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

Anna Guo and Razieh Nabi

Dept. of Biostatistics & Bioinformatics, Emory University

EMORY UNIVERSITY

Motivation

- Widely used missing data tools, such as MICE and Amelia, typically rely on the missing-at-random (MAR) assumption.
- Missing-not-at-random (MNAR) mechanism is common in practice, complicating identification and valid analysis.
- A practical framework is needed to unify identification, estimation, and inference over diverse missing data models, including MNAR models.

Missing Data DAG Models

- $X = \{X_1, \cdots, X_K\}$: variables with missing values,
- $R = \{R_1, \cdots, R_K\}$, $R_i = 1$: observed, $R_i = 0$: missing,
- $X^* = \{X_1^*, \cdots, X_K^*\}$: $X_k^* = X_k$ if $R_k = 1$, and $X_k^* = ?$ if $R_k = 0$.
- Missing data DAG models: distributions $p(X, R, X^*)$ that factorizes as:

$$\prod_{X_k \in X} p(X_k \mid \mathrm{pa}_{\mathcal{G}}(X_i)) \times \prod_{R_k \in R} p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k)) \times \prod_{X_k^* \in X^*} p(X_k^* \mid \mathrm{pa}_{\mathcal{G}}(X_k^*)).$$

- Propensity score: $\pi_k(\mathrm{pa}_{\mathcal{G}}(R_k)) = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))$.
- Missingness mechanism: $p(R \mid X)$, Full law: $p(X, R)$, Target law: $p(X)$, Observed data law: $p(X^*, R)$.
- Scope: we focus on mDAGs with an unrestricted target law, imposing assumptions only on the missingness mechanism.
- Aim: identify the target law, which is possible if $\pi_k|_{R=1}$ is identified for each $R_k \in R$:

$$p(X) = p(X, R = 1) \, / \, p(R = 1 \mid X).$$

Associational irrelevancy

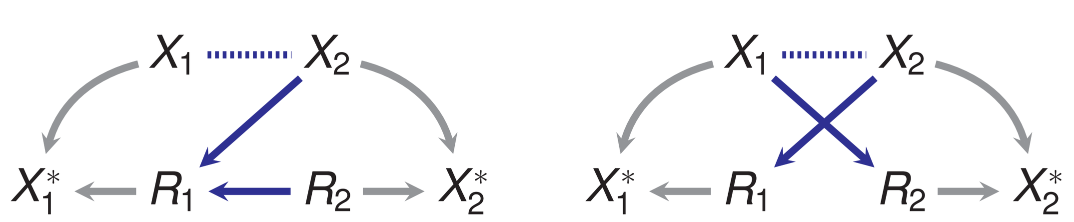


Figure: (a) mDAG for the "block-conditional MAR" model [2]; (b) mDAG for the "block-parallel" model [1]. Target law is identified by associational irrelevancy.

- Missing variables in $\mathrm{pa}_{\mathcal{G}}(R_k)$ are the main challenge in identifying π_k.

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

- Associational irrelevancy relies on the local Markov property:

$$R_k \perp \mathrm{nd}_{\mathcal{G}}(R_k) \setminus \mathrm{pa}_{\mathcal{G}}(R_k) \mid \mathrm{pa}_{\mathcal{G}}(R_k)$$

1. Fig. 1(a): $\pi_1 = p(R_1 = 1 \mid X_2, R_2)$. Let $\mathcal{S}_1^r = \{R_2\}$. $\pi_1|_{\mathcal{S}_1^r=1}$ is identified.
2. Fig. 1(b): $\pi_1 = p(R_1 = 1 \mid X_2) = p(R_1 = 1 \mid X_2, R_2 = 1)$. Thus, $\mathcal{S}_1^r = \emptyset$; π_1 is identified.

