Non-ignorable Missing Data:Old and New

James Robins

The Problem of Missing Data:

Full Data:

$$\mathbf{L} = \overline{L}_K = (L_0, L_1, ..., L_K)$$

Observed Data:

$$\mathbf{O} = \left(R, L_{(R)} = L_{obs}\right),$$

$$\mathbf{R} = (R_0, R_1, ..., R_K)$$

$R_j = 1$ if L_j observed and 0 otherwise

$L_{(\mathbf{R})}$ are the observed components of L

Alternative Notation Related to Causality:

.

$$\mathbf{O} = (\mathbf{R}, L^* = L(R)), L^* = \left(L_0^*, L_1^*, ..., L_K^*\right),$$

$$L_j^* = L_j(\mathbf{R}) = L_j\left(R_j\right) = R_j L_j$$

$\cdot$ $L_j^* = L_j$ if $R_j = 1$, $L_j^* = \cdot$ if $R_j = 0$ is equivalent alternative since R_j observed.

$\cdot$Missingness indicators $\mathbf{R}$ always observed. Not always required in missing data.

·A Central Theme of this talk: Missing Data As Causal Inference Explains Much of What is Old and What is New. Alternative Notation Useful for this

· **Goals:** Given n iid observations O and a model $\mathcal{M}_{(\mathbf{R},\mathbf{L})}$ for the joint distribution of $(\mathbf{R},\mathbf{L})$, draw inferences concerning $F_{\mathbf{L}}$ of L and possibly $F_{\mathbf{R}|L}$ based on data O and the implied model $\mathcal{M}_O$ for its distribution F_O.

· I do not want to concentrate on complications (though real and interesting) due to continuity and measurability issues when discussing identification so I will assume discrete distributions when considering identification results.

· Inference generally continuous

Strategies for NMAR Motivated By The Following MAR Result:

$\cdot$**Definition:** Thehe nonparametric MAR model

$\mathcal{M}_{NP,(\mathbf{R},\mathbf{L})}^{MAR} = \mathcal{M}_{NP,(\mathbf{L})} \bigotimes \mathcal{M}_{NP,MAR,(\mathbf{R}|\mathbf{L})}$is the set of distributions for $(\mathbf{R},\mathbf{L})$ such that $pr\left(\mathbf{R}=\mathbf{r}|\mathbf{L}\right) \in \mathcal{M}_{NP,MAR,(\mathbf{R}|\mathbf{L})}{}^{ie}$

$$pr\left(\mathbf{R}=\mathbf{r}|\mathbf{L}\right) = pr\left(\mathbf{R}=\mathbf{r}|\mathbf{L}_{(\mathbf{r})}\right) = \pi\left(\mathbf{r},\mathbf{L}_{(\mathbf{r})}\right)$$

is a function $\pi\left(\mathbf{r},\mathbf{L}_{(\mathbf{r})}\right)$ of $\mathbf{L}$ only through $\mathbf{L}_{(\mathbf{r})}$. Throughout we restrict to the positive version of the model in which

$$pr\{\mathbf{R}=\mathbf{1}|L\} > 0 \ wp1$$

$\cdot$**Theorem (Gill,Van der Laan, JMR):** Under the above positivity condition, the model $\mathcal{M}_{NP,O}^{MAR}$ for the observed data implied by $\mathcal{M}_{NP,O}^{MAR}$ includes all distributions F_O.

Futher F_L and $F_{R|L}$ are identified.

We therefore say the model $\mathcal{M}_{NP,(\mathbf{R},\mathbf{L})}^{MAR}$ is non-parametric just identified (NPI) because

(i) it is non parametric for F_O

(ii) it identifies F_L even though it does not restrict the joint distributions F_O of the observed data or F_L of the full data

(iii) it is just identified because the model is not empirically testable based on the observed data O owing to excluding no law F_O

· Models like NP MAR that (for positive distributions) provide identification everywhere are to be distinguished from models that are generically identified i.e. that ar identified except at exceptional laws $F_{(R,L)}$.

·MAR is a Set of Conditional Independences:

$$\text{MAR: } I\left(\mathbf{R}=\mathbf{r}\right) \amalg L_{(\mathbf{r}^c)} | L_{(\mathbf{r})}$$

·Often Interested In Lower Dimensional Functionals $\psi\left(F_L\right)$ such as $E[L_3]$ in which case we need not demand identification of F_L but only of the functional.

