Identification In Missing Data Models Represented By Directed Acylic Graphs

Rohit Bhattacharya[†*], Razieh Nabi[†*], Ilya Shpitser[†], James M. Robins[‡]
{*rbhattacharya@*, *rnabi@*, *ilyas@cs.*}jhu.edu, *robins@hsph.harvard.edu*

[*] Equal contribution; [†] Dept. of Computer Science, Johns Hopkins University; [‡] Harvard T. H. Chan School of Public Health

Motivation

- Many popular missing data models can be expressed as factorizations according to a DAG.
- Recent work [2, 4] proposed identification strategies for these models based on causal inference methods.
- We show that these methods are unable to identify a large space of identifiable target distributions. We propose, and illustrate via examples, a new method that fixes based on a partial order, uses selection bias on missingness, and treats missing variables as hidden.

Missing Data Models of a DAG

- **Target law** $p(\mathbf{X}^{(1)}, \mathbf{O})$ over
  - Potentially missing random variables $\{X_1^{(1)} \ldots, X_k^{(1)}\}$
  - Observed random variables $\{O_1, \ldots, O_m\}$.
- **Nuisance law** $p(\mathbf{R}|\mathbf{X}^{(1)}, \mathbf{O})$ over
  - Missingness indicators $\mathbf{R} \equiv \{R_1, \ldots, R_k\}$.
- **Deterministic factors** $p(\mathbf{X}|\mathbf{X}^{(1)}, \mathbf{R})$
  - $X_i \equiv X_i^{(1)}$ if $R_i = 1$ and $X_i \equiv ?$ if $R_i = 0$.
- Missing data models of a DAG $\mathcal{G}$

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

- By chain rule of probability,

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

Fixability And Fixing In Causal Inference

- Consider a graph $\mathcal{G}$ with random variables $\mathbf{V}$, fixed variables $\mathbf{W}$
- $V \in \mathbf{V}$ is **fixable** if $\text{de}_{\mathcal{G}}(V) \cap \text{dis}_{\mathcal{G}}(V) = \{V\}$
- Graphical fixing operator $\phi_V(\mathcal{G}) \equiv$ CADMG $\mathcal{G}'(\mathbf{V} \setminus \{V\} | \mathbf{W} \cup \{V\})$ with edges into V removed.
- Probabilistic fixing operator $\phi_V(q_{\mathbf{V}}; \mathcal{G})$ yields a new kernel

$$q'_{\mathbf{V} \setminus \{V\}}(\mathbf{V} \setminus \{V\}, \mathbf{W} \cup \{V\}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(V | \text{mb}_{\mathcal{G}}(V), \mathbf{W})}.$$

