Handling Nonmonotone Missing Data with Available Complete-Case Missing Value Assumption

Gang Cheng, Yen-Chi Chen, Maureen A. Smith, and Ying-Qi Zhao

February 7, 2024

Why did I choose this paper?

1. I think that this paper answers some important questions about how to analyze data that is nonmonotone and missing-not-at-random.
2. The paper does a good job motivating the semiparametric model and how one can combine a response model with an outcome model.
3. The main example shows a distinct use for paper's methodology.

Outline

1. Motivating Example
2. Problem and Notation
3. Single Variable Case
4. Multiple Variable Case
5. Simulations

Motivating Example

Motivating Example

- Electronic health records (EHRs) data set that contains longitudinal information about diabetes patients.
- Primary variable of interest: glycated hemoglobin (HbA1c) measurement.
- A HbA1c level of less than 7% is known to reduce the risk of microvascular complications.

Motivating Example

- However, EHR data is *incomplete* because it is only measured when patients enter a clinic.
- The missingness is likely to be missing-not-at-random (MNAR) because a sicker patient is more likely to come into the clinic.

Monotone and Nonmonotone Data

- The data also contains nonmonotone missing patterns when patients miss visits.
- This data contains quarterly measurements from 10 years worth of data, but the authors only consider analyzing the first year.

Monotone Data

Index	Q1	Q2	Q3	Q4
1	✓	✓	✓	✓
2	✓	✓	✓	
3	✓	✓		
4	✓	✓		

Nonmonotone Data

Index	Q1	Q2	Q3	Q4
1	✓	✓	✓	✓
2	✓		✓	
3	✓		✓	
4	✓			✓

Problem and Notation

Problem

- How are we going to analyze this data?
- Three questions we want to answer:
  1. Single Variable of Interest: $\theta = E[Y_4]$
  2. Multiple Variables: Summary Measures

$$\theta = E[Y_3 Y_4] \text{ or } \theta = \Pr(Y_3 \leq 0.07, Y_4 \leq 0.07)$$

  3. Multiple Variables: Marginal Parametric Model

$$E[Y_4 \mid Y_2, Y_3] = \beta_0 + \beta_1 Y_2 + \beta_2 Y_3.$$

Notation

- Primary variables (L)
- Auxiliary variables (X)
- Parameter of interest $\theta = E[f(L)]$
- Response pattern for L: $A \in \{0,1\}^d$ where $A_j = 1$ if L_j is observed.
- Response pattern for X: $R \in \{0,1\}^p$ where $R_j = 1$ if X_j is observed.

Notation

- We define $R \geq r$ if $R_i \geq r_i$ for all $i \in \{1, \ldots, p\}$.
- For example, $1010 \geq 1000$ but 1010 cannot be compared with 0101.

Single Variable Case

Question 1: Single Variable of Interest

- Since we have a single variable of interest $L \in \mathbb{R}$ and $A \in \{0, 1\}$. We want to estimate $\theta = E[f(L)]$ for some known f.

$$
\begin{aligned}
\theta = E[f(L)] &= \int f(\ell)p(\ell)d\ell \\
&= \underbrace{\int f(\ell)p(\ell, A = 1)d\ell}_{\theta_1} + \\
&\quad \sum_r \underbrace{\int \int f(\ell)p(\ell, x_r, R = r, A = 0)dx_r d\ell}_{\theta_{0,r}} \\
&= \theta_1 + \sum_r \theta_{0,r}
\end{aligned}
$$

Question 1: Single Variable of Interest

- We know that θ_1 is identifiable since $A = 1$.
- To estimate $\theta_{0,r}$ we notice that

$$\begin{aligned}
\theta_{0,r} &= \int \int f(\ell) p(\ell, x_r, R = r, A = 0) dx_r d\ell \\
&= \int \int f(\ell) p(\ell \mid x_r, R = r, A = 0) p(x_r, R = r, A = 0) d\ell dx_r
\end{aligned}$$

Identification

- To identify $\theta_{0,r}$ we need to identify both $p(x_r, R = r, A = 0)$ and $p(\ell \mid x_r, R = r, A = 0)$.
- The first quantity is identifiable from the data.
- The second quantity requires an additional assumption.

