(1)}) \times \text{OR}(r_2, r_3 | l_1^{(1)}),$$

where Z is the normalizing term, and

$$\text{OR}(r_2, r_3 \mid l_1^{(1)}) = \frac{p(r_3 \mid r_2, l_1^{(1)})}{p(r_3 = 1 \mid r_2, l_1^{(1)})} \times \frac{p(r_3 = 1 \mid r_2 = 1, l_1^{(1)})}{p(r_3 \mid r_2 = 1, l_1^{(1)})}.$$

All the terms in above parameterization are identified. This immediately implies the identifiability of the individual propensity scores for R_2 and R_3. This result, in addition to the fact that $p(r_1 \mid r_2, l_1^{(1)})$ is identified, leads to identification of both the target law and the full law, as the missingness process $p(r \mid l^{(1)})$ is also identified for all possible values of the missingness indicators. It is interesting to point out that the m-DAG in Fig. 10(a) is Markov equivalent to the one in Fig. 10(c), which means, the m-DAG model in both examples implies the same set of independence restrictions on the full data law. It is perhaps easier to see how identification in Fig. 10(c) proceeds using techniques discussed in the main draft – the target law is identified via parallel interventions on R_1 and R_3 followed by a sequential intervention on R_2. That is, identification can be obtained via the partial order $\{\{I_{r_1}, I_{r_3}\} < I_{r_3}\}$.

It is worthwhile pointing out that if we add the $L_2^{(1)} \rightarrow R_3$ edge to the m-DAG in Fig. 10, the full law and thus the target law remains identified via the odds ratio parameterization. However, the connection to how target law identification

can be carried out using the described methods in terms of finding partial orders of intervention remains an open problem.

S2. m-DAG Models with Unmeasured Confounders

Previous sections illustrated how identification may be accomplished in missing data models represented by a DAG where all variables are either fully or partially observed. However, just as in standard causal inference problems, most realistic missing data models include variables that are completely unobserved. We represent such models with an m-DAG $\mathcal{G}_m(L, R, L^{(1)}, U)$, where the vertex set U represents unobserved variables. By analogy with restrictions in Section 4, we require that $(L^{(1)} \cup U) \cap \{\text{de}_{\mathcal{G}_m}(R) \cup \text{de}_{\mathcal{G}_m}(L)\} = \emptyset$, i.e., there are no directed paths from any of the missingness indicators or proxy variables pointing towards variables in U or $L^{(1)}$. To clearly distinguish hidden variables from others in the model, we will render edges adjacent to such vertices in red.

In some m-DAGs with hidden variables, straightforward generalizations of identification strategies developed for m-DAGs without hidden variables can be developed. Consider the hidden variable m-DAG in Fig. 11 where U_1, U_2, and U_3 are completely unobserved. Although the joint over all variables in this model still factorizes with respect to this m-DAG, no factors containing unobserved variables in U can be used in identification or estimation strategies for the target $p(l^{(1)})$ or the selection mechanism $p(r \mid l^{(1)})$. Thus, in this setting it is useful to consider a factorization of the marginal model defined over variables that are either fully or partially observed. Recall that under any valid topological ordering on the variables, the ordered local Markov property simplifies each factor $p(v_i \mid \text{past}_{\mathcal{G}_m}(v_i))$ in the chain rule factorization to simply $p(v_i \mid \text{pa}_{\mathcal{G}_m}(v_i))$, as each variable is independent of its past (except parents) given its parents. We now describe an analogue of the ordered local Markov property and

factorization that relies only on partially or fully observed variables in the m-DAG, and demonstrate how this leads to an identification strategy.