Causal irrelevancy

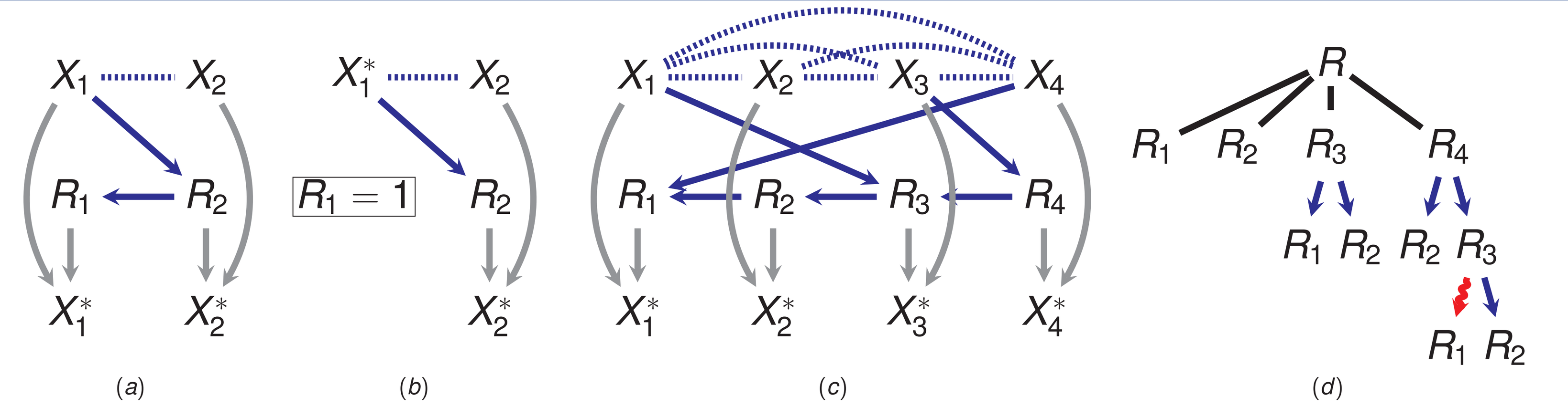


Figure: (a, c) mDAGs illustrating causal irrelevancy; (b) Conditional mDAG for (a) showing the intervention on R_2; (d) Tree structure corresponding to target law identification in (c).

- Identification complicates when $\mathcal{S}_k^x$ contains descendants of R_k.

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

- Associational irrelevancy fails for indicators in $\mathcal{R}_k^p$.
- Causal irrelevancy serves as an alternative, which explores independence relationships in post-intervention distributions.
- Consider Fig. 2(a): $\pi_1 = p(R_1 = 1|R_2)$ and $\pi_2 = p(R_2 = 1|X_1)$.
  - $\boxed{\pi_2}$: $\mathcal{R}_2^p = \{R_1\}$. $\pi_2 = p(R_2 = 1|X_1^*, \mathrm{do}(R_1 = 1))$ is identified ($\mathcal{S}_2^r = \emptyset$).

$$p(X_1^*, X_2, R_2|\,\mathrm{do}(R_1)) = p(X, R_1 = 1, R_2)/\pi_1(R_2). \quad \text{(post-intervention distribution)}$$

Selection bias, admissible interventions, and pruning operation

- Intervention on R_k imposes selections defined by the selection set.

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

- Consider Fig. 2(c):

$$\pi_1 = p(R_1 = 1 \mid X_4, R_2), \qquad \pi_2 = p(R_2 = 1 \mid R_3)$$
$$\pi_3 = p(R_3 = 1 \mid X_1, R_4), \qquad \pi_4 = p(R_4 = 1 \mid X_3).$$

