Treatment effect estimation with missing attributes

Julie Josse

École Polytechnique, INRIA
Visiting Researcher, Google Brain
Collaborators

Methods: Imke Mayer (PhD X, EHESS), Jean-Philippe Vert (Google Brain), Stefan Wager (Stanford)
Assistance Publique Hopitaux de Paris
Covid data

- 4780 patients (patients with at least one PCR-documented SARS-CoV-2 RNA from a nasopharyngeal sample)
- 119 continuous and categorical variables: **heterogeneous**
- 34 hospitals: **multilevel data**

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Treatment</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>DDI</th>
<th>BP</th>
<th>dead28</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>HCQ</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NM</td>
<td>180</td>
<td>yes</td>
</tr>
<tr>
<td>Pitie</td>
<td>AZ</td>
<td>76</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>131</td>
<td>no</td>
</tr>
<tr>
<td>Beaujon</td>
<td>HCQ+AZ</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>270</td>
<td>145</td>
<td>yes</td>
</tr>
<tr>
<td>Pitie</td>
<td>HCQ</td>
<td>80</td>
<td>f</td>
<td>NR</td>
<td>NR</td>
<td>107</td>
<td>no</td>
</tr>
<tr>
<td>HEGP</td>
<td>none</td>
<td>66</td>
<td>m</td>
<td>98</td>
<td>5890</td>
<td>118</td>
<td>no</td>
</tr>
</tbody>
</table>

⇒ Estimate causal effect: Administration of the treatment "Hydroxychloroquine" on the outcome 28-day mortality.
Covid data

- 4780 patients (patients with at least one PCR-documented SARS-CoV-2 RNA from a nasopharyngeal sample)
- 119 continuous and categorical variables: **heterogeneous**
- 34 hospitals: **multilevel data**

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Treatment</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>DDI</th>
<th>BP</th>
<th>dead28</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>HCQ</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NM</td>
<td>180</td>
<td>yes</td>
</tr>
<tr>
<td>Pitie</td>
<td>AZ</td>
<td>76</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>131</td>
<td>no</td>
</tr>
<tr>
<td>Beaujon</td>
<td>HCQ+AZ</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>270</td>
<td>145</td>
<td>yes</td>
</tr>
<tr>
<td>Pitie</td>
<td>HCQ</td>
<td>80</td>
<td>f</td>
<td>NR</td>
<td>NR</td>
<td>107</td>
<td>no</td>
</tr>
<tr>
<td>HEGP</td>
<td>none</td>
<td>66</td>
<td>m</td>
<td>98</td>
<td>5890</td>
<td>118</td>
<td>no</td>
</tr>
</tbody>
</table>

⇒ **Estimate causal effect**: Administration of the **treatment”Hydroxychloroquine”** on the **outcome** 28-day mortality.
**Observational data: non random assignment**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>survived</th>
<th>deceased</th>
<th>$\text{Pr}(\text{survived} \mid \text{treatment})$</th>
<th>$\text{Pr}(\text{deceased} \mid \text{treatment})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCQ</td>
<td>497 (11.4%)</td>
<td>111 (2.6%)</td>
<td>0.817</td>
<td>0.183</td>
</tr>
<tr>
<td>HCQ+AZI</td>
<td>158 (3.6%)</td>
<td>54 (1.2%)</td>
<td>0.745</td>
<td>0.255</td>
</tr>
<tr>
<td>none</td>
<td>2699 (62.1%)</td>
<td>830 (19.1%)</td>
<td>0.765</td>
<td>0.235</td>
</tr>
</tbody>
</table>

Mortality rate 23% - for HCQ 18% - non treated 24%: treatment helps?

![Age distribution comparison](image)

Comparison of the distribution of Age between HCQ and non treated.

Severe patients (with higher risk of death) are less likely to be treated. If control group does not look like treatment group, difference in response may be **confounded** by differences between the groups.
Potential outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect

- \( n \) iid samples \((X_i, W_i, Y_i(1), Y_i(0)) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}\)
- Individual causal effect of the treatment: \( \Delta_i \triangleq Y_i(1) - Y_i(0) \)
- Missing problem: \( \Delta_i \) never observed (only observe one outcome/indiv)