Let $V = L^{(1)} \cup R \cup L$ denote the set of all partially and fully observed variables in $\mathcal{G}_m$. We define the *district* of $V_i \in V$ as the set of all variables $V_j \in V$ such that there exists a path connecting V_i and V_j that consists of only red edges, where any unmeasured variable $U_k \in U$ along the path is not a collider and any variable $V_k \in V$ along the path is a collider. We will use $\text{dis}_{\mathcal{G}_m}(V_i)$ to denote the district of V_i in $\mathcal{G}_m$; by convention $\text{dis}_{\mathcal{G}_m}(V_i)$ includes V_i itself. Given any valid topological order on all the variables in $\mathcal{G}_m$ (including unobserved variables) define $(\mathcal{G}_m)_{\overline{V_i}}$ to be the subgraph of $\mathcal{G}_m$ consisting of only the variables that appear before V_i in the topological order (including V_i itself) and the arrows present between these variables – not to be confused with alternative usage of the notation $\mathcal{G}_{\overline{X}}$ employed in the causal graph literature (Pearl, 2000) to represent a graph where incoming edges into X have been deleted. Then, the *Markov pillow* of V_i, denoted as $\text{mp}_{\mathcal{G}_m}(V_i)$, is defined as the district of V_i and the observed parents of the district of V_i (excluding V_i itself) in the subgraph $(\mathcal{G}_m)_{\overline{V_i}}$. That is, $\text{mp}_{\mathcal{G}_m}(V_i) \coloneqq \left\{ \text{dis}_{(\mathcal{G}_m)_{\overline{V_i}}}(V_i) \cup \text{pa}_{(\mathcal{G}_m)_{\overline{V_i}}}\left(\text{dis}_{(\mathcal{G}_m)_{\overline{V_i}}}(V_i) \right) \right\} \cap \left\{ V \setminus V_i \right\}$. We suppress the dependence of the definition of the Markov pillow on the topological order for notational simplicity. Given these definitions, we have the following independence relations among the observed variables in a hidden variable DAG that resemble the ordered local Markov property in fully observed DAGs (Tian and Pearl, 2002; Bhattacharya et al., 2020):

$$V_i \perp\!\!\!\perp \text{past}_{\mathcal{G}_m}(V_i) \cap V \setminus \text{mp}_{\mathcal{G}_m}(V_i) \mid \text{mp}_{\mathcal{G}_m}(V_i). \tag{S2.3}$$

That is, each variable is independent of its observed past given its Markov pillow. Using this observation we can simplify the chain rule factorization according to any

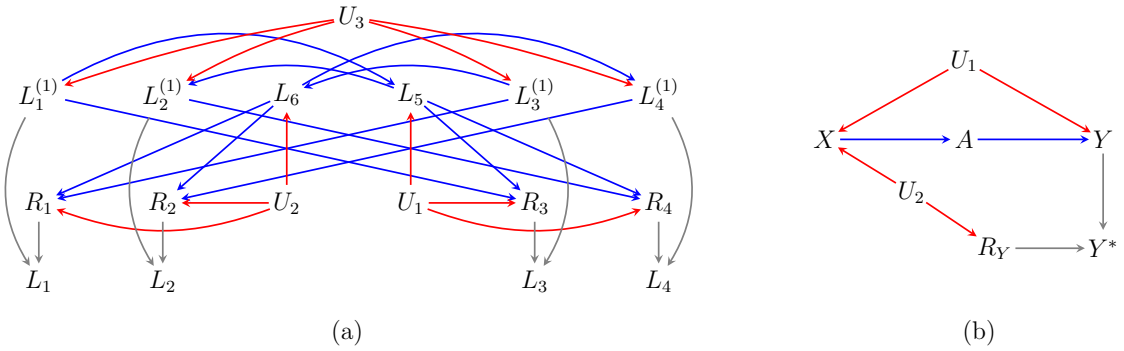


Figure 11: (a) A graph corresponding to a missing data model with hidden variables where identification of the law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}, l_4^{(1)}, l_5, l_6)$ is possible; (b) An example where the causal effect is identified but the target law is not identified.