·A variation independent model $\mathcal{M}_{(\mathbf{R},\mathbf{L})} = \mathcal{M}_{(\mathbf{L})} \bigotimes \mathcal{M}_{(\mathbf{R}|\mathbf{L})}$ with $\mathcal{M}_{(\mathbf{R}|\mathbf{L})}$ satisfying MAR is said to be ignorable since the likelihood factorizes as

$$f\left(L_{(R)};\theta\right)\pi\left(\mathbf{r},\mathbf{L}_{(\mathbf{r})};\gamma\right)$$

where θ indexes laws F_L in $\mathcal{M}_{(\mathbf{L})}$ and γ indexes laws in $\mathcal{M}_{(\mathbf{R}|\mathbf{L})}$.

·When MAR may not to be reasonable , JMR, Scharfstein and Rotnitzky introduced a philosophy of sensitivity analysis based on non-ignorable NPI models centered at an ignorable model.

·**Simplest Example**: $L = (Y, W)$, $O = (W, R, RY)$ with W high dim with cont components

$Model\ \mathcal{A}$ is the semiparametric model for $F_{(R,L)}$ satisfying the sole restriction that

$$pr\left[R = 1|W, Y\right] = \{\phi\left(\gamma^*\left(W\right) + \alpha^* Y\right)\}$$

where ϕ is a known, smoothly increasing distribution function, $\gamma^*\left(W\right)$ is a unknown unrestricted function and α^* is an unknown parameter. In particular F_L is unrestricted.

·Model $\mathcal{B}$ is the submodel of A in which α^* is known.

·If $\alpha^* = 0$ Model $\mathcal{B}$ is $\mathcal{M}^{MAR}_{NP,(\mathbf{R},\mathbf{L})}$

·Model $\mathcal{C}1$ is the submodel of A in which

$$\gamma^* (W) = \gamma (W; \upsilon^*)$$

with $\gamma(\cdot;\cdot)$ a known function and υ^* an unknown vector parameter and

$$f_Y (y|w) = f (y|w; \beta^*)$$

with $f(y|w;\beta)$ a known function and β^* an unknown vector parameter.

·Model $\mathcal{C}2$ differs from $\mathcal{C}2$ in the we replace the last parametric model by

$$f_Y (y|w, R = 1) = f_1 (y|w; \beta^*)$$

with $f_1(y|w;\beta)$ a known function and β^* an unknown vector parameter.

Models $\mathcal{C}1$ and $\mathcal{C}2$ are the same if and only if $\alpha^* = 0$

Theorem: Assumimg

$$pr\{R = 1|L\} > 0 \ wp1$$

(i) Models A and B contain all distributions F_O, but C does not so models A and B cannot be empirically tested. In particular, under model A any value of the selection parameter is compatible with the observed data distribution F_O.

(ii) Under model A, the distribution F_O that generated the data does not identify $\gamma^*(\cdot), \alpha^*$, F_L or any functional of F_L

(iii) Under model $B = B(\alpha^*)$, the distribution F_O that generated the data identifies $\gamma^*(\cdot)$ and $F_L, pr\{\mathbf{R} = \mathbf{1}|L\}$ even though the model $B(\alpha^*)$ left both $\gamma^*(\cdot)$ and F_L unrestricted.

$\cdot$Model $B = B(\alpha^*)$ is a NPI model since it places not restrictions on F_O but identifies F_L even though F_L is NP.

$\cdot$Under models C1 and C2 $\gamma^*(\cdot), \alpha^*$, F_L all tend to be identified but identification comes through the functionl form of the model $\gamma^*(W)$ and either $f_Y(y|w)$ or $f_Y(y|w, R = 1)$.

·**Philosophy of Sensitivity Analysis**:

·Generally parametric or semiparametric models for

$\gamma^{*}(W), f_{Y}(y|w), f_{Y}(y|w, R=1)$ not based on subject matter knowledge.

·Not good to get identification that way.

·Eric and Miao based on IV methods clear exceptions.

·From model A results, we conclude that α^{*} is not NP identified.

This combined with model B results suggests a sensitivity analysis strategy.

Sensitivity Analysis Strategy :

·For each value of α^* assume it is the truth.

·Let $F_L(\alpha^*, F_O)$ be the unique F_L implied by α^* and the F_O that generated the data.

·Plot $\psi(\alpha^*) = \psi\{F_L(\alpha^*, F_O)\}$ as a function of α^*.