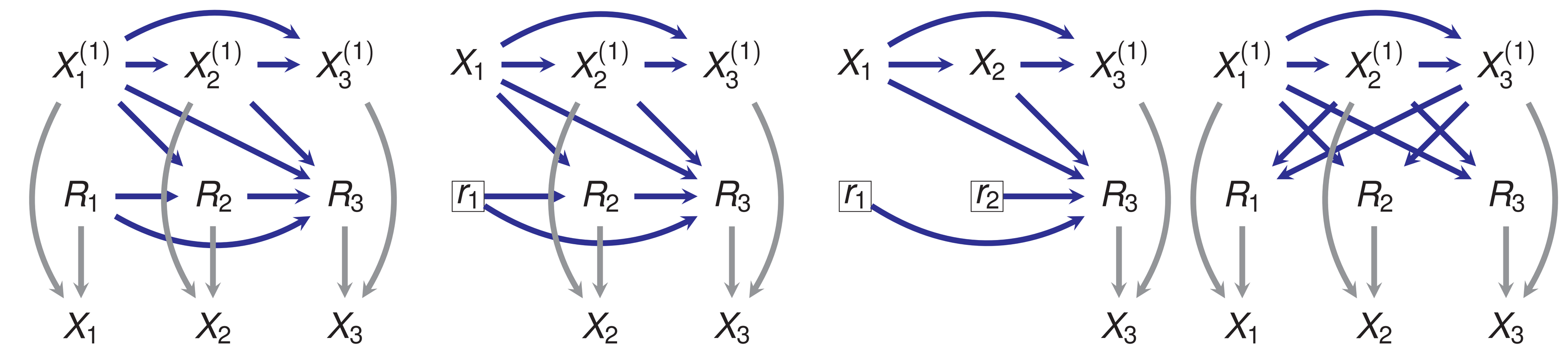
Fixability And Fixing In Missing Data

- For $\mathbf{Z} \subseteq \mathbf{D}_{\mathbf{Z}} \in \mathcal{D}(\mathcal{G})$, let $\mathbf{R}_{\mathbf{Z}} = \{R_j | X_j^{(1)} \in \mathbf{Z} \cup \text{mb}_{\mathcal{G}}(\mathbf{Z}), R_j \notin \mathbf{Z}\}$, and $\text{mb}_{\mathcal{G}}(\mathbf{Z}) \equiv (\mathbf{D}_{\mathbf{Z}} \cup \text{pa}_{\mathcal{G}}(\mathbf{D}_{\mathbf{Z}})) \setminus \mathbf{Z}$. We say $\mathbf{Z}$ is fixable in $\mathcal{G}(\mathbf{V} \setminus \mathbf{X}_{\mathbf{U}}^{(1)}, \mathbf{W})$ if
  - $\text{de}_{\mathcal{G}}(\mathbf{Z}) \cap \mathbf{D}_{\mathbf{Z}} \subseteq \mathbf{Z}$,
  - $\mathbf{S} \cap \mathbf{Z} = \emptyset$, where $\mathbf{S}$ are selected variables,
  - $\mathbf{Z} \perp\!\!\!\perp (\mathbf{S} \cup \mathbf{R}_{\mathbf{Z}}) \setminus \text{mb}_{\mathcal{G}}(\mathbf{Z}) | \text{mb}_{\mathcal{G}}(\mathbf{Z})$.

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

Sequential And Parallel Fixing

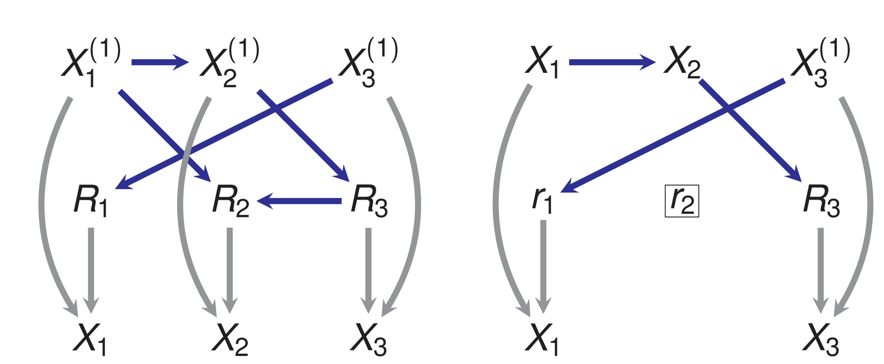
- (a) and (d) are examples of DAGs where existing theory is sufficient for identification of the target law.



