# What is a good imputation under MAR missingness

**Julie Josse**
Head of the Inria-Inserm team **PreMeDICaL**:
"**Pre**cision **Me**dicine by **D**ata **I**ntegration & **Ca**usal **L**earning"

April 16, 2024

Näf et al. (2024)
(https://arxiv.org/abs/2403.19196)

Jeffrey Näf

# Traumabase project: decision support for trauma patients

- ▷ 30 000 French trauma patients[1]
- ▷ 250 features from the accident site to the hospital discharge
- ▷ 30 hospitals
- ▷ 4000 new patients/ year

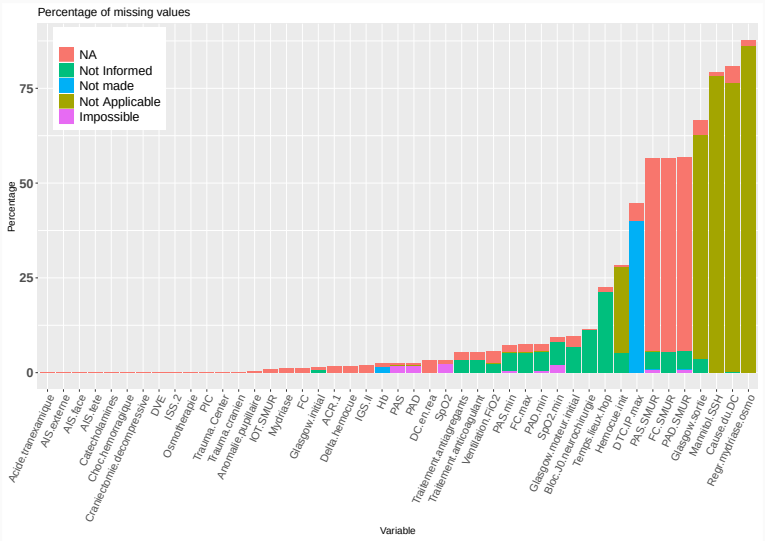
| Center | Accident | Age | Sex | Weight | Lactactes | BP | **TXA.** | $Y$ |
|---|---|---|---|---|---|---|---|---|
| Beaujon | fall | 54 | m | 85 | NM | 180 | treated | 0 |
| Pitie | gun | 26 | m | NR | NA | 131 | untreated | 1 |
| Beaujon | moto | 63 | m | 80 | 3.9 | 145 | treated | 1 |
| Pitie | moto | 30 | w | NR | Imp | 107 | untreated | 0 |
| HEGP | knife | 16 | m | 98 | 2.5 | 118 | treated | 1 |
| ⋮ | | | | | | | | ⋱ |

⇒ **Estimate causal effect**: Administration of the **treatment** *tranexamic acid (TXA)*, given within 3 hours of the accident, on the **outcome** $Y$ *28 days intra hospital mortality* for trauma brain patients

TXA decreases mortality for extra-cranial bleeding. Effect for intra-cranial bleeding? (detected by CT scan). TXA is one of the first treatments given

[1] `www.traumabase.eu - https://www.traumatrix.fr/`

# Missing data: important bottleneck in statistical practice



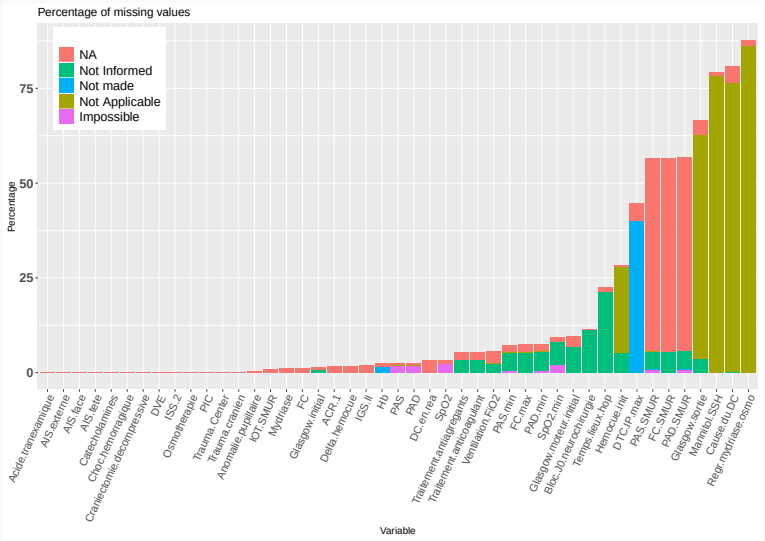

*"One of the ironies of Big Data is that missing data play an ever more significant role"*[2]

[2] Zhu, Wang, Samworth. High-dimensional PCA with heterogeneous missingness. *JRSSB*. 2022.

[3] J., et al. `https://rmisstastic.netlify.com/` - Tutorial JJ diableret.

[4] R Taskview https://cran.r-project.org/web/views/MissingData.html

# Missing data: important bottleneck in statistical practice




*"One of the ironies of Big Data is that missing data play an ever more significant role"*[2]

Complete-case analysis, often not a good idea! What are the alternatives?