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Treatment</th>
<th>Outcome(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Y(0)</td>
<td>Y(1)</td>
</tr>
<tr>
<td>( X_1 )</td>
<td>( X_2 )</td>
<td>( X_3 )</td>
</tr>
<tr>
<td>1.1</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>-6</td>
<td>45</td>
<td>F</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>M</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>-2</td>
<td>52</td>
<td>M</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Survived</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>Dead</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Survived</td>
</tr>
<tr>
<td>-2</td>
<td>52</td>
<td>Survived</td>
</tr>
</tbody>
</table>
Potential outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect

- $n$ iid samples $(X_i, W_i, Y_i(1), Y_i(0)) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i \equiv Y_i(1) - Y_i(0)$
- Missing problem: $\Delta_i$ never observed (only observe one outcome/indiv)

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Treatment</th>
<th>Outcome(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$X_2$</td>
<td>$X_3$</td>
</tr>
<tr>
<td>1.1</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>-6</td>
<td>45</td>
<td>F</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>52</td>
<td>M</td>
</tr>
</tbody>
</table>

Average treatment effect (ATE): $\tau \equiv \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment.

ATE=0.05: mortality rate in the treated group is 5% points higher than in the control group. So, on average the treatment increases the risk of dying.
Assumption for ATE identifiability in observational data

**Unconfoundedness - selection on observables**

\[
\{ Y_i(0), Y_i(1) \} \perp \!
\!
\perp W_i \mid X_i
\]

Treatment assignment \( W_i \) is random conditionally on covariates \( X_i \).

Measure enough covariates to capture dependence between \( W_i \) and outcomes.

- Observed outcome: \( Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0) \)

**Unconfoundedness - graphical model**

Unobserved confounders make it impossible to separate correlation and causality when correlated to both the outcome and the treatment.

ATE not identifiable without assumption: it is not a sample size problem!
Assumption for ATE identifiability in observational data

Propensity score: probability of treatment given observed covariates.

<table>
<thead>
<tr>
<th>Propensity score - overlap assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e(x) \triangleq \mathbb{P}(W_i = 1 \mid X_i = x) \quad \forall x \in \mathcal{X}$.</td>
</tr>
</tbody>
</table>

We assume overlap, i.e. $\eta < e(x) < 1 - \eta$, $\forall x \in \mathcal{X}$ and some $\eta > 0$.

Left: Non smoker and never treated  
Right: Smokers and all treated

If proba to be treated when smoker $e(x) = 1$, how to estimate the outcome for smokers when not treated $Y(0)$? How to extrapolate if total confusion?
Inverse-propensity weighting estimation of ATE

Average treatment effect (ATE): \( \tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)] \)
Propensity score: \( e(x) \triangleq P(W_i = 1 | X_i = x) \)

**IPW estimator (Horvitz-Thomson, survey)**

\[
\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left( \frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right)
\]

⇒ Balance the differences between the two groups
⇒ Consistent estimator of \( \tau \) as long as \( \hat{e}(\cdot) \) is consistent.

*Credit: S. Athey*
Doubly robust ATE estimation

Model Treatment on Covariates $e(x) \triangleq \Pr(W_i = 1 \mid X_i = x)$
Model Outcome on Covariates $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \mid X_i = x]$

**Augmented IPW - Double Robust (DR)**

\[
\hat{\tau}_{AIPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)
\]

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

Possibility to use **any (machine learning) procedure** such as random forests, deep nets, etc. to estimate $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ without harming the interpretability of the causal effect estimation.

**Properties - Double Machine Learning (Chernozhukov et al., 2018)**

If $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ converge at the rate $n^{1/4}$ then
\[
\sqrt{n} \left( \hat{\tau}_{DR} - \tau \right) \xrightarrow{d} \mathcal{N}(0, V^*), \quad V^* \text{ semiparametric efficient variance.}
\]
Missing values

Percentage of missing values

Variable

- age
- gender
- num_hospitals
- on_corticoids
- period
- asthma
- cancer
- chemotherapy
- radiotherapy
- chronic obstructive pulmonary disease
- chronic respiratory failure
- diabetes
- dyslipidemia
- heart arrhythmia
- hematological malignancies
- ischemic heart disease
- kidney disease
- obesity
- smoker
- CREAT value
- CRP value
- PNN value
- LYM value
- TP value
- GOS_PaCO2 value
- GOS_PaO2 value
- weigh_kg
- LDH value
- DDI value

Percentage

0
20
40
60

Variable

- days_since_first_case_datetime
- num_hospitals
- on_corticoids
- period
- asthma
- cancer
- chemotherapy
- radiotherapy
- chronic obstructive pulmonary disease
- chronic respiratory failure
- diabetes
- dyslipidemia
- heart arrhythmia
- hematological malignancies
- ischemic heart disease
- kidney disease
- obesity
- smoker
- CREAT value
- CRP value
- PNN value
- LYM value
- TP value
- GOS_PaCO2 value
- GOS_PaO2 value
- weigh_kg
- LDH value
- DDI value
Deleting rows with missing values?