valid topological order on the observed variables as,

$$p(v) = \prod_{v_k \in V} p(v_k \mid \text{past}_{\mathcal{G}_m}(v_k)) = \prod_{v_k \in V} p(v_k \mid \text{mp}_{\mathcal{G}_m}(v_k)). \tag{S2.4}$$

We now apply the above factorization to study the identification of the target laws $p(l^{(1)})$ in hidden variable m-DAGs. The properties of missing data graphs, as we described them, namely that for every L_k, $\text{pa}_{\mathcal{G}_m}(L_k) = \{L_k^{(1)}, R_k\}$, and $\{\text{de}_{\mathcal{G}_m}(R) \cup \text{de}_{\mathcal{G}_m}(L)\} \cap L^{(1)} = \emptyset$, implies that a version of the g-formula holds for identifying $p(l^{(1)})$ under topological orderings where variables in $R \cup L$ come after variables in $L^{(1)}$. That is, under a topological ordering defined on the partially and fully observed variables we have,

$$p(l^{(1)}) = \left. \frac{p(l, r)}{\prod_{r_k \in R} p(r_k \mid \text{past}_{\mathcal{G}_m}(r_k))} \right|_{r=1}, \tag{S2.5}$$

We have defined each $\text{past}_{\mathcal{G}_m}(r_k)$ here as containing only observed or missing (but not hidden) variables. However, though the above g-formula does not contain any hidden variables, it still may not necessarily yield identification, unless additional structure of the model can be exploited.

As an example, consider the model in Fig. 11. Fix a topological ordering U_1, U_2, U_3, $L_1^{(1)}, L_3^{(1)}, L_5, L_6, L_2^{(1)}, L_4^{(1)}, R_1, R_2, R_3, R_4, L_1, L_2, L_3, L_4$. Considering the subsequence of this ordering on just the observed variables the g-formula for $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}, l_4^{(1)}, l_5, l_6)$ is

$$\frac{p(l_1, l_2, l_3, l_4, l_5, l_6, r_1, r_2, r_3, r_4)}{p(r_1|\,\text{past}_{\mathcal{G}_m}(r_1)) \times p(r_2|\,\text{past}_{\mathcal{G}_m}(r_2)) \times p(r_3|\,\text{past}_{\mathcal{G}_m}(r_3)) \times p(r_4|\,\text{past}_{\mathcal{G}_m}(r_4))}\Bigg|_{r=1}.$$

Using the independence relations described in (S2.3), we have that each $p(r_k \mid \text{past}_{\mathcal{G}_m}(r_k))$ simplifies as $p(r_k \mid \text{mp}_{\mathcal{G}_m}(r_k))$. The propensity scores for each missingness indicator then simplifies under the proposed topological order on $\mathcal{G}_m$ as follows:

$$p(r_1 \mid \text{past}_{\mathcal{G}_m}(r_1))|_{r=1} = p(r_1 \mid l_3^{(1)}, l_6)|_{r=1} = p(r_1 = 1 \mid l_6, l_3, r_3 = 1),$$
$$p(r_2 \mid \text{past}_{\mathcal{G}_m}(r_2))|_{r=1} = p(r_2|l_3^{(1)}, l_4^{(1)}, l_6, r_1)|_{r=1} = p(r_2 = 1 \mid l_3, l_4, l_6, r_1 = r_3 = r_4 = 1),$$
$$p(r_3 \mid \text{past}_{\mathcal{G}_m}(r_3))|_{r=1} = p(r_3 \mid l_5, l_1^{(1)}, r_1)|_{r=1} = p(r_3 \mid l_5, l_1, r_1 = 1),$$
$$p(r_4 \mid \text{past}_{\mathcal{G}_m}(r_4))|_{r=1} = p(r_4 \mid l_1^{(1)}, l_2^{(1)}, l_5, r_3)|_{r=1} = p(r_4 \mid l_1, l_2, l_5, r_3 = r_1 = r_2 = 1).$$

Since these terms are all functions of observed data law, $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}, l_4^{(1)}, l_5,\ l_6)$ is identified.

