Sufficient Identification Conditions and Semiparametric Estimation under Missing Not at Random Mechanisms

Anna Guo†, Jiwei Zhao‡, and Razieh Nabi†

† Dept. of Biostatistics & Bioinformatics, Emory University

‡ Dept. of Biostatistics & Medical Informatics, University of Wisconsin-Madison

EMORY UNIVERSITY

Motivation

- Analyzing data missing-not-at-random (MNAR) is challenging.
- Existing approaches for identification and inference under MNAR mechanisms rely on fully observed variables and/or untestable model assumptions.
- A more relaxed set of statistical assumptions is needed.

Missing Data DAG Models

- $L = \{X, Y\}$: variables with missing values,
- $R = \{R_x, R_y\}$: binary indicators, $R_i = 1$: observed, $R_i = 0$: missing,
- $L^* = \{X^*, Y^*\}$, deterministically defined proxy variables: $L_i^* = L$ if $R_i = 1$, and $L_i^* = ?$ if $R_i = 0$.
- Missing data DAG models: a set of distributions $p(L, R, L^*)$ that factorizes as:

$$\prod_{L_i \in L} p(L_i \mid \mathrm{pa}_{\mathcal{G}}(L_i)) \times \prod_{R_i \in R} p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i)) \times \prod_{L_i^* \in L^*} p(L_i^* \mid \mathrm{pa}_{\mathcal{G}}(L_i^*)).$$

- Missingness mechanism: $p(R \mid L)$, Full law: $p(L, R)$, Target law: $p(L)$, Observed data law: $p(L^*, R)$.

Criss-Cross MNAR Model

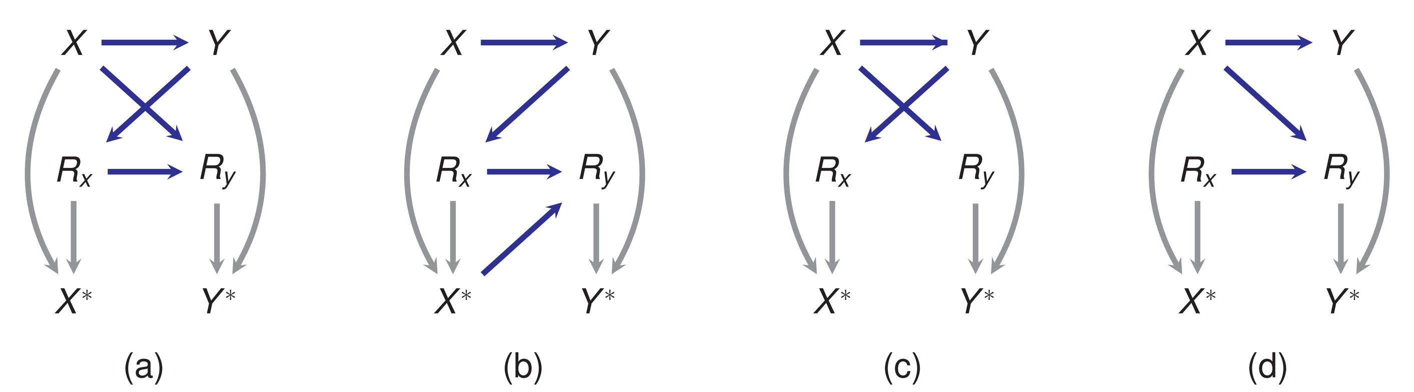


Figure: (a) Criss-cross MNAR model; (b) Permutation model [2]; (c) Block-parallel model [1]; (d) Block-conditional MAR model [3].

1. Neither full law nor target law is identifiable (proof via a bivariate normal counterexample)
2. Only known structure thus far that prevents target law identification. (besides $L_i \to R_i$ self-censoring)
3. Super model of several popular MNAR models shown above.

(a) $R_x \perp X \mid Y$; $R_y \perp Y \mid X, R_x$.

(b) $R_x \perp X \mid Y$; $R_y \perp Y, X \mid R_x, X^* \Rightarrow R_y \perp Y \mid R_x = 1, X$; $R_y \perp Y, X \mid R_x = 0$.