“One of the ironies of Big Data is that missing data play an ever more significant role” (R. Samworth, 2019)

An $n \times p$ matrix, each entry is missing with probability 0.01

$p = 5 \implies \approx 95\%$ of rows kept

$p = 300 \implies \approx 5\%$ of rows kept
Missing (informative) values in the covariates

Straightforward – but often biased – solution is complete-case analysis.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Treatment</th>
<th>Outcome(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$X_2$</td>
<td>$X_3$</td>
</tr>
<tr>
<td>NA</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>-6</td>
<td>45</td>
<td>NA</td>
</tr>
<tr>
<td>0</td>
<td>NA</td>
<td>M</td>
</tr>
<tr>
<td>NA</td>
<td>32</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>M</td>
</tr>
<tr>
<td>-2</td>
<td>NA</td>
<td>M</td>
</tr>
</tbody>
</table>

→ Often not a good idea! What are the alternatives?

Three families of methods - different assumptions

- Unconfoundedness with missingness + (no) missing values mechanisms Mayer, J., Wager, Sverdrup, Moyer, Gauss. AOAS 2020.
- Classical unconfoundedness + classical missing values mechanisms
1. Unconfoundedness with missing + (no) missing hypothesis

Unconfoundedness: \( \{Y_i(1), Y_i(0)\} \perp W_i \mid X \) not testable from the data.
\( \Rightarrow \) Doctors give us the DAG (covariates relevant for either treatment decision and for predicting the outcome)

Unconfoundedness with missing values: \( \{Y_i(1), Y_i(0)\} \perp W_i \mid X^* \)
\( X^* \triangleq (1 - R) \odot X + R \odot \text{NA}; \) with \( R_{ij} = 1 \) if \( X_{ij} \) is missing, \( 0 \) otherwise.
\( \Rightarrow \) Doctors decide to treat a patient based on what they observe/record. We have access to the same information as the doctors.
Under 1: Double Robust with missing values

AIPW with missing values

\[ \hat{\tau}^* \triangleq \frac{1}{n} \sum_i \left( \hat{\mu}^*_1(X_i) - \hat{\mu}^*_0(X_i) + W_i \frac{Y_i - \hat{\mu}^*_1(X_i)}{\hat{e}^*_1(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}^*_0(X_i)}{1 - \hat{e}^*_1(X_i)} \right) \]

Generalized propensity score (Rosenbaum and Rubin, 1984)

\[ e^*(x^*) \triangleq \mathbb{P}(W = 1 | X^* = x^*) \]

One model per pattern: \[ \sum_{r \in \{0,1\}} \mathbb{E} \left[ W | X_{obs}(r), R = r \right] 1_{R=r} \]

⇒ Supervised learning with missing values. \(^1\)

- Mean imputation is consistent with a universally consistent learner.
- Missing Incorporate in Attributes (MIA) for trees methods.

\(^1\)consistency of supervised learning with missing values J., Prost, Scornet, Varoquaux. JMLR 2020
Under 1: Double Robust with missing values

AIPW with missing values

\[ \hat{\tau}^* \triangleq \frac{1}{n} \sum_{i} \left( \hat{\mu}^*_1(X_i) - \hat{\mu}^*_0(X_i) + W_i \frac{Y_i - \hat{\mu}^*_1(X_i)}{e^*(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}^*_0(X_i)}{1 - e^*(X_i)} \right) \]

Generalized propensity score \((\text{Rosenbaum and Rubin, 1984})\)

\[ e^*(x^*) \triangleq P(W = 1 \mid X^* = x^*) \]

One model per pattern: \( \sum_{r \in \{0, 1\}} d \mathbb{E} [W \mid X_{\text{obs}(r)}, R = r] 1_{R=r} \)

\( \Rightarrow \) Supervised learning with missing values. \( ^1 \)

- Mean imputation is consistent with a universally consistent learner.
- Missing Incorporate in Attributes (MIA) for trees methods.

Implemented in \texttt{grf} package: combine two non-parametrics models, forests (conditional outcome and treatment assignment) adapted to any missing values with MIA.