Inferential aim: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$
Matrix completion aim: **Predict the missing values** - low rank approx.
Predictive aim: **Predict an outcome** with missing values in covariates
Rmistatic $>$ 150 packages,[3][4]

[2] Zhu, Wang, Samworth. High-dimensional PCA with heterogeneous missingness. *JRSSB*. 2022.
[3] J., et al. `https://rmisstastic.netlify.com/` - Tutorial JJ diableret.
[4] R Taskview https://cran.r-project.org/web/views/MissingData.html

# Missing values alter causal analyses

| Covariates | | | Treatment | Outcome(s) | |
|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | W | Y(0) | Y(1) |
| NA | 20 | F | 1 | ? | 200 |
| -6 | 45 | NA | 0 | 10 | ? |
| 0 | NA | M | 1 | ? | 150 |
| NA | 32 | F | 1 | ? | 100 |
| 1 | 63 | M | 1 | 15 | ? |
| -2 | NA | M | 0 | 20 | ? |

## Both causal and missing assumptions

1. Classical unconfoundedness + classical missing values mechanisms[5]
2. Unconfoundedness with missing + (no) missing values mechanisms[6]
3. Latent unconfoundedness + MCAR[7]

[5] Seaman and White. IPW with missing predictors of treatment assignment, *Communications in Statistics, Theory & Methods*. 2014.

[6] Mayer, Wager, **J.** Doubly robust estimation with incomplete confounders. *AOAS*. 2020.

[7] Kallus et al. Causal inf. with noisy & missing covariates via matrix factorization. *Neurips*. 2018.

# 1. Popular multiple imputation for estimating treatment effect

| $X_1$ | $X_2$ | $X_3$ | ... | W | Y(0) | Y(1) |
|---|---|---|---|---|---|---|
| NA | 20 | 10 | ... | 1 | ? | 200 |
| -6 | 45 | NA | ... | 1 | 10 | ? |
| 0 | NA | 30 | ... | 0 | ? | 150 |
| NA | 32 | 35 | ... | 0 | ? | 100 |
| -2 | NA | 12 | ... | 0 | 20 | ? |

1) Generate $M$ plausible values for each missing value

| $X_1$ | $X_2$ | $X_3$ | ... | W | Y | $X_1$ | $X_2$ | $X_3$ | ... | W | Y | $X_1$ | $X_2$ | $X_3$ | ... | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 20 | 10 | ... | 1 | 200 | -7 | 20 | 10 | ... | 1 | 200 | 7 | 20 | 10 | ... | 1 | 200 |
| -6 | 45 | 6 | ... | 1 | 10 | -6 | 45 | 9 | ... | 1 | 10 | -6 | 45 | 12 | ... | 1 | 10 |
| 0 | 4 | 30 | ... | 0 | 150 | 0 | 12 | 30 | ... | 0 | 150 | 0 | -5 | 30 | ... | 0 | 150 |
| -4 | 32 | 35 | ... | 0 | 100 | 13 | 32 | 35 | ... | 0 | 100 | 2 | 32 | 35 | ... | 0 | 100 |
| -2 | 15 | 12 | ... | 0 | 20 | -2 | 10 | 12 | ... | 0 | 20 | -2 | 20 | 12 | ... | 0 | 20 |

2) Estimate Average Treatment Effect on each imputed data set with IPW: $\hat{\tau}_m$

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

## Consistency of multiple imputation with IPW[8]

Assume: **MAR** Proba to have missing depends on observed values

Classical **unconfoundedness** $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$,

Propensity Score and model for $(X \mid Y, W)$ correctly specified,

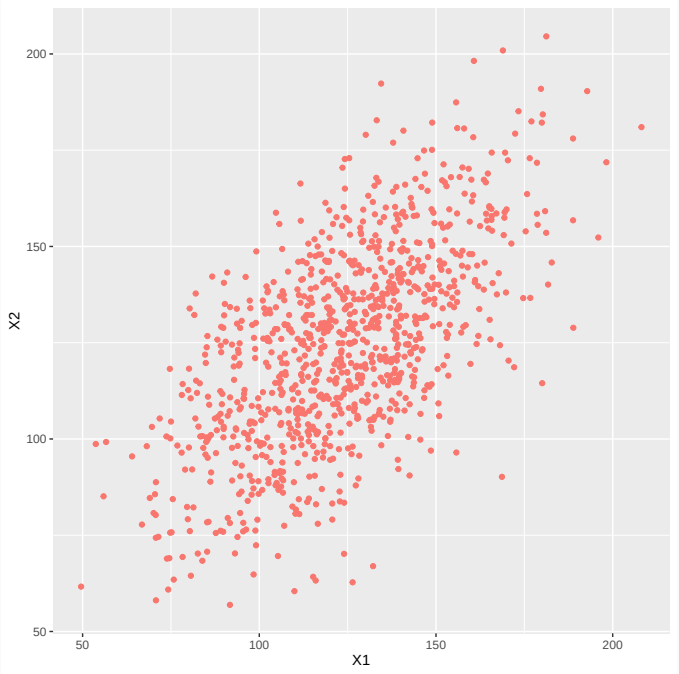
$\Rightarrow$ Multiple imputation (using $(X, W, Y)$) with IPW is consistent

[8] Seaman and White. 2014. IPW with missing predictors of treatment assignment, *Communications in Statistics, Theory & Methods*.