Assumption

- The available complete-case missing value (ACCMV) assumption says the following:

$$p(\ell \mid x_r, R = r, A = 0) = p(\ell \mid x_r, R \geq r, A = 1).$$

Assumption

- The available complete-case missing value (ACCMV) assumption says the following:

$$p(\ell \mid x_r, R = r, A = 0) = p(\ell \mid x_r, R \geq r, A = 1).$$

- The ACCMV assumption should be contrasted with the complete-case missing values (CCMV) assumption [2] which is

$$p(\ell \mid x_r, R = r, A = 0) = p(\ell \mid x_r, R = 1_p, A = 1).$$

IPW Estimation

- First notice that the ACCMV assumption is equivalent to

$$\frac{\Pr(R = r, A = 0 \mid X_r, L)}{\Pr(R \geq r, A = 1 \mid X_r, L)} = \frac{\Pr(R = r, A = 0 \mid X_r)}{\Pr(R \geq r, A = 1 \mid X_r)} := O_r(X_r).$$

- We can estimate the odds ratio, $O_r(X_r)$, using logistic regression.

IPW Estimation

- Then we have

$$
\begin{aligned}
\theta_{0,r} &= \int\int f(\ell)p(\ell, x_r, R = r, A = 0)dx_r d\ell \\
&= \int\int f(\ell)\frac{p(\ell, x_r, R = r, A = 0)}{p(\ell, x_r, R \geq r, A = 1)}p(\ell, x_r, R \geq r, A = 1)dx_r d\ell \\
&= \int\int f(\ell)\frac{\Pr(R = r, A = 0 \mid x_r, \ell)}{\Pr(R \geq r, A = 1 \mid x_r, \ell)}p(\ell, x_r, R \geq r, A = 1)dx_r d\ell \\
&= \int\int f(\ell)O_r(x_r)p(\ell, x_r, R \geq r, A = 1)dx_r d\ell \\
&= E[f(L)O_r(X_r)I(R \geq r, A = 1)].
\end{aligned}
$$

IPW Estimation

- Then we have the IPW estimator

$$\hat{\theta}_{0,r,IPW} = n^{-1} \sum_{i=1}^{n} f(L_i) O_r(X_{ir}; \hat{\alpha}_r) I(R_i \geq r, A_i = 1).$$

- This can be combined with an estimated mean for θ_1 and create

$$\hat{\theta}_{IPW} = n^{-1} \sum_{i=1}^{n} f(L_i) I(A_i = 1) \left(1 + \sum_{r} O_r(X_{ir}; \hat{\alpha}_r) I(R_i \geq r) \right)$$

Theory

Theorem

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{IPW} - \theta) \xrightarrow{d} N(0, \sigma^2_{IPW}).$$

Regression Estimation

- We can also use the ACCMV assumption directly to get an outcome model approach.

$$
\begin{aligned}
\theta_{0,r} &= \int\int f(\ell)p(\ell \mid x_r, R = r, A = 0)p(x_r, R = r, A = 0)d\ell dx_r \\
&= \int\int f(\ell)p(\ell \mid x_r, R \geq r, A = 1)p(x_r, R = r, A = 0)d\ell dx_r \\
&= \int m_{r,0}(x_r)p(x_r, R = r, A = 0)dx_r \\
&= E[m_{r,0}(X_r)I(R = r, A = 0)]
\end{aligned}
$$

where $m_{r,0}(x_r) = E[f(\ell) \mid X_r = x_r, R \geq r, A = 1]$.

Regression Estimation

- Then we can construct a regression estimator with

$$\hat{\theta}_{0,r,Reg} = n^{-1} \sum_{i=1}^{n} m_{r,0}(X_{i,r}, \hat{\beta}_r) I(R_i = r, A_i = 0).$$

- This can be combined with the estimated mean for θ_1 to get

$$\hat{\theta}_{Reg} = n^{-1} \sum_{i=1}^{n} (f(L_i)A_i + m_{R_i,0}(X_{i,R_i}, \hat{\beta}_{R_i})(1 - A_i)).$$

Theory

Theorem

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{Reg} - \theta) \xrightarrow{d} N(0, \sigma^2_{Reg}).$$

Double Robust Estimation

- Since the IPW and regression estimator might not be efficient, we combine the two to get a double robust estimator.

Theorem

Under the ACCMV assumption, the efficient influence function of estimating $\theta_{0,r}$ is

$$(f(L) - m_{r,0}(X_r))O_r(X_r)I(R \geq r, A = 1) + m_{r,0}(X_r)I(R = r, A = 0) - \theta_{0,r}.$$

Double Robust Estimation

- Hence, an efficient estimator of $\theta_{0,r}$ is

$$\begin{aligned}
&\hat{\theta}_{0,r,DR} \\
&= n^{-1}\sum_{i=1}^{n}\{(f(L_i) - \hat{m}_{r,0}(X_{i,r}))\hat{O}_r(X_{i,r})I(R_i \geq r, A_i = 1) \\
&\quad + \hat{m}_{r,0}(X_{i,r})I(R_i = r, A_i = 0)\}
\end{aligned}$$

- This lead to the double robust estimator

$$\hat{\theta}_{DR} = \sum_{r}\hat{\theta}_{0,r,MR} + \hat{\theta}_1$$

where $\hat{\theta}_1 = n^{-1}\sum_{i=1}^{n} f(L_i)I(A_i = 1)$.