\( \hat{\tau}_{\text{AIPW}}^* \) is \( \sqrt{n} \)-consistent, asymptotically normal given the product of RMSE of the nuisance estimates decay as \( o(n^{-1/2}) \) \( \text{Mayer et al. AOAS 2020} \)

\( ^1 \) consistency of supervised learning with missing values J., Prost, Scornet, Varoquaux. \( \text{JMLR 2020} \)
2. Classical unconfoundedness + missing values mechanism

Aparté on missing values mechanisms taxonomy *(Rubin, 1976)*

![Graphs showing Systolic Blood Pressure vs Gravity score (GCS)]

**MCAR** - **MAR** - **MNAR**

**Orange**: missing values for Systolic Blood Pressure - Gravity index (GCS) is always observed

MCAR (completely at random): Proba to be missing does not depend on SBP neither on gravity
MAR: Proba depends on gravity (we do not measure for too severe patients)
MNAR (not at random): Proba depends on SBP (low SBP not measured)
Consistency of IPW with missing values \cite{Seaman2014}

Assume **Missing At Random (MAR)** mechanism. Multiple imputation (MICE using \(X^*, W, Y\)) with IPW on each imputed data is consistent when Gaussian covariates and logistic/linear treatment/outcome model

\[
\begin{array}{cccccc}
X_1^* & X_2^* & X_3^* & \ldots & W & Y \\
NA & 20 & 10 & \ldots & 1 & \text{survived} \\
-6 & 45 & NA & \ldots & 1 & \text{survived} \\
0 & NA & 30 & \ldots & 0 & \text{died} \\
NA & 32 & 35 & \ldots & 0 & \text{survived} \\
-2 & NA & 12 & \ldots & 0 & \text{died} \\
1 & 63 & 40 & \ldots & 1 & \text{survived} \\
\end{array}
\]

1) Generate \(M\) plausible values for each missing value

2) Estimate ATE on each imputed data set: \(\hat{\tau}_m, \text{Var}(\hat{\tau}_m)\)

3) Combine the results (Rubin’s rules): \(\hat{\tau} = \frac{1}{M} \sum_{m=1}^{M} \hat{\tau}_m\)

\[
\text{Var}(\hat{\tau}) = \frac{1}{M} \sum_{m=1}^{M} \text{Var}(\hat{\tau}_m) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\tau}_m - \hat{\tau})^2
\]
3. Latent unconfoundedness + missing values mechanism

**Latent confounding assumption**

The covariates \( X \) are **noisy (incomplete) proxies** of the true **latent confounders** \( Z \) (Kallus et al., 2018; Louizos et al., 2017).

\[
X^* \triangleq (1 - R) \odot X + R \odot NA; \text{ with } R_{ij} = 1 \text{ if } X_{ij} \text{ is missing, } 0 \text{ otherwise}
\]

Observed outcome: \( Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0) \)
3. Latent unconfoundedness + missing values mechanism

**Latent confounding assumption**

The covariates $X$ are **noisy (incomplete) proxies** of the true **latent confounders** $Z$ (Kallus et al., 2018; Louizos et al., 2017).

$$X^* \triangleq (1 - R) \odot X + R \odot NA; \text{ with } R_{ij} = 1 \text{ if } X_{ij} \text{ is missing, } 0 \text{ otherwise}$$

Observed outcome: $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$

![Diagram showing the relationships between $Z$, $X^*$, $X$, $W$, $R$, and $Y$]

**Matrix Factorization as a pre-processing step**

- Assume data are generated as a low-rank structure corrupted by noise. Estimate $Z$ using matrix completion from $X^*$ (softimpute types).

- Plug $\hat{Z}$ in regression model of outcome on treatment and confounders:
  $$Y = \tau W + Z \beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \text{ (or in the (A)IPW estimators)}$$

- Kallus et al. (2018) show that $\hat{\tau}$ is a consistent estimator under MCAR of the Average Treatment Effect.
# 3. Latent unconfoundedness + missing values mechanism

## Latent confounding assumption

Covariates $X_{n \times d}$ proxies of the latent confounders $Z_{n \times q}$.

$X^* \triangleq (1 - R) \odot X + R \odot NA$; with $R_{ij} = 1$ if $X_{ij}$ is missing, 0 otherwise.