# Single imputation by the mean

▷ $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$

| $\mathbf{X}_1$ | $\mathbf{X}_2$ |
|---|---|
| -0.56 | -1.93 |
| -0.86 | -1.50 |
| ..... | ... |
| 2.16 | 0.7 |
| 0.16 | 0.74 |




$\mu_{x_2} = 0$

$\sigma_{x_2} = 1$

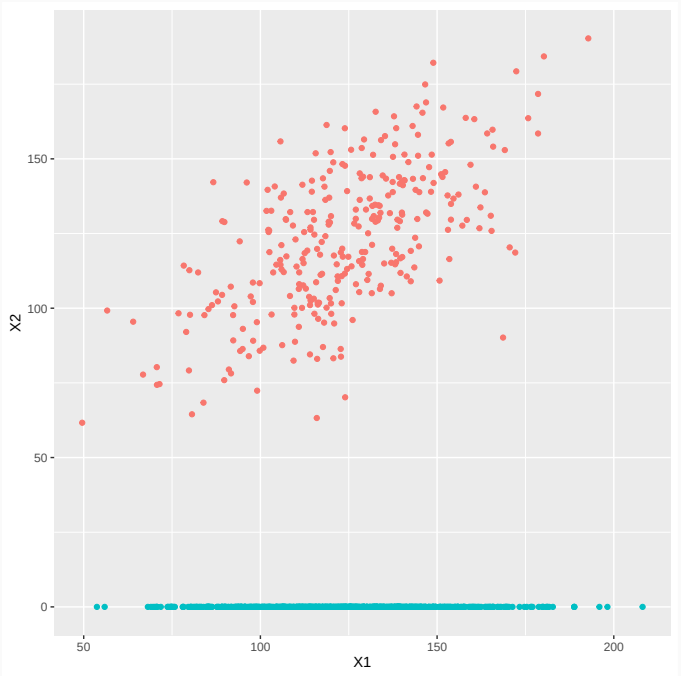
$\rho = 0.6$

| |
|---|
| $\hat{\mu}_{x_2} = -0.01$ |
| $\hat{\sigma}_{x_2} = 1.01$ |
| $\hat{\rho} = 0.66$ |

# Single imputation by the mean

- $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$
- 70 % of missing entries completely at random on $X_2$

| $\mathbf{X}_1$ | $\mathbf{X}_2$ |
|---|---|
| -0.56 | NA |
| -0.86 | NA |
| ..... | ... |
| 2.16 | 0.7 |
| 0.16 | NA |



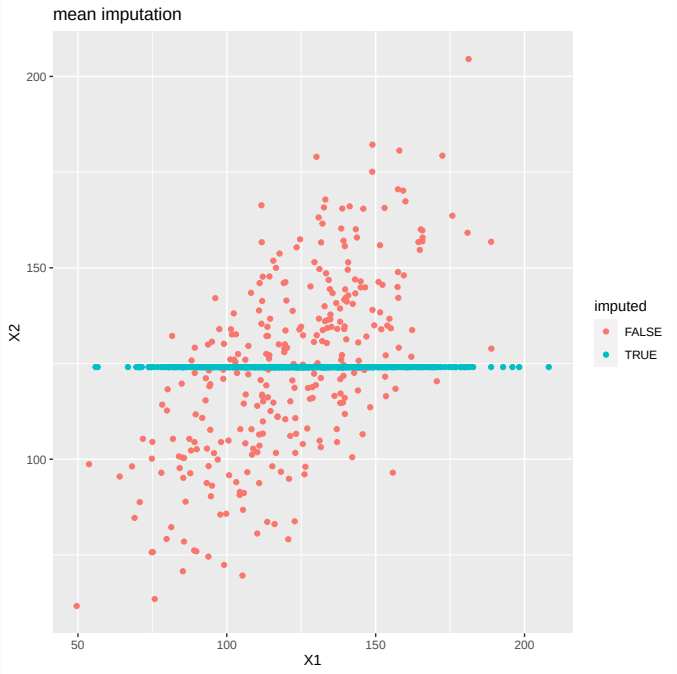

$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

| |
|---|
| $\hat{\mu}_{x_2} = 0.18$ |
| $\hat{\sigma}_{x_2} = 0.9$ |
| $\hat{\rho} = 0.6$ |

# Single imputation by the mean

- $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$
- 70 % of missing entries completely at random on $X_2$
- Estimate parameters on the mean imputed data

| $\mathbf{X}_1$ | $\mathbf{X}_2$ |
|---|---|
| -0.56 | **0.01** |
| -0.86 | **0.01** |
| ..... | ... |
| 2.16 | 0.7 |
| 0.16 | **0.01** |




$\mu_{x_2} = 0$

$\sigma_{x_2} = 1$

$\rho = 0.6$

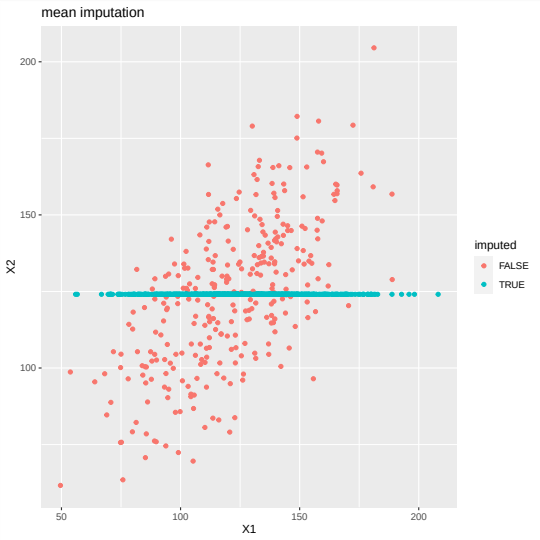
| |
|---|
| $\hat{\mu}_{x_2} = 0.01$ |
| $\hat{\sigma}_{x_2} = 0.5$ |
| $\hat{\rho} = 0.30$ |