·This is feature not a bug that we have to assume α^* known since the data offer no information if F_L and $\gamma^*(\cdot)$ left unspecified

·Estimation: Hence we only need estimate F_O from iid data on O.

·Given a functional $\psi(F_L)$, say $E_{F_L}[Y]$ if W is high dimensional the NP estimator of F_O is undefined.

·The usual method is to estimate $\psi(F_L)$ under working parametric or semi-parametric models (whose dimension may increase with sample size) with the goal to make estimation of $\psi(F_L)$ as robust as possible:

·Consistent at a large submodel of model B(α^*) with second order bias otherwise.

·Thus we search for so called DR estimators which if they exist will do so for only some parametrizations.

·It was surprising to us in 1999 that DR estimators existed in such nonignorable models as the likelihood no longer factors as

$$f\left(L_{(R)};\theta\right)\pi\left(\mathbf{R},\mathbf{L}_R;\gamma\right)$$

·It turns out that if (and essentially only if) we take ϕ to be expit (ie the logistic distribution) in

$$pr\left[R = 1|W, Y\right] = \left\{\phi\left(\gamma^*\left(W\right) + \alpha^* Y\right)\right\}$$

and use the models

$$\gamma^*\left(W\right) = \gamma\left(W; \upsilon^*\right)$$

$$f_Y\left(y|w, R = 1\right) = f_1\left(y|w; \beta^*\right)$$

then estimators based on solving the efficient influence function for the NP model B(α^*) will be CAN for $E_{F_L}\left[Y\right]$ in the union model in which one but not necessarily both of the above working models is correct.

·We also obtain a bias that is a expected value of the product of the error of each model in the limit.

·Explosion of DR estimators with better properties but same basic idea

α^* is a lousy sensitivity parameter because it is on the odds ratio scale so hard for subject matter experts to give a range. See Scharfstein et al Jasa discussion paper 1999.

· **Game Played**:

·Took a missingess model for $F_{R|L}$ defined by conditonal independence (ie fundamental non-parametric structural features)

$$\text{MAR: } I\left(\mathbf{R}=\mathbf{r}\right) \amalg L_{(\mathbf{r}^c)}|L_{(\mathbf{r})}$$

that is NPI subject to positivity constraints.

·An example

$$\text{MAR: } I\left(\mathbf{R}=\mathbf{r}\right) \amalg L_{(\mathbf{r}^c)}|L_{(\mathbf{r})}$$

·Then got rid of those independencies via an unidentified sensitivity parameter that we treat as known but vary in a sensitivity analysis. The model with the known sensitivity parameter remains NPI.

$\cdot$The parameter quantifies on some scale the dependence of $I\left(\mathbf{R}=\mathbf{r}\right)$ on $L_{(\mathbf{r}^C)}$ given $L_{(\mathbf{r})}$.

·When we start from a NPI MAR model we go from MAR to MNAR.

·But we could start with a NMAR NPI model defined by conditional independencies.

·Note we are restricting to models that identify the entire joint F_L of L which is not needed if only interested in a specific functional that depends on only a subset of the components of L.

·**NPI Ignorable Past-Nonignorable Future Missing (IPNFM) Model (formally a Permutation Missingness Model)**

·**Definition:** The nonparametric (IPNFM) model

$\mathcal{M}_{NP,(\mathbf{R},\mathbf{L})}^{IPNFM} = \mathcal{M}_{NP,(\mathbf{L})} \bigotimes \mathcal{M}_{NP,IPNFM,(\mathbf{R}|\mathbf{L})}$ associated with an ordering

$\overline{L}_K = (L_0, L_1, ..., L_K)$ of L (e.g. temporal) is the set of distributions for $(\mathbf{R}, \mathbf{L})$ restricted by $pr\left(\mathbf{R} = \mathbf{r}|\mathbf{L}\right) \in \mathcal{M}_{NP,IPNFM,(\mathbf{R}|\mathbf{L})}$

·That is, for each k

$$pr\left(R_k = 1|\overline{R}_{k-1}, L\right) = pr\left(R_k = 1|\overline{O}_{k-1}, \underline{L}_{k+1}\right)$$

where

· $\overline{O}_{k-1} = \left(\overline{R}_{k-1}, \overline{L}^*_{k-1}\right)$,is the observed past data where $L^*_j = R_j L_j$

·$\underline{L}_{k+1} = (L_{k+1}, ..., L_K)$ is the future unobserved full data .