![Diagram of causal model](image)

- MissDeepCausal (MDC) **Mayer, J., Raimundo, Vert, 2020.**
  - Assume a Deep Latent Variable Model instead of linear factor analysis
  - Leverage VAE with MAR values (Mattei and Frellsen, 2019). Imputing NA with 0 maximizes an ELBO of the observed log-likelihood.
  - Draw $(Z^{(j)})_{1 \leq j \leq B}$ from the posterior distribution $P(Z|X^*)$ (using importance sampling with $Q(Z|X^*)$ for proposal).

**MDC-Multiple Imputation**: estimate ATE on each $(Z^{(j)})$

**MDC-process** plug-in $\hat{Z}(x^*) \triangleq \mathbb{E}[Z|X^* = x^*]$ in classical estimators

Flexible with promising empirical results.
### Methods to do causal inference with missing values

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Missingness</th>
<th>Unconfoundedness</th>
<th>Models for ((W, Y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>multivariate normal</td>
<td>general</td>
<td>M(C)AR general</td>
<td>Missing Latent Classical logistic-linear non-param.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1. (SA)EM</strong>(^2)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>1. Mean.GRF</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>1. MIA.GRF</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>2. Mult. Imp.</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>3. MatrixFact.</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>3. MissDeep-Causal</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Methods & assumptions on data generating process (models for covariates, outcome, treatment), missing values mechanism and identifiability conditions.

✓: can be handled  ❌: not applicable in theory

(✓): empirical results and ongoing work on theoretical guarantees

(❌): no theoretical guarantees but heuristics.

\(^2\)Use of EM algorithms for logistic regression with missing values. [Jiang et al. (2020)]
Simulations: no overall best performing method.

- 10 covariates generated with Gaussian mixture model \( X_i \sim \mathcal{N}_d(\mu(c_i), \Sigma(c_i)) | C_i = c_i, C \) from a multinomial distribution with three categories.
- Unconfoundedness on complete/observed covariates, 30% NA
- Logistic-linear for \((W, Y), \logit(e(X_i.)) = \alpha^T X_i., Y_i \sim \mathcal{N}(\beta^T X_i. + \tau W_i, \sigma^2)\)

**Figure 1:** Estimated with AIPW and true ATE \(\tau = 1\)

\[ \rightarrow \text{GRF-MIA is asymptotically unbiased under unconfoundedness despite missingness.} \]
\[ \rightarrow \text{Multiple imputation requires many imputations to remove bias.} \]
Simulations: no overall best performing method.

- 100 covariates generated with a DLVM model, latent confounding ($q = 3$): $Z_i \sim \mathcal{N}_q(0, \sigma_z)$, covariates $X_i$ sampled from $\mathcal{N}_d(\mu(Z), \Sigma(Z))$, where $(\mu(Z), \Sigma(Z)) = (V \tanh(UZ + a) + b, \text{diag}\{\exp(\eta^T \tanh(UZ + a) + \delta)\})$ with $U, V, a, b, \delta, \eta$ drawn from standard Gaussian and uniform distributions.

- 30% MCAR, $n = 1000$.

- Logistic-linear for $(W, Y)$, $\logit(e(Z_i.)) = \alpha^T Z_i., Y_i \sim \mathcal{N}(\beta^T Z_i. + \tau W_i, \sigma^2)$

Figure 1: Estimated with AIPW and true ATE $\tau = 1$.

→ MDC empirically unbiased if number of features ($d$) $>>$ dim of the latent space ($q$)

Tuning: variance of the prior of $Z$ and $\hat{q}$ chosen by cross-validation using the ELBO
Results for Covid Patients

33 covariates, 26 confounders. 4137 patients.

ATE estimations ($\times$100): effect of Hydroxychloroquine on 28day mortality

(y-axis: estimation approach, solid: Doubly Robust AIPW, dotted: IPW), (x-axis: ATE estimation with CI)

The obtained value corresponds to the difference in percentage points between mortality rates in treatment and control.

Light Blue: unadjusted (-5.3)
Conclusion and perspectives

Take-away messages

- **Missing attributes** alter causal analyses. Performance of methods depends on the underlying assumptions.
Conclusion and perspectives

Take-away messages

- **Missing attributes** alter causal analyses. Performance of methods depends on the underlying assumptions

Further details in original papers


Future work

- Coupling of observational data and RCT data
- Heterogeneous treatment effects
- Architecture of neural nets with missing values
- More with MNAR data
Missing value website

More information and details on missing values: **R-miss-tastic** platform. (Mayer et al., 2019)

→ Theoretical and practical tutorials, popular datasets, bibliography, workflows (in R and in python), active contributors/researchers in the community, etc.

rmissstastic.netlify.com

Interested in contribute to our platform? Feel free to contact us!
MERCI
References


Observational data: non random assignment

⇒ Treatment assignment $W$ depends on covariates $X$.
Distribution of covariates of treated and control are different.
1. Unconfoundedness despite missingness

Adapt the initial assumptions s.t. treatment assignment is unconfounded given only the observed covariates and the response pattern.