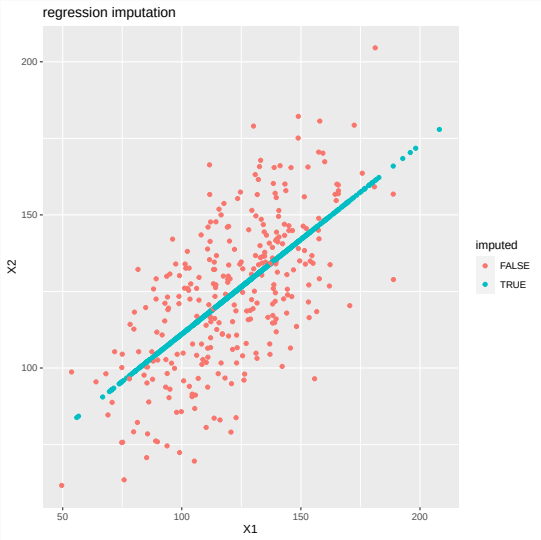
Mean imputation deforms joint and marginal distributions

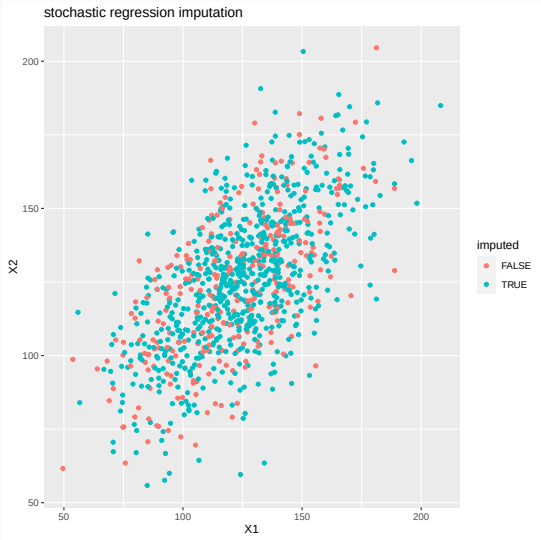
# Objective: to impute while preserving distribution

Assuming a bivariate gaussian distribution $x_{i2} = \beta_0 + \beta_1 x_{i1} + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

- ▷ Regression imputation: Estimate $\beta$ (here with complete data) and impute $\hat{x}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} \Rightarrow$ variance underestimated and correlation overestimated
- ▷ Stochastic reg. imputation: Estimate $\beta$ and $\sigma$ - impute from the predictive $\hat{x}_{i2} \sim \mathcal{N}\left(\beta_0 + \hat{\beta}_1 x_{i1}, \hat{\sigma}^2\right) \Rightarrow$ preserve distributions




$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

| |
|---|
| 0.01 |
| 0.5 |
| 0.30 |

| |
|---|
| 0.01 |
| 0.72 |
| 0.78 |

| |
|---|
| 0.01 |
| 0.99 |
| 0.59 |

# Impute while preserving distribution. Multivariate case

## Assuming a joint distribution

- Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
- Low rank : $X_{n\times d} = \mu_{n\times d} + \varepsilon\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\ \sigma^2\right)$ with $\mu$ of low rank
  - $\Rightarrow$ Powerful in recommendation system: Netflix prize 90% of missing
  - $\Rightarrow$ Different regularization depending on noise regime[9]
  - $\Rightarrow$ Count data,[10] ordinal data, categorical data, blocks/multilevel data[11]
- Using optimal transport,[12] deep generative models (GAIN,[13] MIWAE,[14] etc.)

[9] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*. 2016.
[10] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[11] **J.** et al. Imputation of mixed data with multilevel SVD. *JCGS*. 2018.
[12] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[13] Yoon et al. GAIN: Missing data imputation using generative adversarial nets. *ICML*. 2018.
[14] Mattei & Frellsen. Miwae: Deep generative model. & imput. of incomplete data. *ICML*. 2018.
[15] Stekhoven & Bühlmann. MissForest–non-parametric imputation for mixed data. *Bioinfo. 2012.*

# Impute while preserving distribution. Multivariate case

## Assuming a joint distribution

- Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
- Low rank : $X_{n\times d} = \mu_{n\times d} + \varepsilon\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\ \sigma^2\right)$ with $\mu$ of low rank
  - $\Rightarrow$ Powerful in recommendation system: Netflix prize 90% of missing
  - $\Rightarrow$ Different regularization depending on noise regime[9]
  - $\Rightarrow$ Count data,[10] ordinal data, categorical data, blocks/multilevel data[11]
- Using optimal transport,[12] deep generative models (GAIN,[13] MIWAE,[14] etc.)

## Iterating conditional models (joint distribution implicitly defined)

- with multinomial, Poisson regression (ICE: Imputation by Chained Equations)
- iterative imputation of each variable by random forests[15]

---

[9] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*. 2016.
[10] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[11] **J.** et al. Imputation of mixed data with multilevel SVD. *JCGS*. 2018.
[12] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[13] Yoon et al. GAIN: Missing data imputation using generative adversarial nets. *ICML*. 2018.
[14] Mattei & Frellsen. Miwae: Deep generative model. & imput. of incomplete data. *ICML*. 2018.
[15] Stekhoven & Bühlmann. MissForest–non-parametric imputation for mixed data. *Bioinfo. 2012.*

# Missing values mechanism: Rubin's taxonomy[16,17]

- Random Variables:
  - $X^\star \in \mathbb{R}^d$: complete unavailable data, $X \in \mathbb{R}^d$: observed data with NA
  - $M \in \{0,1\}^d$: missing pattern, or mask, $M_j = 1$ if and only if $X_j$ is missing

- Realizations: For a pattern $m$, $o(x,m) = (x_j)_{j \in \{1,\dots,d\}:m_j=0}$ the observed elements of $x$ and while $o^c(x,m) = (x_j)_{j \in \{1,\dots,d\}:m_j=1,}$ the missing elements.