S3. Identification of the Full Law

All of our examples so far have focused on identification of the target law, or equivalently the missingness mechanism evaluated at 1, i.e., $p(R = 1 \mid \text{pa}_{\mathcal{G}_m}(r))$. If identification of the full law is of interest (for instance for model selection purposes as in Gain and Shpitser (2018) and Tu et al. (2019)), the missingness mechanism $p(r \mid \text{pa}_{\mathcal{G}_m}(r))$, for all $r \in \{0,1\}^K$ must be identified. It is possible that in certain missing data DAG models, the target law is identified whereas the full law is not. For instance, in Fig. 6(a), $p(r_2 \mid R_1 = 0, l_1^{(1)})$ is not identified, and in Fig. 7(a), $p(r_1 \mid R_2 = 0, l_2^{(1)})$ is not identified, though the target law is identified in both cases.

Both examples have a special colluder structure $L_j^{(1)} \rightarrow R_i \leftarrow R_j$ in common. Bhattacharya et al. (2019) show that the presence of colluders in a graph always implies the full law of the corresponding missing data model is not identified.

Nabi et al. (2020) studied identification of the full law in missing data DAG models, and provided the first completeness result in a subclass of missing data DAGs where the proxy variables L are childless. They show the missingness mechanism $p(r \mid l^{(1)})$ that is Markov relative to a missing data DAG $\mathcal{G}_m$, where Ls are childless, is identified *if and only if* $\mathcal{G}_m$ does not contain self-censoring edges and colluders. The identification is given via an odds ratio parameterization (Chen, 2007) of the missingness mechanism. An example of identification with odds ratio parameterization is provided in Appendix S1. Nabi et al. (2020) drew an important connection between missing data models of a DAG $\mathcal{G}_m$ that are devoid of self-censoring and colluders, and the itemwise conditionally independent nonresponse (ICIN) model described in (Shpitser, 2016b; Sadinle and Reiter, 2017) (the ICIN model is referred to as the "no self-censoring" model in Shpitser (2016b)). As a substantive model, the ICIN model implies that no partially observed variable directly determines its own missingness, and is defined by the restrictions that for every pair $L_k^{(1)}, R_k$, it is the case that $L_k^{(1)} \perp\!\!\!\perp R_k \mid R_{-k}, L_{-k}^{(1)}$.

The no-self-censoring and no-colluder assumptions imply that $L_i^{(1)}$ is not in the *Markov blanket* of R_i, where the Markov blanket is defined as $\text{mb}_{\mathcal{G}_m}(V_i) = \text{pa}_{\mathcal{G}_m}(V_i) \cup \text{ch}_{\mathcal{G}_m}(V_i) \cup \text{pa}_{\mathcal{G}_m}(\text{ch}_{\mathcal{G}}(V_i))$. Given the local Markov property, $V_i \perp\!\!\!\perp V \setminus \text{mb}_{\mathcal{G}_m}(V_i) \mid \text{mb}_{\mathcal{G}_m}(V_i)$. If the full law is identified, then the target law is guaranteed to be identified. For instance, since there is no self-censoring edges or colluder structures in Figs. 4(a) and 5(a), we can immediately conclude that the full law and hence the target law are identified. However, the reverse is not necessarily true – that is if the full law is not identified (due to presence of colluders or self-censoring edges), the target

law might still be identified as discussed in examples related to Figs. 6(a) and 7(a). Nabi et al. (2020) generalized this theory to scenarios where some variables are not just missing, but completely unobserved. They proposed necessary and sufficient graphical conditions that must hold in a missing data DAG model with unmeasured confounders to permit identification of the full law. They defined a *colluding path* between $L_k^{(1)}$ and R_k as a path where every collider is a variable in $L^{(1)} \cup R$ and every non-collider is a variable in U. They showed that in the absence of such paths, the odds ratio parameterization can be used to identify the full law, while their presence results in non-identification.