Notations

mask $R \in \{0, 1\}^d$, $R_{ij} = 1$ when $X_{ij}$ is missing and 0 otherwise

$X^* \triangleq (1 - R) \odot X + R \odot NA \in \{\mathbb{R} \cup NA\}^d$

Unconfoundedness despite missingness

$$\{Y_i(1), Y_i(0)\} \perp \perp W_i \mid X^*$$

CIT: $W_i \perp \perp X_i \mid X^*_i, R_i$ or CIO: $Y_i(w) \perp \perp X_i \mid X^*_i, R_i$ for $w \in \{0, 1\}$
Mean imputation

- \((x_i, y_i) \sim \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})\)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.56</td>
<td>-1.93</td>
</tr>
<tr>
<td>-0.86</td>
<td>-1.50</td>
</tr>
<tr>
<td>.....</td>
<td>...</td>
</tr>
<tr>
<td>2.16</td>
<td>0.7</td>
</tr>
<tr>
<td>0.16</td>
<td>0.74</td>
</tr>
</tbody>
</table>

\(\hat{\mu}_y = -0.01\)
\(\hat{\sigma}_y = 1.01\)
\(\hat{\rho} = 0.66\)

\(\mu_y = 0\)
\(\sigma_y = 1\)
\(\rho_{xy} = 0.6\)
Mean imputation

- $(x_i, y_i) \sim \mathcal{N}_2(\mu_x, \mu_y, \Sigma_{xy})$
- 70% of missing entries completely at random on $Y$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.56</td>
<td>NA</td>
</tr>
<tr>
<td>-0.86</td>
<td>NA</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2.16</td>
<td>0.7</td>
</tr>
<tr>
<td>0.16</td>
<td>NA</td>
</tr>
</tbody>
</table>

$\hat{\mu}_y = 0.18$
$\hat{\sigma}_y = 0.9$
$\hat{\rho}_{xy} = 0.6$

$\mu_y = 0$
$\sigma_y = 1$
$\rho_{xy} = 0.6$
Mean imputation

- \((x_i, y_i) \sim \mathcal{N}_2(\mu_x, \mu_y, \Sigma_{xy})\)
- 70% of missing entries completely at random on \(Y\)
- Estimate parameters on the mean imputed data

Mean imputation deforms joint and marginal distributions

<table>
<thead>
<tr>
<th>(X)</th>
<th>(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.56</td>
<td>0.01</td>
</tr>
<tr>
<td>-0.86</td>
<td>0.01</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>2.16</td>
<td>0.7</td>
</tr>
<tr>
<td>0.16</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\hat{\mu}_y &= 0.01 \\
\hat{\sigma}_y &= 0.5 \\
\hat{\rho} &= 0.30
\end{align*}
\]
Mean imputation is bad for estimation

Ecological data: \( n = 69000 \) species - 6 traits. Estimated correlation between Pmass & Rmass \( \approx 0 \) (mean imputation) or \( \approx 1 \) (EM PCA)

Imputation methods

- by regression takes into account the relationship: Estimate $\beta$ - impute 
  $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Rightarrow \text{variance underestimated and correlation overestimated}$$
- by stochastic reg: Estimate $\beta$ and $\sigma$ - impute from the predictive 
  $$y_i \sim \mathcal{N}(x_i\hat{\beta}, \hat{\sigma}^2) \Rightarrow \text{preserve distributions}$$

Here $\hat{\beta}, \hat{\sigma}^2$ estimated with complete data, but MLE can be obtained with EM

<table>
<thead>
<tr>
<th>Mean imputation</th>
<th>Regression imputation</th>
<th>Stochastic regression imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_y = 0$</td>
<td>$\sigma_y = 1$</td>
<td>$\rho_{xy} = 0.6$</td>
</tr>
<tr>
<td>0.01</td>
<td>0.5</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.72</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.59</td>
</tr>
</tbody>
</table>
Imputation methods for multivariate data