Theory

Theorem

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{DR} - \theta) \xrightarrow{d} N(0, \sigma^2_{eff}).$$

Multiple Variable Case

Question 2: Multiple Variables of Interest

- Now, we consider the problem when $L \in \mathbb{R}^d$ is multivariate.
- In this setup, the complete case is $A = 1_d$.
- We also introduce a new notation $\bar{a} = 1_d - a$ and $\bar{r} = 1_p - r$.

ACCMV Assumption

- For a multivariate L, the ACCMV assumption is revised to

$$p(\ell_{\bar{a}} \mid \ell_a, x_r, A = a, R = r) = p(\ell_{\bar{a}} \mid \ell_a, x_r, A = 1_d, R \geq r).$$

- This is equivalent to

$$\frac{\Pr(R = r, A = a \mid x_r, \ell)}{\Pr(R \geq r, A = 1_d \mid x_r, \ell)} = \frac{\Pr(R = r, A = a \mid x_r, \ell_a)}{\Pr(R \geq r, A = 1_d \mid x_r, \ell_a)} := O_{r,a}(x_r, \ell_a).$$

IPW Estimation

- To handle the multivariate case, notice that we have

$$\theta = E[f(L)] = \sum_{r,a} E[f(L)I(A = a, R = r)] = \sum_{r,a} \theta_{r,a}.$$

- $a = 1_d$, $\theta_{r,a}$ is identifiable and

$$\theta_{r,a} = E[f(L)I(A = a, R = r)].$$

- When $a \neq 1_d$, we can derive

$$\theta_{r,a} = E[f(L)O_{r,a}(X_r, L_a)I(A = 1_d, R \geq r)].$$

IPW Estimation

- Furthermore, we can express

$$
\begin{aligned}
&\sum_{r,a\neq 1_d} O_{r,a}(X_r, L_a) I(A = 1_d, R \geq r) \\
&= \sum_r I(A = 1_d, R = r) \sum_{\tau \leq r, a \neq 1_d} O_{\tau,a}(X_\tau, L_a) \\
&= \sum_r Q_r(X_r, L) I(R = r, A = 1_d)
\end{aligned}
$$

where $Q_r(X_r, L) = \sum_{\tau \leq r, a \neq 1_d} O_{\tau,a}(X_\tau, L_a)$.

IPW Estimation

- Then,

$$
\begin{aligned}
\theta &= E[f(L)] \\
&= \sum_{r,a\neq 1_d} E[f(L)O_{r,a}(X_r, L_a)I(A = 1_d, R \geq r)] \\
&\quad + \sum_r E[f(L)I(A = 1_d, R = r)] \\
&= E\Big[f(L)\sum_r (1 + Q_r(X_r, L)I(R = r, A = 1_d))\Big].
\end{aligned}
$$

IPW Estimation

Then IPW estimation has three steps

1. Estimate the individual odds $O_{r,a}$
2. Compute the total weights Q_r.

$$\hat{Q}_r(X_r, L) = \sum_{\tau \le r} \sum_{a \ne 1_d} \hat{O}_{\tau,a}(X_\tau, L_a).$$

3. Apply the IPW Approach

$$\hat{\theta}_{IPW} = n^{-1} \sum_{i=1}^{n} f(L_i)(\hat{Q}_{R_i}(X_{i,R_i}, L_i) + 1) I(A_i = 1_d).$$

Theory

Theorem

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{IPW} - \theta) \xrightarrow{d} N(0, \sigma^2_{IPW}).$$

Regression Estimation

- Similar to the single variable case, the ACCMV assumption implies that

$$\theta_{r,a} = E[m_{r,a}(X_r, L_a) I(R = r, A = a)]$$

where $m_{r,a}(X_r, L_a) = E[f(L) \mid L_a, X_r, R \geq r, A = 1_d]$.