Often, instead of identifying the entire full law or target law, we might simply be interested in a simple outcome mean or a causal effect. There are plenty of examples where such parameters are indeed identified, but the underlying joint distribution is not. For instance, consider the graph in Fig. 11(b), which is discussed in Mohan and Pearl (2021). The outcome is missing due to a common unmeasured confounder with pre-treatment variables X. The causal effect of A on Y here is indeed identified, even though the target law is not identified. Briefly, the target law is not identified due to the presence of a colluding path between $Y^{(1)}$ and R_Y, which prevents identification of $p(y^{(1)} \mid a, x)$ (Mohan and Pearl, 2021; Nabi et al., 2020). However, the model encodes the following independence restrictions which enable identification of the causal effect: $Y^{(a,R_Y=1)} \perp\!\!\!\perp R_Y$ and $Y^{(a,R_Y=1)} \perp\!\!\!\perp A \mid X, R_Y$, where $Y^{(a,R_Y=1)}$ denotes the potential outcome when A is set to some value a and had we, in fact, been able to observe it. Such counterfactual independence restrictions are often read using d-separation rules applied to single-world intervention graphs (SWIGs). A detailed discussion on how missing data graphical models which contain counterfactuals relate to SWIGs is left to future work. The counterfactual distribution $p(y^{(a,R_Y=1)})$ is identified as:

$$p(y^{(a,R_Y=1)}) = p(y^{(a,R_Y=1)} \mid R_Y = 1)$$

$$= \sum_x p(y^{(a,R_Y=1)} \mid x, R_Y = 1) \times p(x \mid R_Y = 1)$$
$$= \sum_x p(y^{(a,R_Y=1)} \mid x, A = a, R_Y = 1) \times p(x \mid R_Y = 1)$$
$$= \sum_x p(y \mid x, A = a, R_Y = 1) \times p(x \mid R_Y = 1).$$

The first equality follows from $Y^{(a,R_Y=1)} \perp\!\!\!\perp R_Y$, the second from rules of probability, the third from $Y^{(a,R_Y=1)} \perp\!\!\!\perp A \mid X, R_Y$, and the final equality follows from consistency.

S4. Additional Results and Proofs

S4.1 Proposition 1

Proof. Since $\mathrm{de}_{\mathcal{G}}(R_i) \cap L^{(1)}$, the vertex set $L^{(1)}$ is ancestral in $\mathcal{G}_m$ This implies $p(l^{(1)})$ is equal to

$$\sum_{r \cup l} p(l, r, l^{(1)}) = \sum_{r \cup l} \big(\prod_{v_k \in r \cup l} p(v_k \mid \mathrm{pa}_{\mathcal{G}_m}(v_k)) \big) \times \big(\prod_{v_k \in l^{(1)}} p(v_k \mid \mathrm{pa}_{\mathcal{G}_m}(v_k)) \big) = \prod_{v_k \in l^{(1)}} p(v_k \mid \mathrm{pa}_{\mathcal{G}_m}(v_k)).$$

Further, using Bayes rule, we conclude the second equality in (4.6) by noting that $p(l, r, l^{(1)})|_{r=1} = p(l, r)|_{r=1}$ (by consistency) and

$$p(r, l \mid l^{(1)})|_{r=1} = \prod_{l_k \in l} p(l_k \mid r_k = 1, l_k^{(1)}) \times \prod_{r_k \in r} p(r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k))|_{r=1} = \prod_{r_k \in r} p(r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k))|_{r=1}.$$