$$x^\star = (1,2,3,8,5)$$
$$x = (1, \mathrm{NA}, -3, 8, \mathrm{NA})$$
$$m = (0,1,0,0,1)$$
$$o(x,m) = (1,3,8), \qquad o^c(x^\star, m) = (2,5)$$

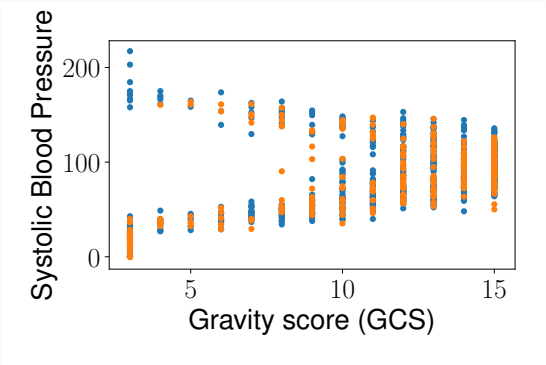
[16] Rubin. Inference and missing data. *Biometrika.* 1976.
[17] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.

# Missing values mechanism: Rubin's taxonomy[16,17]

- Random Variables:
  - $X^\star \in \mathbb{R}^d$: complete unavailable data, $X \in \mathbb{R}^d$: observed data with NA
  - $M \in \{0,1\}^d$: missing pattern, or mask, $M_j = 1$ if and only if $X_j$ is missing

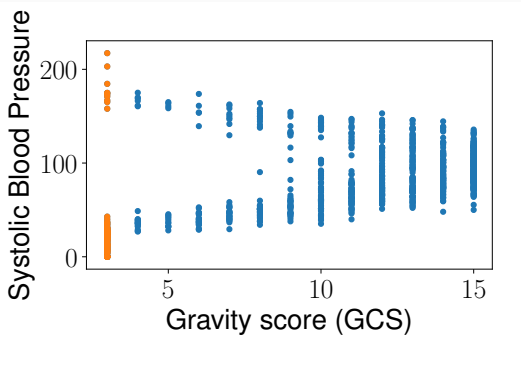
For a pattern $m$, $o(x,m) = (x_j)_{j \in \{1,\dots,d\}:m_j=0}$ the observed elements of $x$ and while $o^c(x,m) = (x_j)_{j \in \{1,\dots,d\}:m_j=1}$, the missing elements.

Ex: Simulated missing values according to the 3 mechanisms (Orange points will be missing) in Systolic Blood Pressure - GCS is always observed



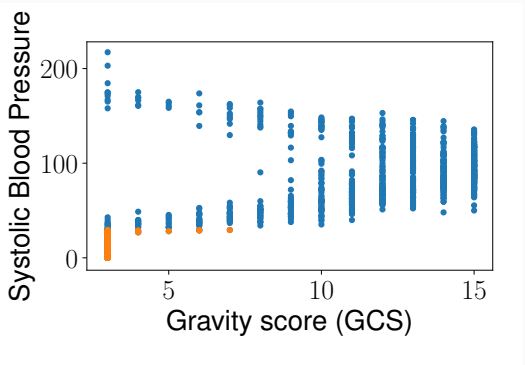

**Missing Completely at Random (MCAR)**

$m \in \mathcal{M}, x \in \mathcal{X}$,
$\mathbb{P}(M = m|x) = \mathbb{P}(M = m)$




**Missing at Random (MAR)**

$\forall m \in \mathcal{M}, x \in \mathcal{X}$
$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x,m))$




**Missing Not At Random (MNAR)**

If not MAR: it is MNAR

[16] Rubin. Inference and missing data. *Biometrika.* 1976.

[17] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.

# Two views to model the joint distribution of $(X, M)$

**Selection Model**[18]**:** $p^*(M = m, x) = \mathbb{P}(M = m \mid x)p^*(x)$

**Definition (SM-MAR)**

$$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x, m)) \text{ for all } m \in \mathcal{M}, x \in \mathcal{X}.$$

The proba. of any $m$ occurring only depends on the obs part of $x$.

**Pattern Mixture Model**[19]**:** $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

**Definition (PMM-MAR)**

$$p^*(o^c(x, m) \mid o(x, m), M = m) = p^*(o^c(x, m) \mid o(x, m)).$$

for all $m \in \mathcal{M}, x \in \mathcal{X}$. The conditional distrib. of missing given obs. in each pattern is equal to the unconditional one.[20]

[18] Heckman. Sample selection bias as a specification error. *Econometrica*. 1979
[19] Little. Pattern-mixture models for multivariate incomplete data. *JASA*. 1993
[20] Molenberghs et al. Every MNAR model has a MAR counterpart with equal fit. *JRSSB*. 2008

# Two views to model the joint distribution of $(X, M)$

**Selection Model**[18]**:** $p^*(M = m, x) = \mathbb{P}(M = m \mid x)p^*(x)$

**Definition (SM-MAR)**

$$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x, m)) \text{ for all } m \in \mathcal{M}, x \in \mathcal{X}.$$

The proba. of any $m$ occurring only depends on the obs part of $x$.

