CAUSAL AND COUNTERFACTUAL VIEWS OF MISSING DATA MODELS

Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, James M. Robins

Emory University, Williams College,

Johns Hopkins University, Harvard University

Supplementary Material

The supplementary materials are organized as follows. Section S1 describes identification using an odds ratio parameterization, different from m-DAG factorization. Section S2 extends the concepts of m-DAGs to m-DAGs with hidden variables. Section S3 expands the target law identification arguments in the manuscript to full law identification. Section S4 contains all our proofs.

S1 Identification via an 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 in some cases via an odds ratio parameterization of conditional distributions (Chen, 2007; Shpitser, 2023).

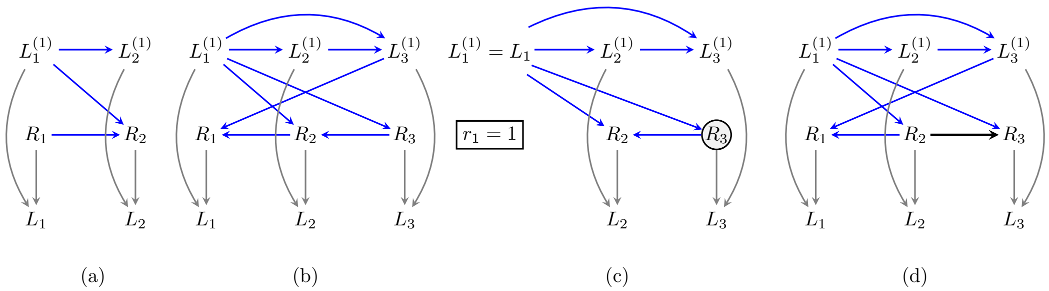
This parameterization yields a sound and complete algorithm for the identification of the full law $p(r, l^{(1)})$ in graphical missing data models, but does not yield a complete algorithm for identification of the target law $p(l^{(1)})$. A simple example of a model where the target law is identified but the full law is not is shown in Fig. 1(a). In fact, deriving a sound and complete algorithm for the identification of the target law by *any* method – whether by the parameterization described below, or methods based on the g-formula described in the main draft – is currently an open problem.

Consider the m-DAG in Fig. 1(b). 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:

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

We can use the independencies encoded in the m-DAG and consistency in



(a) (b) (c) (d)

Figure 1: (a) A simple model where the target law $p(l_1^{(1)}, l_2^{(1)})$ is identified, but the full law $p(l_1^{(1)}, l_2^{(1)}, r_1, r_2)$ is not; (b) Example of an m-DAG used to illustrate target law identification with odds ratio parameterization of the missingness selection model; (c) Graph derived from (b) representing an intervention on R_1 and the induced selection bias on R_3; (d) An m-DAG that is Markov equivalent to the m-DAG in (b).

missing data models to identify the propensity score of R_1 as follows:

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

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 graph in Fig. 1(c), 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 probability distribution $p(r_3 \mid r_2, l_1^{(1)})$ 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 probability distribution $p(r_2, r_3 \mid \text{pa}_{\mathcal{G}}(r_2, r_3)) = p(r_2, r_3 \mid l_1^{(1)})$ as follows

(Chen, 2007),

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

where Z is the normalizing term, and

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

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. 1(b) is Markov equivalent to the one in Fig. 1(d), 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. 1(d) 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 $\{\{I_{r_1}, I_{r_3}\} < I_{r_3}\}$.

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. 2 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 $p(v_i \mid \text{past}_{\prec}(v_i))$ 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, $\text{mp}_{\mathcal{G}_m}(V_i) := \big\{ \text{dis}_{(\mathcal{G}_m)_{\overline{V_i}}}(V_i) \cup \text{pa}_{(\mathcal{G}_m)_{\overline{V_i}}}\big(\text{dis}_{(\mathcal{G}_m)_{\overline{V_i}}}(V_i)\big)\big\} \cap \big\{V \setminus V_i\big\}$. 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., 2022):

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

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 valid topological order on the observed variables as,

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

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

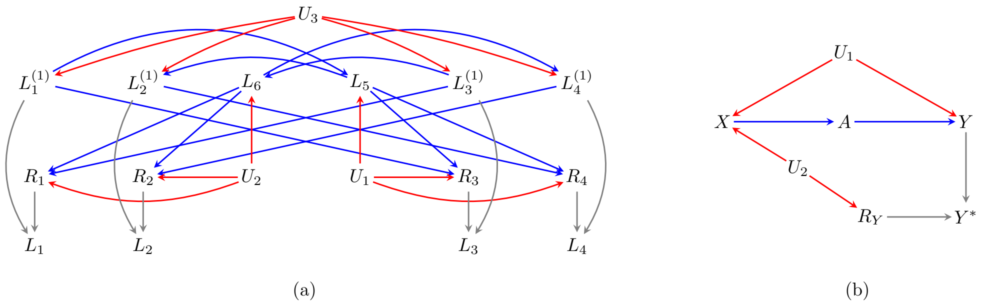


Figure 2: (a) A graph corresponding to a missing data model with hidden variables where identification of the law $p(l_1^{(1)}, l_2^{(1)}, l_3^{(1)}, l_4^{(1)}, l_5, l_6)$ is possible; (b) An example where the causal effect is identified but the target law is not identified.