4. Permits missing values for all variables.

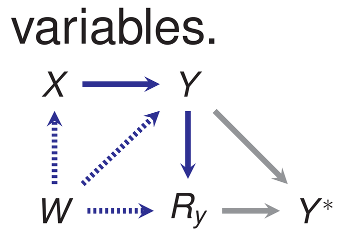


Figure: Shadow variable setup considered in Wang et al, 2014

Identification

Partial Identification & Testability

- $p(X \mid Y)$ is nonparametrically identifiable $\Rightarrow$ testability of $X \to Y$.

$$p(X \mid Y) = p(X \mid Y, R_x = 1) = \frac{p(X, Y, R_x = 1)}{\int p(x, Y, R_x = 1)\, dx}$$

$$p(X, Y, R_x = 1) = \frac{p(X, Y, R_x = 1, R_y = 1)}{p(R_y = 1 \mid R_x = 1, X, Y)} = \frac{p(X, Y, R_x = 1, R_y = 1)}{p(R_y = 1 \mid R_x = 1, X)}.$$

Target Law Identification

- Nonparametric identification of $X|Y$ sheds light on identifying the target law: for two distinct points of X: x_1 and x_0

$$\frac{p(x_1 \mid y)}{p(x_0 \mid y)} = \frac{p(y \mid x_1)}{p(y \mid x_0)} \times \frac{p(x_1)}{p(x_0)}.$$

- Exponential family

$$p(x) \sim \exp\left\{\frac{x\eta_x - b_x(\eta_x)}{\Phi_x} + c_x(x;\ \Phi_x)\right\}$$

$$p(y \mid x) \sim \exp\left\{\frac{y\eta - b(\eta)}{\Phi} + c(y;\ \Phi)\right\}, g(\mu(\eta)) = \alpha + \beta x.$$

- Assume X takes $k+1$ distinct values $x_0, x_1, \cdots, x_k$. Let $\varphi = [g \circ \mu]^{-1}$ and $\zeta = b \circ \varphi$,

$$\phi_i(\theta) = \{\varphi(\alpha + x_i\beta) - \varphi(\alpha + x_0\beta)\}/\Phi$$

$$\zeta_i(\theta) = \frac{-\zeta(\alpha + x_1\beta) + \zeta(\alpha + x_0\beta)}{\Phi} + \frac{\eta_x(x_1 - x_0)}{\Phi_x} + c(x_1;\ \Phi_x) - c(x_0;\ \Phi_x)$$

$$J = \partial(\Phi, Z)/\partial\theta.$$

- Target law $p(X, Y)$ is ID if: (i) $k \geq dim(\theta)$, and (ii) J has full rank.
- Generalizable to non-exponential family and multivariate X.

Full Law Identification

- **Completeness condition**: $\forall h(X)$ with finite mean, $\mathbb{E}\{h(X) \mid Y\} = 0$ implies $h(X) = 0$ a.s.
  - Exponential family is a special case.
- Full law $p(X, Y, R_x, R_y)$ is ID if: (i) J is full rank, and (ii) completeness condition holds.

Bivariate Normal Example

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right].$$

- Target law and full law are ID: one of $\{\mu_1, \mu_2\}$ + any of $\{\sigma_1, \sigma_2, \rho\}$
- Pseudo-likelihood with logistic regression: $v_k = (x_i - x_k)\, |y_i - y_k|$

$$u_k = \begin{cases} 1 & \text{if } y_i - y_k > 0 \\ 0 & \text{if } y_i - y_k < 0. \end{cases}$$

Estimation and Inference

Pseudo-likelihood

- Order statistics: $\tilde{X} = (x_{(1)}, \ldots, x_{(n)})$

$$p(x_1, \ldots, x_n \mid r_{x_1} = r_{y_1} = 1, \ldots, r_{x_n} = r_{y_n} = 1, y_1, \ldots, y_n, \tilde{X})$$

$$= \frac{\prod_{i=1}^n p(x_i \mid y_i)}{\sum_{\text{permutation of } x} \prod_{i=1}^n p(x_{(i)} \mid y_i)} \text{ complexity of order } n!$$

$$\approx \prod_{i<k} \frac{p(x_i \mid y_i)\, p(x_k \mid y_k)}{p(x_i \mid y_i)\, p(x_k \mid y_k) + p(x_i \mid y_k)\, p(x_k \mid y_i)}$$

$$= \prod_{i<k} \frac{1}{1 + Q(x_i, y_i; x_k, y_k)}, \quad Q = OR^{-1}.$$

- Model specification: $p(X \mid Y)$

Generalized Estimating Equations

- GEE

$$\mathbb{E}\left[\frac{R_x \times R_y}{p(R_y = 1 \mid R_x = 1, X)} \times f(Y) \times (X - E(X \mid Y; \theta))\right] = 0$$