·$\mathcal{M}_{NP,IPNFM,(\mathbf{R}|\mathbf{L})}$ is equivalently defined by

$$R_k \amalg \overline{L}_k | \overline{O}_{k-1}, \underline{L}_{k+1}$$

·The model can be represented in a so-called missingness graph of Mohan and Pearl which is a statistical DAG.

·A statistical DAG is a model that specifies that each variable on the graph is independent of its non-descendents given its parents.

·Using d-separation, we can see that in the 2 variable case

$$R_1 \amalg (L_0, L_1) \,|R_0, L_0^*$$
$$R_0 \amalg L_0|L_1$$

We throughout assume the positivity condition that

$$1 > pr\left(R_k = 1|\overline{O}_{k-1}, \underline{L}_{k+1}\right) > 0$$

so that at each time a subject has a positive probability to change their visit status as this precludes monotone missing data.

·This is a visit process not a censoring process.. Error in my 1997 paper re positivity.

·**Theorem (JMR, 1997):** Suppose the above positivity condition holds

The model $\mathcal{M}_{NP,\mathbf{O}}^{IPNFM}$ for the observed data implied by $\mathcal{M}_{NP,(\mathbf{R},\mathbf{L})}^{IPNFM}$ includes all distributions F_O.

Futher F_L and $F_{R|L}$ are identified.

·Hence the model NPI on distributions for distributions satisfying the positivity condition.

· $IPNFM$ model not substantively realistic for longitudinal data as it says for the last occassion $R_K \amalg \overline{L}_K | \overline{O}_{K-1}$ which does not depend on unobserved $L's$

· Robins 1997 give a substabtive non-longitudinal example re HIV testing where the $IPNFM$ model is plausible.

· Model also used without recognizing it when analyzing Cox model with censoring with missing covariates

·Robins et al 1999 describe how to add a non-identified sensitivity parameter encoding the magnitude of the violation of $R_k \amalg \overline{L}_k | \overline{O}_{k-1}, \underline{L}_{k+1}$ on a particular scale.

Estimation of $E[h(L_0, L_1)]$ under $IPNFM$ model

$\cdot h(L_0, L_1) = I(L_0 = l_0, L_1 = l_1)$ gives $E[h(L_0, L_1)] = f(l_0, l_1)$

$\cdot$Obtain Identifying Formula $E[h(L_0, L_1)]$

$$
\begin{aligned}
& E[h(L_0, L_1)] \\
= \;& E[\frac{R_1 R_0}{pr\{R_1 = 1 | \; R_0 = 1, L_1, L_0\} pr\{R_0 = 1 | \; L_1, L_0\}} h(L_0, L_1)] \\
= \;& E[\frac{R_1 R_0}{pr\{R_1 = 1 | \; R_0 = 1, L_0\} pr\{R_0 = 1 | \; L_1\}} h(L_0, L_1)]
\end{aligned}
$$

$\cdot$Thus need to identify $pr\{R_0 = 1 | \; L_1\}$.

$\cdot pr\{R_0 = 1|\ L_1\}$ is given by $pr\{R_0 = 1|\ L_1; \gamma\left(F_O\right)\}$ solving

$$E\left[\frac{R_1\left\{R_0 - pr\{R_0 = 1|\ L_1; \gamma\}q\left(L_1\right)\right\}}{pr\{R_1 = 1|\ R_0, R_0 L_0\}}\right] = 0$$

for a user supplied vector function $q\left(L_1\right)$ of γ which is the dim of the cardinality of the support of L_1

$\cdot$ Immediately leads to estimator baed on specifying parametric models for $pr\{R_1 = 1|\ R_0, R_0 L_0\}$ and $pr\{R_0 = 1|\ L_1\}$.

$\cdot$Greater robustness: We can get 2^K robustness as follows.

$\cdot$We suppose we are in a world with L_1 always observed.

-Full data still L_0, L_1 but now

–observed data $O_{pseudo} = R_0, R_0 L_0, L_1$

with $R_0 \amalg L_0 | L_1$as in $IPNFM$.