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

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

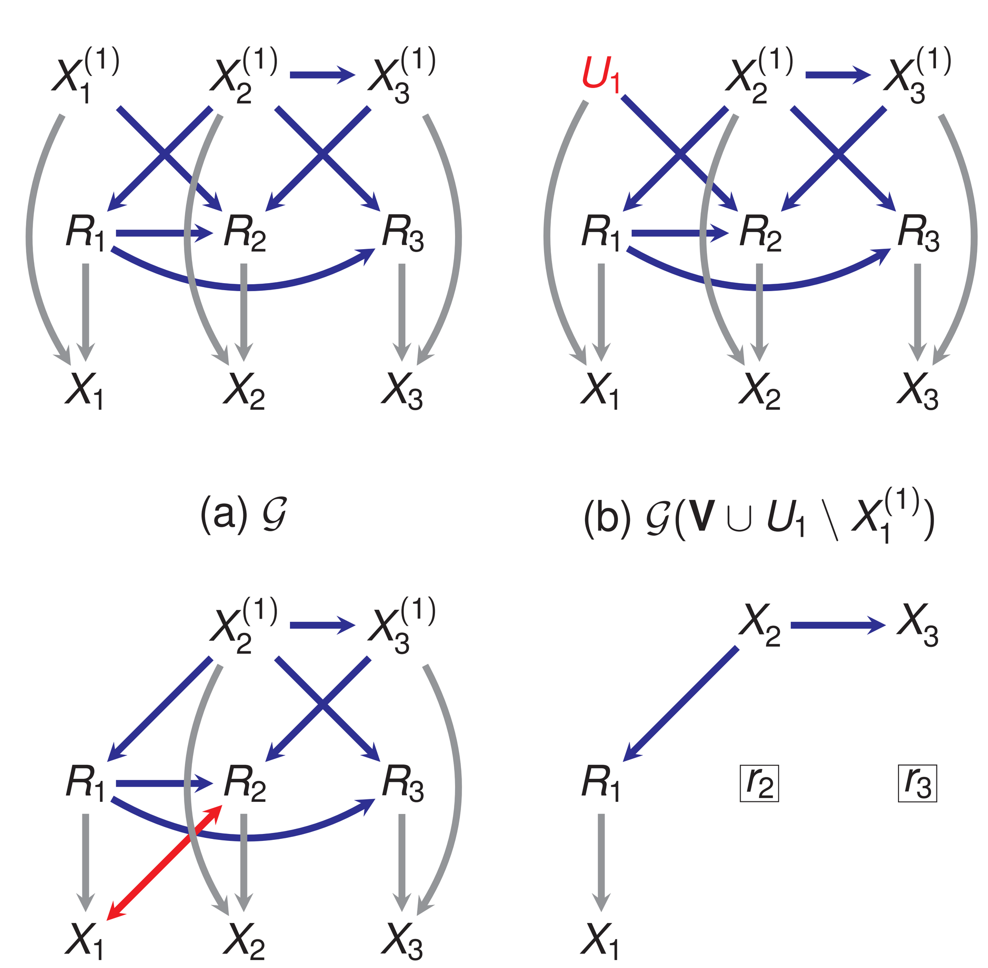
- The target law in (a) is obtained by fixing on a partial order where R_1, R_2 are incompatible and $R_2 \prec R_3$.