  - $\boxed{\pi_1, \pi_2}$ is identified, with $\mathcal{S}_1 = \{R_4\}$ and $\mathcal{S}_2 = \emptyset$.
  - $\boxed{\pi_3}$: $\mathcal{R}_3^p = \{R_1\}$. Three strategies for placing interventions:
    - i. R_1 ($\mathcal{S}_1 = \{R_4\}$): $\mathcal{S}_3^r = \{R_4\}$; $\pi_3|_{\mathcal{S}_3^r=1} = p(R_3|X_1, R_4 = 1, \mathrm{do}(R_1 = 1))$.
    - ii. R_2 ($\mathcal{S}_2 = \emptyset$): $\mathcal{S}_3^r = \emptyset$; $\pi_3 = p(R_3|X_1, R_4, R_1 = 1, \mathrm{do}(R_2 = 1))$.
    - ✓iii. R_1, R_2: $\mathcal{S}_3^r = \{R_4\}$. Algorithm proceeds with *maximal* interventions.
  - $\boxed{\pi_4}$: intervening on R_1 is inadmissible, as $\mathcal{S}_1 = \{R_4\}$ selects on R_4.
  - To handle $\mathcal{R}_4^p = \{R_3\}$, we intervene on R_3 and *re-identify* its propensity score after pruning the intervention on R_1: $\mathcal{S}_3 = \{R_1, \not{R_4}\}$.
  - Intervene on R_2 to remove selection from $\mathcal{S}_3$, given that $R_4 \perp R_1 \mid X_3$.
- Selection from R_1 propagates to R_3 following the rule:

$$\mathcal{S}_{j\downarrow k} := \mathcal{S}_j \cap \mathrm{pa}_{\mathcal{G}}(R_k). \quad \text{(selection propagation rule)}$$

General Identification Algorithm

- Identify propensity scores sequentially in any given mDAG.
- Track selection bias and characterize its propagation.
- Determine which interventions require pruning.

Estimation and Inference

- Inverse propensity weighted estimating equations for each π_k.
- To estimate $\pi_3|_{\mathcal{S}_3^r=1}$ with parameter θ_3, define IPW weight:

$$W_3(\theta_{\mathcal{T}_3}) := \prod_{R_i \in \mathcal{T}_3} \mathbb{I}(R_i = 1)/\pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r=1}, \text{ where } \mathcal{T}_3 = \{R_1, R_2\}$$

  - Estimate θ_3 by solving

$$P_n \Psi_3(X^*, R; \theta_3, \hat{\theta}_{\mathcal{T}_3}) = 0, \text{ where}$$
$$\Psi_3(X^*, R; \theta_3, \hat{\theta}_{\mathcal{T}_3}) := \mathbb{I}(R_4 = 1)\, W_3(\hat{\theta}_{\mathcal{T}_3})\, f_3(\mathrm{pa}_{\mathcal{G}}(R_3)) \left\{R_3 - \pi_3(\mathrm{pa}_{\mathcal{G}}(R_3); \theta_3)\right\}.$$

  - Alternatively, use weighted regression:

$$R_3 \sim \mathrm{pa}_{\mathcal{G}}(R_3), \text{with weight} = \mathbb{I}(R_4 = 1)\, W_3(\hat{\theta}_{\mathcal{T}_3}).$$

- The variance of $\hat{\theta}_3$ has a sandwich form. Let $\boldsymbol{\theta}_3 = (\theta_1, \theta_2, \theta_3)$ and $\boldsymbol{\Psi}(\boldsymbol{\theta}_3) = (\Psi_1, \Psi_2, \Psi_3)$.