**Pattern Mixture Model**[19]**:** $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

**Definition (PMM-MAR)**

$$p^*(o^c(x, m) \mid o(x, m), M = m) = p^*(o^c(x, m) \mid o(x, m)).$$

for all $m \in \mathcal{M}, x \in \mathcal{X}$. The conditional distrib. of missing given obs. in each pattern is equal to the unconditional one.[20]

**Proposition (SM-MAR is equivalent to PMM-MAR)**

[18] Heckman. Sample selection bias as a specification error. *Econometrica*. 1979

[19] Little. Pattern-mixture models for multivariate incomplete data. *JASA*. 1993

[20] Molenberghs et al. Every MNAR model has a MAR counterpart with equal fit. *JRSSB*. 2008

# MAR with shift in marginal distribution between patterns

- Gaussian PMM: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$ and **a shift**:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ NA & x_{2,2} \end{pmatrix}, \mathbf{M} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}.$$

# MAR with shift in marginal distribution between patterns

- Gaussian PMM: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$ and **a shift**:

$$(X_1, X_2) \mid M = m_1 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right) (X_1, X_2) \mid M = m_2 \sim N\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right).$$

# MAR with shift in marginal distribution between patterns

• Gaussian PMM: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$ and **a shift**:

$$(X_1, X_2) \mid M = m_1 \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right) (X_1, X_2) \mid M = m_2 \sim N\left(\begin{pmatrix}5\\5\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right).$$

• Not identifiable without constraints. PMM-MAR: the conditional distrib. of $X_1 \mid X_2$ in each pattern is equal to the unconditional one

$$\underbrace{p^*(x_1 \mid x_2, M = m_1)}_{p^*(o^c(x,m_2)|o(x,m_2),M=m_1)} = \underbrace{p^*(x_1 \mid x_2, M = m_2)}_{p^*(o^c(x,m_2)|o(x,m_2),M=m_2)} = N(x_2, 1)(x_1) = p^*(x_1 \mid x_2).$$

# MAR with shift in marginal distribution between patterns

• Gaussian PMM: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$ and **a shift**:

$$(X_1, X_2) \mid M = m_1 \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right) (X_1, X_2) \mid M = m_2 \sim N\left(\begin{pmatrix}5\\5\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right).$$

• Not identifiable without constraints. PMM-MAR: the conditional distrib. of $X_1 \mid X_2$ in each pattern is equal to the unconditional one

$$\underbrace{p^*(x_1 \mid x_2, M = m_1)}_{p^*(o^c(x,m_2)\mid o(x,m_2), M=m_1)} = \underbrace{p^*(x_1 \mid x_2, M = m_2)}_{p^*(o^c(x,m_2)\mid o(x,m_2), M=m_2)} = N(x_2, 1)(x_1) = p^*(x_1 \mid x_2).$$

**Definition (Conditional indep. MAR - CIMAR)**

$$p^*(o^c(x, m) \mid o(x, m), M = m') = p^*(o^c(x, m) \mid o(x, m), M = m'')$$

for all $m, m', m'' \in \mathcal{M}, x$.equivalent to $o^c(X, M) \mid o(X, M) \perp\!\!\!\perp M$

# MAR with shifts in conditional distribution between patterns

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}, \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}$$

## CIMAR

$p^*(x_1, x_2 \mid x_3, M = m_1) = p^*(x_1, x_2 \mid x_3, M = m_2) = p^*(x_1, x_2 \mid x_3, M = m_3)$

Distrib. of $X_1, X_2 \mid X_3$ is not allowed to change from one pattern to another, though the marginal distrib. of $X_3$ can change. CIMAR allows to learn the conditional distrib. from any pattern.

## PMM-MAR

$p^*(x_1, x_2 \mid x_3, M = m_3) = p^*(x_1, x_2 \mid x_3)$

**Both distrib. of observed variables and conditional ones can change from pattern to pattern.**

## MCAR

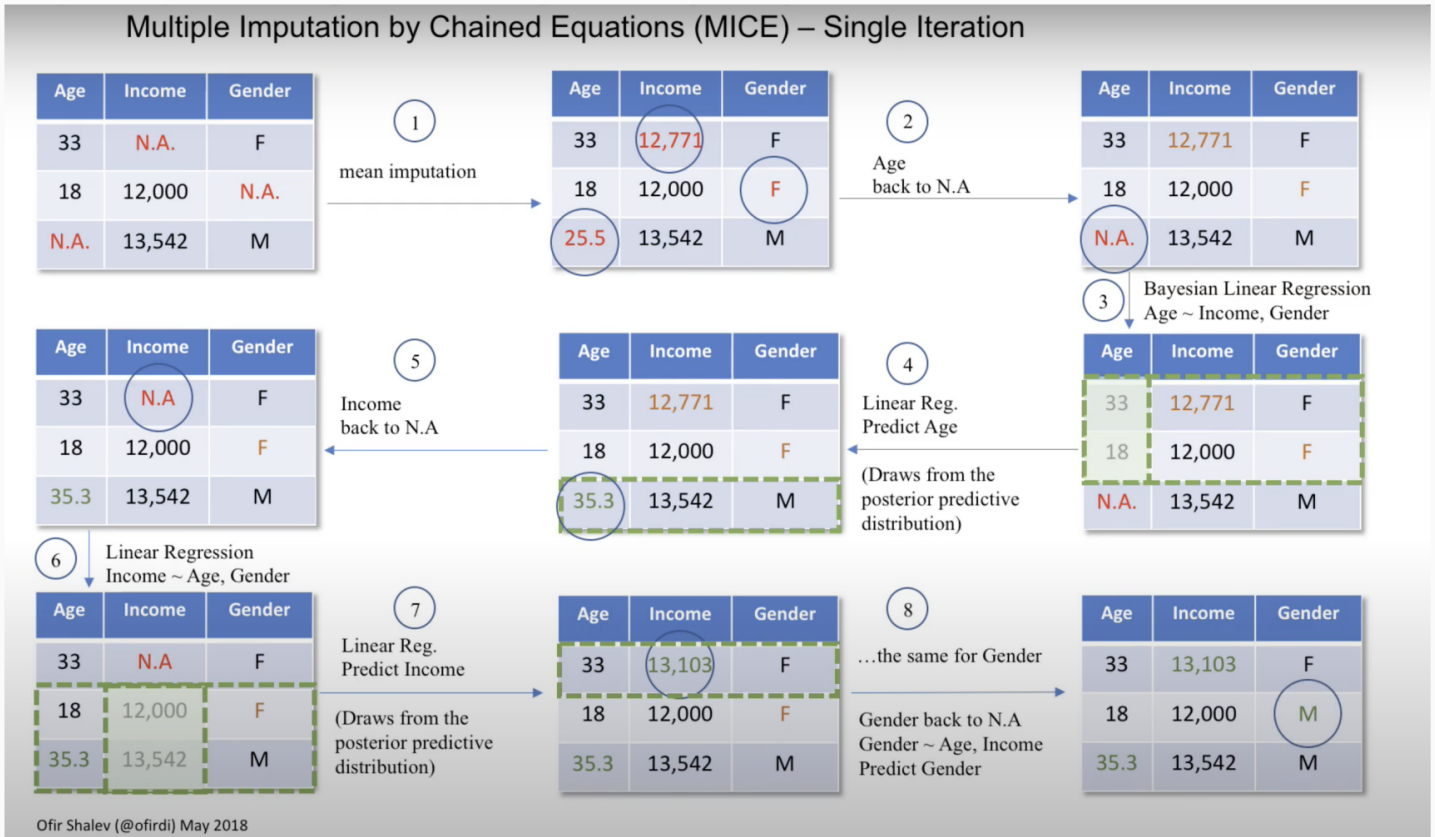
No change is allowed.