$\cdot$Then IF based on O_{pseudo} is

$$
\begin{aligned}
& if\left(O_{pseudo}\right) \\
= \;& u\left(O_{pseudo}\right) - E\left[h\left(L_0, L_1\right)\right] \\
& u\left(O_{pseudo}\right) \\
= \;& \frac{R_0 h\left(L_0, L_1\right)}{pr\{R_0 = 1|L_1\}} - \left\{\frac{R_0}{pr\{R_0 = 1|L_1\}} - 1\right\} E\left[h\left(L_0, L_1\right)|L_1, R_0 = 1\right]
\end{aligned}
$$

$\cdot$Now consider $O_{pseudo} = Full_{pseudo}$ as the (pseudo) full data and O as the observed data. Then the influence function of $E[u\left(Full_{pseudo}\right)] = E\left[h\left(L_0, L_1\right)\right]$ based on data O is

$$\frac{R_1 u\left(Full_{pseudo}\right)}{pr\{R_1 = 1| \ R_0, R_0 L_0\}} \\ -\left\{\frac{R_1}{pr\{R_1 = 1| \ R_0, R_0 L_0\}} - 1\right\} E\left[u\left(Full_{pseudo}\right)|R_0, R_0 L_0, R_1 = 1\right] \\ -E\left[h\left(L_0, L_1\right)\right]$$

$\cdot$ Note by using $u\left(Full_{pseudo}\right)$ rather than $\frac{R_0 h(L_0, L_1)}{pr\{R_0 = 1|L_1\}}$ in the last display we guarantee that we do not have to add another term due to the unknown $pr\{R_0 = 1|L_1\}$.

To obtain 2^2 multiple robustness we need DR estimators of the parametric working models $pr\{R_0 = 1|L_1; \tau\}$ and $E\left[h\left(L_0, L_1\right)|L_1, R_0 = 1; \lambda\right]$ for $pr\{R_0 = 1|L_1\}$ and $E\left[h\left(L_0, L_1\right)|L_1, R_0 = 1\right]$

which we obtain analogous to above.

·**Causal Inference Point of View**.

·Given treatments $A_0, A_1, ..., A_m$, and response Y_m measured after A_m,

·$Y_m(\overline{a}_m)$ is the value of Y_m if contrary to fact we interveneed and set treatment to $\overline{a}_m$.

·The observed Y_m is defined to be $Y_m\left(\overline{A}_m\right)$

·We observe $\overline{A}_K, Y_K$

If $Y_m\left(\overline{a}^{**}_{m-1}, a_m\right) = Y_m\left(\overline{a}^{*}_{m-1}, a_m\right)$, we can write $Y_m(a_m)$ has no direct effect $\overline{a}_{m-1}$ not through a_m on Y_m

Even more $Y_m(\overline{a}_K) = Y_m(a_m)$

Lets try this with missing data.

·Let $L_j^*\left(r_j=1\right)$ be the value of L_j^* that would be recorded if possibly contrary to fact I forced the jth variable to be observed

·Let $L_j^*\left(r_j=0\right)$ be the be the value of L_j^*that would be recorded if possibly contrary to fact I forced the jth variable to be unobserved. We have chosen the convention 0 but we could have chosen ·

·Then $L_j^*=L_j^*\left(R_j\right)$

·By convention we also call $L_j^*\left(r_j=1\right)$ by the name L_j

·Thus we find $L_j^*=R_jL_j$

·Note for missing data $L_j^*\left(\overline{r}_K\right)=L_j^*\left(r_j\right)$

·In graphs by changing intervention we do not need to put counterfactuals.

· Intervention effect of fixing a given variable is identified if all backdoor path blocked by descendants.

·The intervened variable given its parents is replaced in the distribution by indcator the variable takes it fixed vaue.is replaced in the probability distribution. Gives the counterfactual intervention distribution

Example 3 New Mohan and Pearl model : R_1

$$R_1 \amalg (R_0, L_1) \,|L_0$$
$$R_0 \amalg (R_1, L_0) \,|L_1$$

Can't fix as no blocked backdoor paths:

But

$$\begin{aligned} & E[\frac{R_1 R_0}{pr\{R_1 = 1|\ R_0 = 1, L_1, L_0\} pr\{R_0 = 1|\ L_1, L_0\}} h\,(L_0, L_1)] \\ = \ & E[\frac{R_1 R_0}{pr\{R_1 = 1|\ R_0 = 1, L_0\} pr\{R_0 = 1|\ L_1, R_1 = 1\}} h\,(L_0, L_1)] \end{aligned}$$

New result of Mohan and Pearl

Not non parametric.

F_O 8 parametrs. This model has 7.

So not NPI.

Conjecture: My model only graphical NPI model?