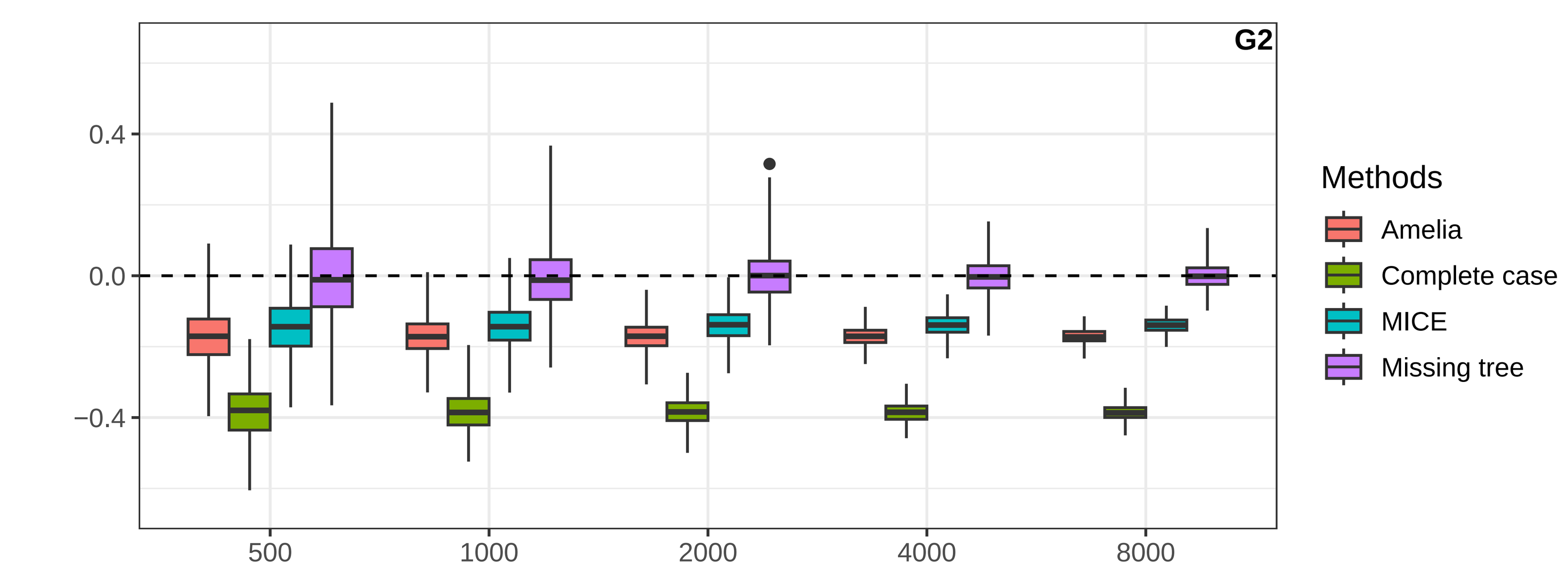
$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_3 - \boldsymbol{\theta}_3) \rightsquigarrow N(0,\ V_3), \qquad V_3 = A_3^{-1} B_3 (A_3^{-1})', \text{ where}$$

$A_3 := \mathbb{E}\{\partial \boldsymbol{\Psi}_3(\boldsymbol{\theta}_3)/\partial \boldsymbol{\theta}_3\}$ and $B_3 := \mathbb{E}\{\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)\boldsymbol{\Psi}_k(\boldsymbol{\theta}_k)'\}$.

- Target-law functionals can be estimated similarly: e.g. $\theta_t = \mathbb{E}(X_3)$:

$$\hat{\theta}_t = P_n\Big(\mathbb{I}(\mathcal{R} = 1)\Big\{\prod_{R_i \in \mathcal{R}} \pi_i(\mathrm{pa}_{\mathcal{G}}(R_i); \theta_i)|_{\mathcal{S}_i^r=1}\Big\}^{-1} X_3\Big), \text{ where } \mathcal{R} = \{R_4, R_3, R_1\}.$$

Simulation: Mean Estimation



Implementation via the `flexMissing` R package

```
library(flexMissing)
G <- make.graph(obs_variables = c(),
         missing_variables = c('X1','X2','X3','X4'),
         missing_indicators = c('R1','R2','R3','R4'),
         di_edges = list(c('R4','R3'),c('R3','R2'),c('R2','R1'),
                         c('X1','R3'),c('X3','R4'),c('X4','R1')))
# check target law ID
ID <- f.ID_algorithm(G) # The target law is identified.
plot(ID)
# estimate propensity scores
propensitys <- f.propensity(graph=G,data=data,ID=ID)
clique <- find_clique(G,data, propensitys, est_R='R3') # find cal R
est <- mean(data$X3*clique$magic_weight) # estimate E(X3)
```

References

[1] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc., 2013.

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