# Fully conditional specification - FCS, (M)ICE

1. Fill NA with plausible values to get an initial completed dataset

2. For $j \in \{1, \ldots, d\}$, $t \geq 1$ use a univariate imputation to sample new imputed values $x_j^{(t+1)} \sim p^*(x_j \mid x_{-j}^{(t)})$, where $x_{-j}^{(t)} = \{x_l^{(t)}\}_{l \neq j}$ the imputed & observed values of other variables except $j$ at the $t$th iteration.

3. Iterate until convergence



# Fully conditional specification under MAR

- Assume $x_{-j}$ is well imputed: we have $p^*(x_{-j})$
- Impute by drawing from the conditional distrib. of $X_j \mid X_{-j}$ **learned from all patterns in which $x_j$ is observed**:

$$h^*(x_j \mid x_{-j}) = \sum_{m \in L_j} \frac{\mathbb{P}(M = m)}{\sum_{m \in L_j} p^*(x_{-j} \mid M = m)\mathbb{P}(M = m)} p^*(x \mid M = m),$$

with $L_j = \{m \in \mathcal{M} : x_j \in o(x, m)\}$ the patterns where $x_j$ is observed

**Theorem: Identifiability (Näf et al., 2024)**

Assume PMM-MAR holds,

$$h^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j}), \text{ for all } x_{-j} \text{ with } p^*(x_{-j}) > 0$$

$\Rightarrow$ In a population setting (perfect estimation), **FCS identifies the right distributions** to impute missing values under MAR.

Remark: Different from learning the conditional distributions from the fully observed data and then impute the missing variables.

# What is a good imputation method?

- ▷ both conditional and marginal **distribution shifts** can occur for different patterns under MAR.
- ▷ conditional shifts are handled with FCS

**An ideal imputation method should**

- ▷ (1) be a distributional regression method,
- ▷ (2) be able to capture nonlinearities in the data,
- ▷ (3) be able to deal with distributional shifts in the observed variables,
- ▷ (4) be fast to fit,
- ▷ (5) the method is able to deal with multivariate responses.

1-3 are crucial for imputation under MAR

4-5 are only relevant to reduce the computational burden.

# What is a good imputation method?

(1) be a distributional regression method,
(2) be able to capture nonlinearities in the data,
(3) be able to deal with distributional shifts in the observed variables,
(4) be fast to fit,
(5) the method is able to deal with multivariate responses.

| Method | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| missForest (Stekhoven & Bühlmann, 2011) | | ✓ | | ✓ | |
| mice-cart (Burgette & Reiter, 2010) | ✓ | ✓ | | ✓ | |
| mice-RF (Doove et al., 2014) | ✓ | ✓ | | ✓ | |
| mice-DRF (Näf et al., 2024) | ✓ | ✓ | | ✓ | ✓ |
| mice-norm.nob (Gaussian) | ✓ | | ✓ | ✓ | ✓ |
| mice-norm.predict (Regresssion) | | | ✓ | ✓ | ✓ |

[21] Cevid et al., Distributional Random Forests. *JMLR*. 2022

# What is a good imputation method?

(1) be a distributional regression method,
(2) be able to capture nonlinearities in the data,
(3) be able to deal with distributional shifts in the observed variables,
(4) be fast to fit,
(5) the method is able to deal with multivariate responses.

| Method | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| missForest (Stekhoven & Bühlmann, 2011) | | ✓ | | ✓ | |
| mice-cart (Burgette & Reiter, 2010) | ✓ | ✓ | | ✓ | |
| mice-RF (Doove et al., 2014) | ✓ | ✓ | | ✓ | |
| mice-DRF (Näf et al., 2024) | ✓ | ✓ | | ✓ | ✓ |
| mice-norm.nob (Gaussian) | ✓ | | ✓ | ✓ | ✓ |
| mice-norm.predict (Regresssion) | | | ✓ | ✓ | ✓ |