Assuming a joint model

- Gaussian distribution: \( x_i \sim \mathcal{N}(\mu, \Sigma) \) (Amelia Honaker, King, Blackwell)
- low rank: \( X_{n \times d} = \mu_{n \times d} + \varepsilon_{ij} \sim iid \mathcal{N}(0, \sigma^2) \) with \( \mu \) of low rank \( k \) (softimpute Hastie & Mazuder; missMDA J. & Husson)
- latent class - nonparametric Bayesian (dpmpm Reiter)
- deep learning using variational autoencoders (MIWAE, Mattei, 2018)

Using conditional models (joint implicitly defined)

- with logistic, multinomial, poisson regressions (mice van Buuren)
- iterative impute each variable by random forests (missForest Stekhoven)

Imputation for categorical, mixed, blocks/multilevel data, etc.

⇒ Missing values plateform \(^5\) J., Mayer., Tierney, Vialaneix

---

\(^5\) https://rmisstastic.netlify.com/
Mean imputation consistent

Learn on the mean-imputed training data, impute the test set with the same means and predict is optimal if the missing data are MAR and the learning algorithm is universally consistent.

Framework - assumptions

- \( Y = f(X) + \varepsilon \)
- \( X = (X_1, \ldots, X_d) \) has a continuous density \( g > 0 \) on \([0, 1]^d\)
- \( \|f\|_\infty < \infty \)
- Missing data MAR on \( X_1 \) with \( R_1 \perp X_1 | X_2, \ldots, X_d \).
- \((x_2, \ldots, x_d) \mapsto P[R_1 = 1 | X_2 = x_2, \ldots, X_d = x_d] \) is continuous
- \( \varepsilon \) is a centered noise independent of \((X, R_1)\)

(remains valid when missing values occur for variables \( X_1, \ldots, X_j \))
Mean imputation consistent

Learn on the mean-imputed training data, impute the test set with the **same means** and predict is optimal if the missing data are MAR and the **learning algorithm is universally consistent**

Mean imputed entry $x' = (x'_1, x_2, \ldots, x_d)$: $x'_1 = x_1 \mathbb{1}_{R_1=0} + \mathbb{E}[X_1] \mathbb{1}_{R_1=1}$

$\tilde{X} = X \odot (1 - R) + NA \odot R$ (takes value in $\mathbb{R} \cup \{NA\}$)

**Theorem**

Prediction with mean is equal to the Bayes function almost everywhere

$$f_{impute}^*(x') = \mathbb{E}[Y|X^* = x^*]$$

Other values than the mean are OK but use the same value for the train and test sets, otherwise the algorithm may fail as the distributions differ
Consistency of supervised learning with NA: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant:
- Need a lot of data (asymptotic result) and a super powerful learner

Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!
Consistency of supervised learning with NA: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time
- With categorical data, just code ”Missing”
- With continuous data, any constant: out of range
- Need a lot of data (asymptotic result) and a super powerful learner

Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!
Consistency: 40% missing values MCAR

- Linear problem (high noise)
- Friedman problem (high noise)
- Non-linear problem (low noise)

Sample size

Explained variance

- Surrogates (rpart)
- Gaussian imputation
- Mean imputation
- MIA
- Bayes rate
- Block (XGBoost)
End-to-end learning with missing values

- Random forests powerful learner
- Trees well suited for empirical risk minimization with missing values: Handle half discrete data $X^*$ that takes values in $\mathbb{R} \cup \{\text{NA}\}$
CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^*$, the threshold $z^*$ which minimises the (quadratic) loss

$$(j^*, z^*) \in \arg \min_{(j, z) \in S} \mathbb{E} \left[ (Y - \mathbb{E}[Y|X_j \leq z])^2 \cdot 1_{X_j \leq z} + (Y - \mathbb{E}[Y|X_j > z])^2 \cdot 1_{X_j > z} \right].$$
CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^*$, the threshold $z^*$ which minimises the (quadratic) loss

$$(j^*, z^*) \in \arg\min_{(j,z) \in S} \mathbb{E} \left[ (Y - \mathbb{E}[Y | X_j \leq z])^2 \cdot 1_{X_j \leq z} ight. \
\left. + (Y - \mathbb{E}[Y | X_j > z])^2 \cdot 1_{X_j > z} \right].$$
CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^*$, the threshold $z^*$ which minimises the (quadratic) loss

$$(j^*, z^*) \in \arg \min_{(j, z) \in S} \mathbb{E} \left[ (Y - \mathbb{E}[Y | X_j \leq z])^2 \cdot 1_{X_j \leq z} ight. + (Y - \mathbb{E}[Y | X_j > z])^2 \cdot 1_{X_j > z} \left. \right].$$
## CART with missing values