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,

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

We have defined each $\text{past}_{\prec}(r_k)$ 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. 2. 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

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

Using the independence relations described in (S2.3), we have that each $p(r_k \mid \text{past}_{\prec}(r_k))$ 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:

$$p(r_1 \mid \text{past}_{\prec}(r_1))|_{r=1} = p(r_1 \mid l_3^{(1)}, l_6)|_{r=1} = p(r_1 = 1 \mid l_6, l_3, r_3 = 1),$$

$$p(r_2 \mid \text{past}_{\prec}(r_2))|_{r=1} = p(r_2 \mid l_3^{(1)}, l_4^{(1)}, l_6, r_1)|_{r=1} = p(r_2 = 1 \mid l_3, l_4, l_6, r_1 = 1, r_3 = 1, r_4 = 1),$$

$$p(r_3 \mid \text{past}_{\prec}(r_3))|_{r=1} = p(r_3 \mid l_5, l_1^{(1)}, r_1)|_{r=1} = p(r_3 \mid l_5, l_1, r_1 = 1),$$

$$p(r_4 \mid \text{past}_{\prec}(r_4))|_{r=1} = p(r_4 \mid l_1^{(1)}, l_2^{(1)}, l_5, r_3)|_{r=1} = p(r_4 \mid l_1, l_2, l_5, r_3 = 1, r_1 = 1, r_2 = 1).$$

Since these terms are all functions of the 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 the model shown in Fig. 6(a) in the main draft, $p(r_2 \mid R_1 = 0, l_1^{(1)})$ is not identified, and in the model in Fig. 7(a) in the main body of the draft, $p(r_1 \mid R_2 = 0, l_2^{(1)})$

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, 2016; Sadinle and Reiter, 2017) (the ICIN model is referred to as the "no self-censoring" model in Shpitser (2016)). 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. 2(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:

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

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

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

Further, using Bayes rule, we conclude the second equality in (4.8) by noting that $p(l, r, l^{(1)})|_{r=1} = p(l, r)|_{r=1}$ (by consistency) and

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

□

S4.2 Lemma 1

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

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

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:

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

where conditioning in $p(v \parallel w)$ is defined as in (3.5).

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 $p(v_i \mid \text{past}_{\mathcal{G}}(v_i) \setminus w \parallel w \cap \text{pa}_{\mathcal{G}}(v_i))$ is only a function of V_i and its direct causes $\text{pa}_{\mathcal{G}}(V_i)$. 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. (2023) 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 iden-*

tified 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:

$$\begin{aligned}
p(l^{(a)}) &= \frac{p(l_1^{(a_1)},\ldots,l_K^{(a_K)},A_1=a_1,\ldots,A_K=a_K)}{p(A_1=a_1,\ldots,A_K=a_K \mid l_1^{(a_1)},\ldots,l_K^{(a_K)})}\\
&= \frac{p(l^{(a)},A_1=a_1,\ldots,A_K=a_K)}{p(A_1=a_1,\ldots,A_K=a_K \mid l^{(a)})}\\
&= \frac{p(l^{(a)},A_1=a_1,\ldots,A_K=a_K)}{\sum\limits_{l^{(1-a)}} p(A_1=a_1,\ldots,A_K=a_K \mid l^{(a)},l^{(1-a)}) \times p(l^{(1-a)} \mid l^{(a)})}\\
&= \frac{p(l^{(a)},A_1=a_1,\ldots,A_K=a_K)}{\sum\limits_{l^{(1-a)}} \prod\limits_{A_k\in A} p(A_k=a_k \mid l^{(a)},l^{(1-a)},A_{-k}=a_{-k}) \times p(l^{(1-a)} \mid l^{(a)})}\\
&= \Bigg\{ \sum_{l^{(1-a)}} \prod_{A_k\in A} p(A_k=a_k | l^{(a)},l^{(1-a)},A_{-k}=a_{-k}) \times p(l^{(1-a)}|l^{(a)})\\
&\qquad \times \mathbb{I}(l_1^{(1-a_1)}=g_1(l_1^{(a_1)}),\ldots,l_K^{(1-a_K)}=g_K(l_K^{(a_K)})) \Bigg\}^{-1}\\
&\qquad\qquad \times p(l^{(a)},A_1=a_1,\ldots,A_K=a_K)\\
&= \frac{p(l^{(a)},A_1=a_1,\ldots,A_K=a_K)}{\prod\limits_{A_k\in A} p(A_k=a_k \mid l_{-k}^{(a)},A_{-k}=a_{-k})}\\
&= \frac{p(l_1,\ldots,l_K,A_1=a_1,\ldots,A_K=a_K)}{\prod\limits_{A_k\in A} p(A_k=a_k \mid l_{-k},A_{-k}=a_{-k})}.
\end{aligned}$$

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. □

Bibliography

Bhattacharya, R., R. Nabi, and I. Shpitser (2022). Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research 23*(295), 1–76.

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.

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

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

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.

Richardson, T. S., R. J. Evans, J. M. Robins, and I. Shpitser (2023). Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics 51*(1), 334–361.

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

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

Shpitser, I. (2023). The Lauritzen-Chen likelihood. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*.

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.

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.