□

S4.2 Lemma 1

Proof. Given a set $R^* \subseteq R$ and the corresponding set of counterfactuals $L^{*(1)}$, the distribution

$$p(l^{(1)} \setminus l^{*(1)}, r \setminus r^*, l \parallel r^* = 1) := \frac{p(l^{(1)} \setminus l^{*(1)}, r, l)}{\prod_{r_k \in r^*} p(r_k \mid \mathrm{pa}_{\mathcal{G}_m}(r_k))} \Bigg|_{R^*=1}$$

factorizes with respect to a *conditional DAG* (CDAG) $\widetilde{\mathcal{G}}_m(L^{(1)} \cup \{R \setminus R^*\} \cup L, R^*)$, which is a DAG containing random vertices $L^{(1)} \cup \{R \setminus R^*\} \cup L$ and fixed vertices R^* with the property that all fixed vertices can only have outgoing directed edges. $\widetilde{\mathcal{G}}_m$ is constructed from original m-DAG $\mathcal{G}_m$ by removing all edges with arrowheads into R^*, marking R^* as fixed vertices, and treating each $L_k^{(1)} \in L^{*(1)}$ as equivalent to its corresponding proxy (by consistency). The CDAG factorization of any $p(v \parallel w)$ with respect to a CDAG $\mathcal{G}(V, W)$ is a straightforward generalization of the DAG factorization:

$$p(v \parallel w) = \prod_{v_i \in v} p(v_i \mid \mathrm{pa}_{\mathcal{G}}(v_i) \setminus w \parallel \mathrm{pa}_{\mathcal{G}}(v_i) \cap w),$$

where conditioning in $p(v \parallel w)$ is defined as in (2.3).

If $p(v \parallel w)$ factorizes with respect to $\mathcal{G}(V, W)$, it obeys the local Markov property which states that for each variable V_i, the distribution $p(v_i \mid \mathrm{past}_{\mathcal{G}}(v_i) \backslash w \parallel w \cap \mathrm{pa}_{\mathcal{G}}(v_i))$ is only a function of V_i and its direct causes $\mathrm{pa}_{\mathcal{G}}(V_i)$. This immediately implies the conclusion, since the kernel $p(l^{(1)} \setminus l^{*(1)}, r \setminus r^*, l \parallel r^* = 1)$ factorizes according to the CDAG $\widetilde{\mathcal{G}}_m(L^{(1)} \cup \{R \setminus R^*\} \cup L, R^*)$ where all direct causes of each $R_k \notin R^*$ are preserved. See Richardson et al. (2017) for more details on conditional DAG factorization.

□

S4.3 Identification under rank preservation in a K variable block-parallel model

In Section 6 we saw how missing data identification strategies can be applied in conjunction with additional assumptions, such as rank preservation, to attain identification in a two variable causal analogue of the block-parallel missing data model. The following theorem shows how this applies to any causal model endowed with rank

preservation that is analogous to a K variable block-parallel model (as well as any sub models of it). For simplicity, we will assume all treatment variables are binary though the result trivially extends to non-binary treatments.

Theorem 1. *Given a causal model that encodes the following independence restrictions: for each* $k \in \{1, \ldots, K\}$

$$A_k \perp\!\!\!\perp L_k^{(0)}, L_k^{(1)}, A_{-k} \mid L_{-k}^{(0)}, L_{-k}^{(1)},$$

and the following rank preservation assumptions: for each $k \in \{1, \ldots, K\}$ *and* $j = \{0, 1\}$ *there exists a bijection* g_k *such that* $L_k^{(1-j)} = g_k(L_k^{(j)})$. *The counterfactual distribution* $p(l_1^{(a_1)}, \ldots l_2^{(a_K)})$, *where each* $a_k \in \{0, 1\}$, *is identified and given by the following functional:*

$$\frac{p(l_1, \ldots, l_K, A_1 = a_1, \ldots, A_K = a_K)}{\prod_{A_k \in A} p(A_k = a_k \mid l_{-k}, A_{-k} = a_{-k})}.$$