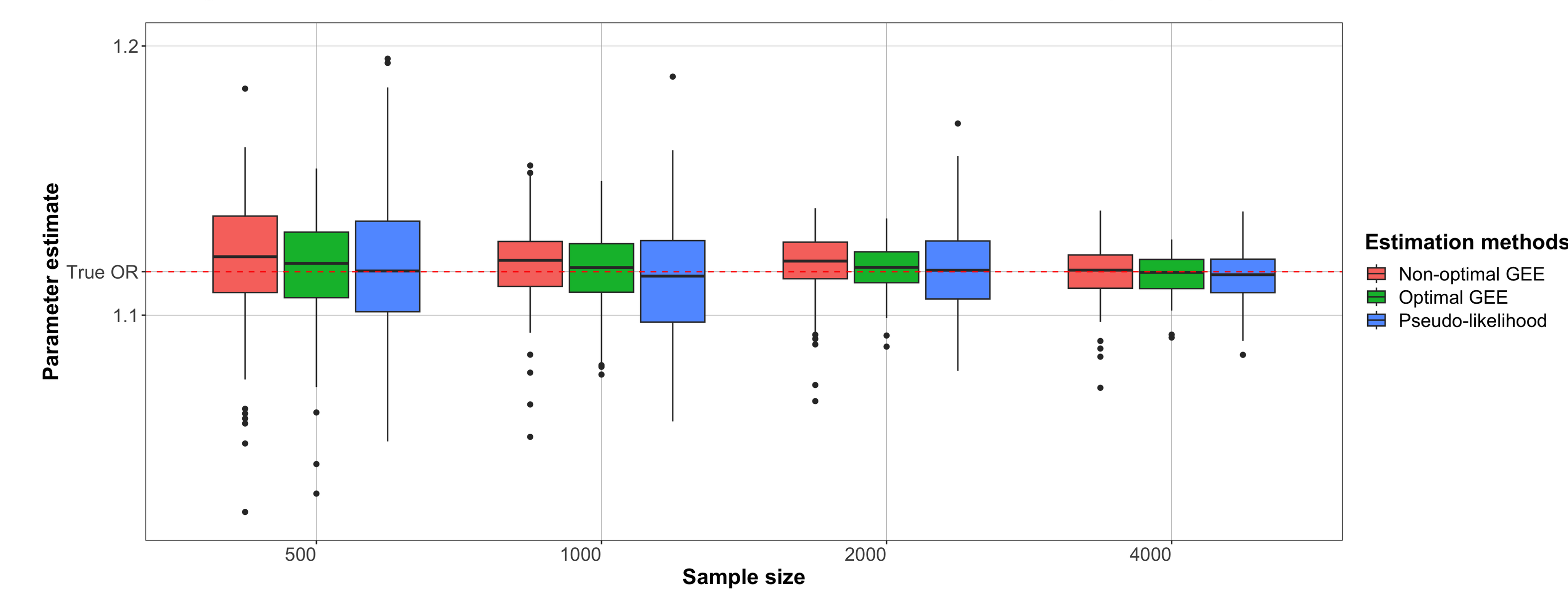
- Optimal GEE

$$f_{opt}(Y) = \left[\mathbb{E}\left\{\frac{(X - \mathbb{E}(X \mid Y; \theta))^2}{p(R_y = 1 \mid R_x = 1, X)} \mid Y\right\}\right]^{-1} \frac{\partial E(X \mid Y; \theta)}{\partial \theta}\bigg|_{\theta = \theta_0}$$

- Model specification: $p(R_y = 1 \mid R_x = 1, X)$ and $E(X \mid Y; \theta)$

Simulation: Odds Ratio Estimation

Illustrating unbiasedness of the estimators and the efficiency of optimal GEE:



Future Work

- Developing a doubly-robust estimation framework.

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

[3] 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.