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

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

Dodging Selection Bias

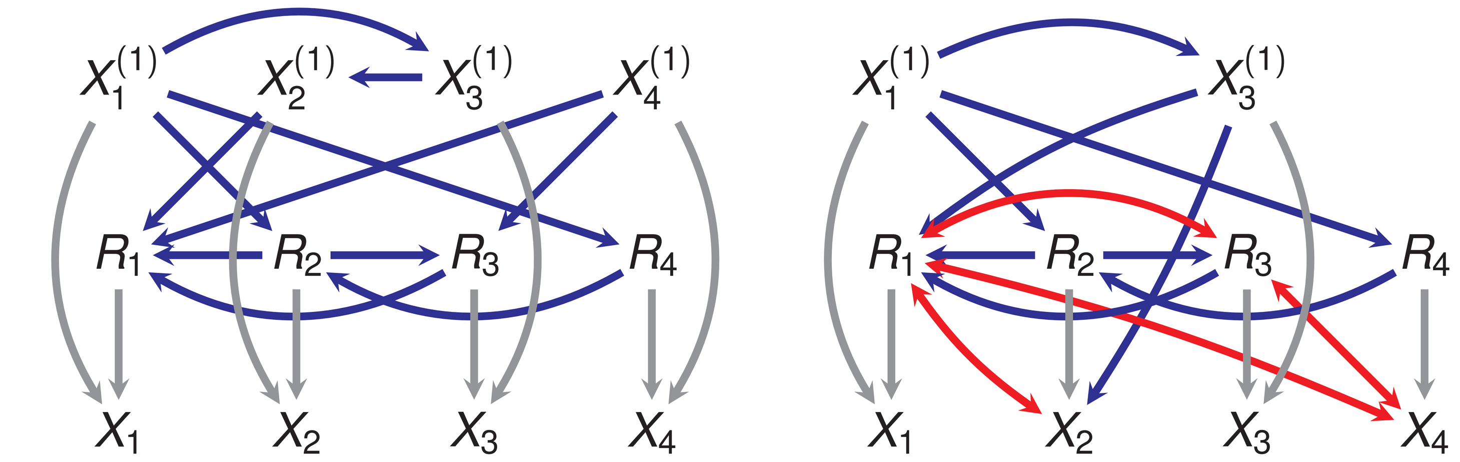


(a) $\mathcal{G}$ (b) $\mathcal{G}(\mathbf{V} \cup U_1 \setminus X_1^{(1)})$

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

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

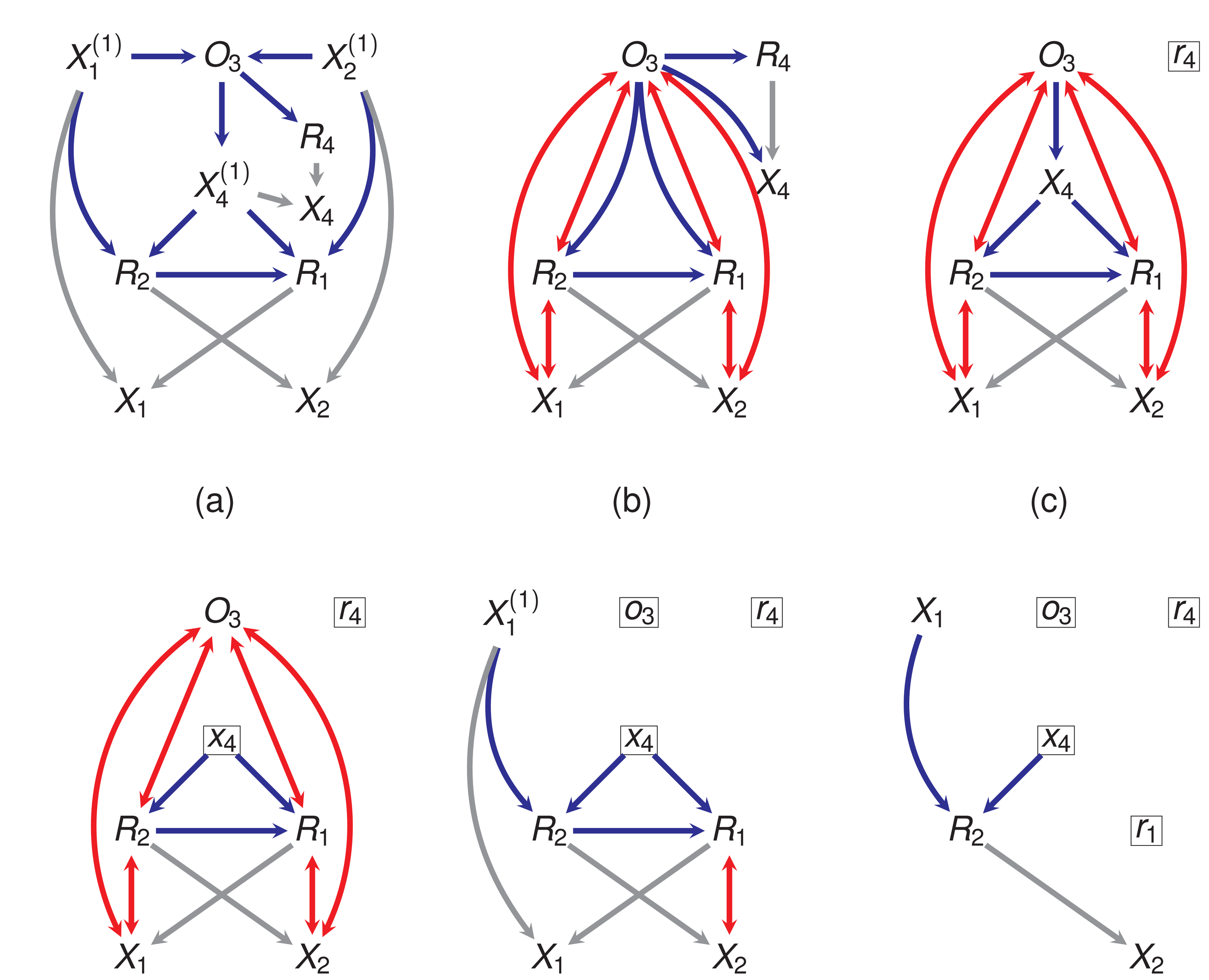
Fixing Sets Of Variables



(a) (b)

Figure: (a) A DAG where the fixing operator must be performed on a set of vertices. (b) A latent projection of a subproblem used for identification of $p(R_4 | X_4^{(1)})$.

Fixing Variables Other Than Rs



(a) (b) (c)

(d) (e) (f)

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

Future Work

- Is the algorithm complete?
- Is there a polynomial time formulation?

References

[1] Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M. Robins. **Identification in missing data models represented by directed acyclic graphs**. In *UAI*, 2019. forthcoming.

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

[3] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

[4] Ilya Shpitser, Karthika Mohan, and Judea Pearl. Missing data as a causal and probabilistic problem. In *UAI*, 2015.