Proof. The counterfactual distribution $p(l_1^{(a_1)}, \ldots, l_k^{(a_K)})$ is identified via the following identities following a very similar strategy to the one used in the main text. In the following we will use $l^{(a)}$ and $l^{(1-a)}$ as short hand for $l_1^{(a_1)}, \ldots, l_K^{(a_K)}$ and $l_1^{(1-a_1)}, \ldots, l_K^{(1-a_K)}$ respectively:

$$\begin{aligned}
p(l^{(a)}) &= \frac{p(l_1^{(a_1)}, \ldots, l_K^{(a_K)}, A_1 = a_1, \ldots, A_K = a_K)}{p(A_1 = a_1, \ldots, A_K = a_K \mid l_1^{(a_1)}, \ldots, l_K^{(a_K)})} \\
&= \frac{p(l^{(a)}, A_1 = a_1, \ldots, A_K = a_K)}{p(A_1 = a_1, \ldots, A_K = a_K \mid l^{(a)})} \\
&= \frac{p(l^{(a)}, A_1 = a_1, \ldots, A_K = a_K)}{\sum_{l^{(1-a)}} p(A_1 = a_1, \ldots, A_K = a_K \mid l^{(a)}, l^{(1-a)}) \times p(l^{(1-a)} \mid l^{(a)})} \\
&= \frac{p(l^{(a)}, A_1 = a_1, \ldots, A_K = a_K)}{\sum_{l^{(1-a)}} \prod_{A_k \in A} p(A_k = a_k \mid l^{(a)}, l^{(1-a)}, A_{-k} = a_{-k}) \times p(l^{(1-a)} \mid l^{(a)})} \\
&= \Bigg\{ \sum_{l^{(1-a)}} \prod_{A_k \in A} p(A_k = a_k | l^{(a)}, l^{(1-a)}, A_{-k} = a_{-k}) \times p(l^{(1-a)} | l^{(a)})
\end{aligned}$$

$$\times \mathbb{I}(l_1^{(1-a_1)} = g_1(l_1^{(a_1)}), \ldots, l_K^{(1-a_K)} = g_K(l_K^{(a_K)}))\Big\}^{-1}$$

$$\times\, p(l^{(a)}, A_1 = a_1, \ldots, A_K = a_K)$$

$$= \frac{p(l^{(a)}, A_1 = a_1, \ldots, A_K = a_K)}{\prod_{A_k \in A} p(A_k = a_k \mid l_{-k}^{(a)}, A_{-k} = a_{-k})}$$

$$= \frac{p(l_1, \ldots, l_K, A_1 = a_1, \ldots, A_K = a_K)}{\prod_{A_k \in A} p(A_k = a_k \mid l_{-k}, A_{-k} = a_{-k})}.$$

The first equality follows from Bayes rule, the second from our notational convention, the third from rules of probability, the fourth from applying the chain rule of factorization and using the independence restrictions implied by the model, the fifth and sixth from rank preservation as well as restrictions encoded by the model, and finally, the last equality follows from consistency. □

References

Bhattacharya, R., R. Nabi, and I. Shpitser (2020). Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*.

Bhattacharya, R., R. Nabi, I. Shpitser, and J. Robins (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.

Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. *biometrics 63*, 413–421.

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

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Ding, P. and F. Li (2018). Causal inference: A missing data perspective. *Statistical Science 33*, 214–237.

Duarte, G., N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser (2021). An automated approach to causal inference in discrete settings. *arXiv preprint arXiv:2109.13471*.

Gain, A. and I. Shpitser (2018). Structure learning under missing data. In *Proceedings of the 9th International Conference on Probabilistic Graphical Models (PGM-2018)*.

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

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.

Horton, N. J. and N. M. Laird (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research 8*(1), 37–50.

Huang, Y. and M. Valtorta (2006). Pearl's calculus of interventions is complete. In *Twenty Second Conference On Uncertainty in Artificial Intelligence*.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, U.K.: Clarendon.