Regression Estimation

1. Estimate the outcome regression model.
2. Using the regression model to impute the missing values,

$$\hat{\theta}_{Reg} = n^{-1} \sum_{i=1}^{n} (f(L_i) I(A_i = 1_d) + \hat{m}_{R_i, A_i}(X_{i,R_i}, L_{A_i}) I(A_i \neq 1_d)).$$

Theory

Theorem

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{Reg} - \theta) \xrightarrow{d} N(0, \sigma^2_{Reg}).$$

Double Robust Estimation

- For exactly the same reasons as before, we can construct a double robust estimator:

$$
\begin{aligned}
&\hat{\theta}_{DR} \\
&= n^{-1}\sum_{i=1}^{n}\Big(\sum_{r,a\neq 1_d}\{(f(L_i)-\hat{m}_{\tau,a}(X_i,r))\hat{O}_{r,a}(X_{i,r},L_{i,a})I(R_i\geq r,A_i=1_d) \\
&\quad+\hat{m}_{r,a}(X_{i,r},L_{i,a})I(R_i=r,A_i=a)\}+f(L_i)I(A_i=1_d)\Big)
\end{aligned}
$$

Theory

Theorem

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{DR} - \theta) \xrightarrow{d} N(0, \sigma^2_{eff}).$$

Simulations

Simulation: Single Variable

- Let $L = Y_3$ and $X = (Y_1, Y_2)$.
- We want to estimate $\theta = E[Y_3]$.
- Let $|r| = \sum_r r_i$ be the number of observed variables in X.

Simulation: Single Variable

We generate data with the following setup:

1. $(L, X_r) \mid A = 1, R = r \sim N\left(\mu_{|r|+1}, \Sigma_{|r|+1}\right)$
2. $X_r \mid A = 0, R = r \sim N\left(\mu_{|r|}, \Sigma_{|r|}\right)$.

- We have $\mu_1 = 1$, $\mu_2 = (1, -1)'$, $\mu_3 = (0, -1, -1)'$, and

$$\Sigma_1 = 1, \Sigma_2 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} \text{ and } \Sigma_3 = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{bmatrix}.$$

Simulation: Single Variable

- We assume that $\Pr(A = j, R = r) = 1/8$ for $j = 0, 1$ and $r \in \{00, 01, 10, 11\}$.
- We run 1000 Monte Carlo samples with the total number of observations equal to $n = 2000$.

Simulation Results

Method	Bias	Coverage of 95% CI
IPW	-0.006	0.778
IPW (mis.)	-0.084	0.536
Reg	-0.001	0.955
Reg (mis.)	0.040	0.857
DR	-0.002	0.931
DR (IPW mis.)	0.000	0.939
DR (Reg mis.)	-0.000	0.920
DR (Both mis.)	0.041	0.870
Complete Case	-0.178	0.001

Simulation: Multiple Variables

- $L = (Y_3, Y_4)$ and $X = (Y_1, Y_2)$
- Goal: $\theta = E[Y_3 Y_4]$

Simulation: Multiple Variables

We generate the data using the following procedure:

1. $(L, X_r) \mid A = 11, R = r \sim N(1_{2+|r|,\Sigma_{2+|r|}})$ for $r \in \{00, 01, 10, 11\}$.
2. $(L_a, X_r) \mid A = a, R = r \sim N(\mu_{1+|r|}, \Sigma_{1+|r|})$ for any $a \in \{01, 10\}$ and any $r \in \{00, 01, 10, 11\}$.
3. $X_r \mid A = 00, R = r \sim N(\mu_{|r|}, \Sigma_{|r|})$ for any $r \in \{01, 10, 11\}$.

where $\Sigma_d = (1/2)I_d + (1/2)1_d 1_d'$ and $\mu_1 = 0.5$, $\mu_2 = 1_2$, and $\mu_3 = 1_3$.

Simulation: Multiple Variables

- We let $\Pr(A = a, R = r) = 1/16$ for $a \in \{00, 01, 10, 11\}$ and for $r \in \{00, 01, 10, 11\}$.

Simulation Results

Method	Bias	Coverage of 95% CI
IPW	-0.000	0.943
IPW (mis.)	0.078	0.852
Reg	-0.001	0.956
Reg (mis.)	-0.048	0.892
DR	-0.001	0.948
DR (IPW mis.)	-0.001	0.949
DR (Reg mis.)	-0.001	0.952
DR (Both mis.)	0.014	0.943
Complete Case	0.131	0.723

Thank You

References

[1] Gang Cheng et al. "Handling Nonmonotone Missing Data with Available Complete-Case Missing Value Assumption". In: *arXiv preprint arXiv:2207.02289* (2022).

[2] Eric J Tchetgen Tchetgen, Linbo Wang, and BaoLuo Sun. "Discrete choice models for nonmonotone nonignorable missing data: Identification and inference". In: *Statistica Sinica* 28.4 (2018), p. 2069.