### Table

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**root**

### Equations

1. Select variable and threshold on observed data
   
   \[
   E \left[ (Y - E[Y | X_j \leq z, R_j = 0]) \cdot 1_{X_j \leq z, R_j = 0} + (Y - E[Y | X_j > z, R_j = NA]) \cdot 1_{X_j > z, R_j = 0} \right].
   \]

2. Propagate observations (2 & 3) with missing values?
   - Probabilistic split: Bernoulli ($\#L / (\#L + \#R)$)
   - Block: Send all to a side by minimizing the error (xgboost, lightgbm)
   - Surrogate split: Search another variable that gives a close partition (rpart)

### Variable selection bias (not a problem to predict):

- partykit package, Hothorn, et al.
CART with missing values

1. Select variable and threshold on observed data

\[ E \left[ \left( Y - E[Y|X_j \leq z, R_j = 0] \right)^2 \cdot 1_{X_j \leq z, R_j = 0} + \left( Y - E[Y|X_j > z, R_j = NA] \right)^2 \cdot 1_{X_j > z, R_j = 0} \right]. \]

2. Propagate observations (2 & 3) with missing values?
   - Probabilistic split: \text{Bernoulli}(\#L/\#L + \#R) \text{ (Rweeka)}
   - Block: Send all to a side by minimizing the error (xgboost, lightgbm)
   - Surrogate split: Search another variable that gives a close partition (rpart)

\[ ^6 \text{ Variable selection bias (not a problem to predict): partykit package, Hothorn, et al.} \]
CART with missing values

1) Select variable and threshold on observed data

\[
\mathbb{E} \left[ (Y - \mathbb{E}[Y|X_j \leq z, R_j = 0])^2 \cdot 1_{X_j \leq z, R_j = 0} + (Y - \mathbb{E}[Y|X_j > z, R_j = NA])^2 \cdot 1_{X_j > z, R_j = 0} \right].
\]

2) Propagate observations (2 & 3) with missing values?

- Probabilistic split: \textit{Bernoulli}(\frac{\#L}{\#L + \#R}) (Rweeka)
- Block: Send all to a side by minimizing the error (xgboost, lightgbm)
- Surrogate split: Search another variable that gives a close partition (rpart)

---

\[\text{Variable selection bias (not a problem to predict): partykit package, Hothorn, et al.}\]
One step: select the variable, the threshold and propagate missing values. Use missingness to make the best possible splits.

\[ f^* \in \arg \min_{f \in \mathcal{P}_{c, \text{miss}}} \mathbb{E}\left[(Y - f(X^*))^2\right], \]

where \( \mathcal{P}_{c, \text{miss}} = \mathcal{P}_{c, \text{miss}, L} \cup \mathcal{P}_{c, \text{miss}, R} \cup \mathcal{P}_{c, \text{miss}, \text{sep}} \) with

1. \( \mathcal{P}_{c, \text{miss}, L} \rightarrow \{\{X_j^* \leq z \lor X_j^* = \text{NA}\}, \{X_j^* > z\}\} \)
2. \( \mathcal{P}_{c, \text{miss}, R} \rightarrow \{\{X_j^* \leq z\}, \{X_j^* > z \lor X_j^* = \text{NA}\}\} \)
3. \( \mathcal{P}_{c, \text{miss}, \text{sep}} \rightarrow \{\{X_j^* \neq \text{NA}\}, \{X_j^* = \text{NA}\}\} \).

- Missing values treated like a category (well to handle \( \mathbb{R} \cup \text{NA} \))
- Good for informative pattern (\( R \) explains \( Y \))
- Implementation trick: duplicate the incomplete columns, and replace the missing entries once by \(+\infty\) and once by \(-\infty\) (J. Tibshirani)\(^7\)

Target model/pattern: \( \mathbb{E}[Y|X^*] = \sum_{r \in \{0,1\}^d} \mathbb{E}[Y|X_{\text{obs}(r)}, R = r] \mathbb{1}_{R=r} \)

Does not require the missing data to be MAR.

\(^7\)Implemented for conditional forests partykit, generalized random forest grf, scikitlearn