Li, L., C. Shen, X. Li, and J. M. Robins (2013). On weighting approaches for missing data. *Statistical methods in medical research 22*(1), 14–30.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association 88*(421), 125–134.

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

Malinsky, D., I. Shpitser, and E. J. Tchetgen Tchetgen (2021). Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, 1–9.

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

Mattei, A., F. Mealli, and B. Pacini (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics 70*(2), 278–288.

McKnight, P. E., K. M. McKnight, S. Sidani, and A. J. Figueredo (2007). *Missing data: A gentle introduction*. Guilford Press.

Miao, W., P. Ding, and Z. Geng (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association 111*(516), 1673–1683.

Miao, W. and E. J. Tchetgen Tchetgen (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika 103*(2), 475–482.

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

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

Moreno-Betancur, M. and M. Chavance (2016). Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. *Statistical Methods in Medical Research 25*(4), 1471–1489.

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

Nabi, R., R. Bhattacharya, and I. Shpitser (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)*.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2 ed.). Cambridge University Press.

Richardson, T. S., R. J. Evans, J. M. Robins, and I. Shpitser (2017). Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*.

Robins, J. M. (1986). 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.

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. and A. Rotnitzky (2001). Comment on inference for semiparametric models: Some questions and an answer, by PJ Bickel and J. Kwon. *Statistica Sinica 11*, 920–936.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000). Sensitivity analysis for selec-

tion bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pp. 1–94. Springer.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (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., J. M. Robins, and D. O. Scharfstein (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.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons.

Saadati, M. and J. Tian (2019). Adjustment criteria for recovering causal effects from missing data. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Sadinle, M. and J. P. Reiter (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika 104*(1), 207–220.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research 8*(1), 3–15.

Schafer, J. L. and J. W. Graham (2002). Missing data: our view of the state of the art. *Psychological methods 7*(2), 147.

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

Scharfstein, D. O., J. J. Lee, A. McDermott, A. Campbell, E. Nunes, A. G. Matthews, and I. Shpitser (2021). Markov-restricted analysis of randomized trials with non-monotone missing binary outcomes: Sensitivity analysis and identification results. *arXiv preprint arXiv:2105.08868*.

Scharfstein, D. O., R. Nabi, E. H. Kennedy, M.-Y. Huang, M. Bonvini, and M. Smid (2021). Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *Journal of the American Statistical Association 94*, 1096–1146.

Shpitser, I. (2016a). Consistent estimation of functions of data missing non-monotonically and not at random. *Advances in Neural Information Processing Systems 29*, 3144–3152.

Shpitser, I. (2016b). 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., K. Mohan, and J. Pearl (2015). Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pp. 802–811. AUAI Press.

Shpitser, I. and J. Pearl (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.

Shpitser, I. and E. Sherman (2018). Identification of personalized effects associated with

causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*.

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

Tchetgen, E. J. T. (2009). Commentary: A simple implementation of doubly robust estimation in logistic regression with covariates missing at random. *Epidemiology 20*(3), 391–394.

Tchetgen, E. J. T., L. Wang, and B. Sun (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica 28*(4), 2069.

Thoemmes, F. and N. Rose (2013). Selection of auxiliary variables in missing data problems: Not all auxiliary variables are created equal. Technical report, R-002, Cornell University.

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

Tian, J. (2015). Missing at random in graphical models. In *Artificial Intelligence and Statistics*, pp. 977–985. PMLR.

Tian, J. and J. Pearl (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, Volume 18, pp. 519–527. AUAI Press, Corvallis, Oregon.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data* (1st edition ed.). Springer-Verlag New York.

Tu, R., C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, and K. Zhang (2019). Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1762–1770. PMLR.

Vansteelandt, S., A. Rotnitzky, and J. Robins (2007). Estimation of regression models

for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika 94*(4), 841–860.

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

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

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