- ▷ mice-cart/RF estimate a tree, a forest, on observed data and then draw imputations from the leaves (approx conditional distribution) whereas distributional forest[21] is a distributional method

[21] Cevid et al., Distributional Random Forests. *JMLR*. 2022

# Forests generalize poorly outside of the training set

Ex: Variables income & age with MAR missing values in income

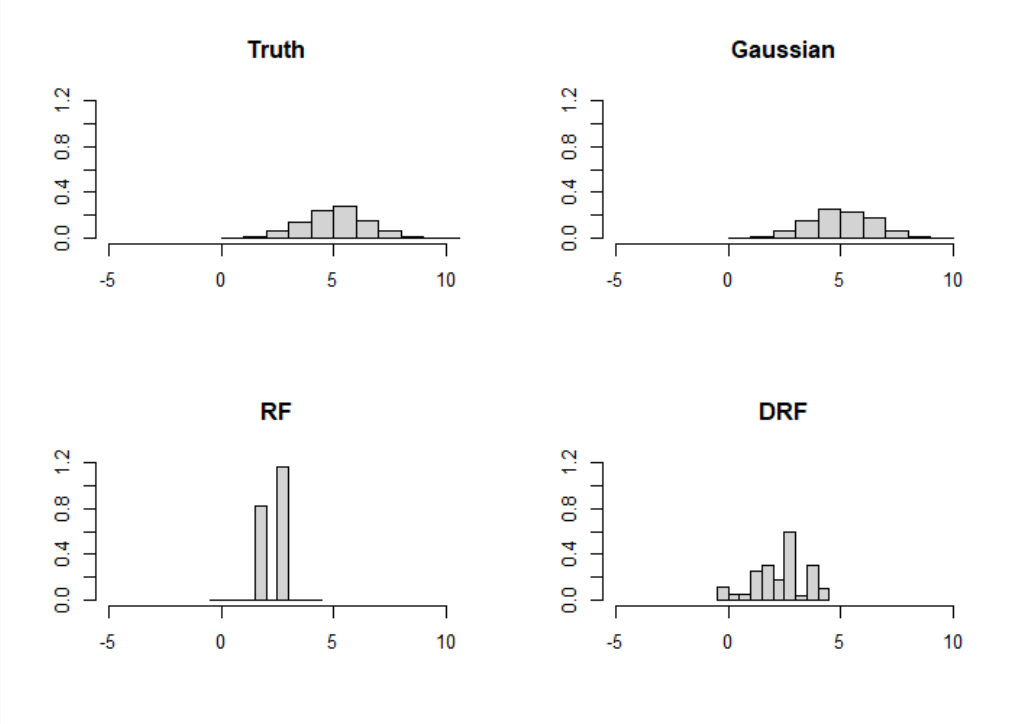
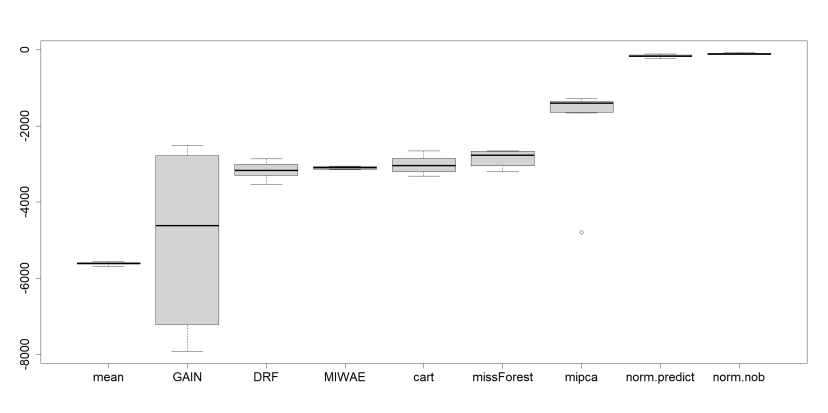



**Figure 2:** True distribution against a draw from different imputation methods.

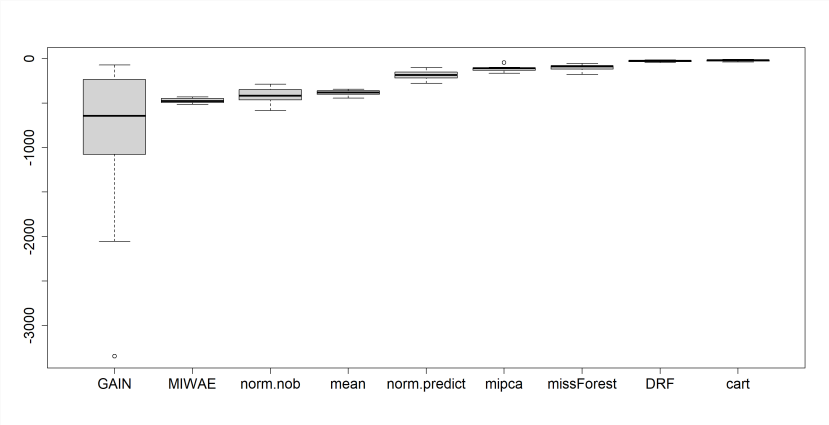
DRF, a distributional method $>$ mice-RF but fails to deal with the covariate shift (centering $\approx 2$ instead of 5).

Finding an imputation method that meets (1) - (5) is still an open problem!

# Empirical study: ranking with energy scores and not RMSE




Gaussian relation with shifts




Non linear relation with shifts

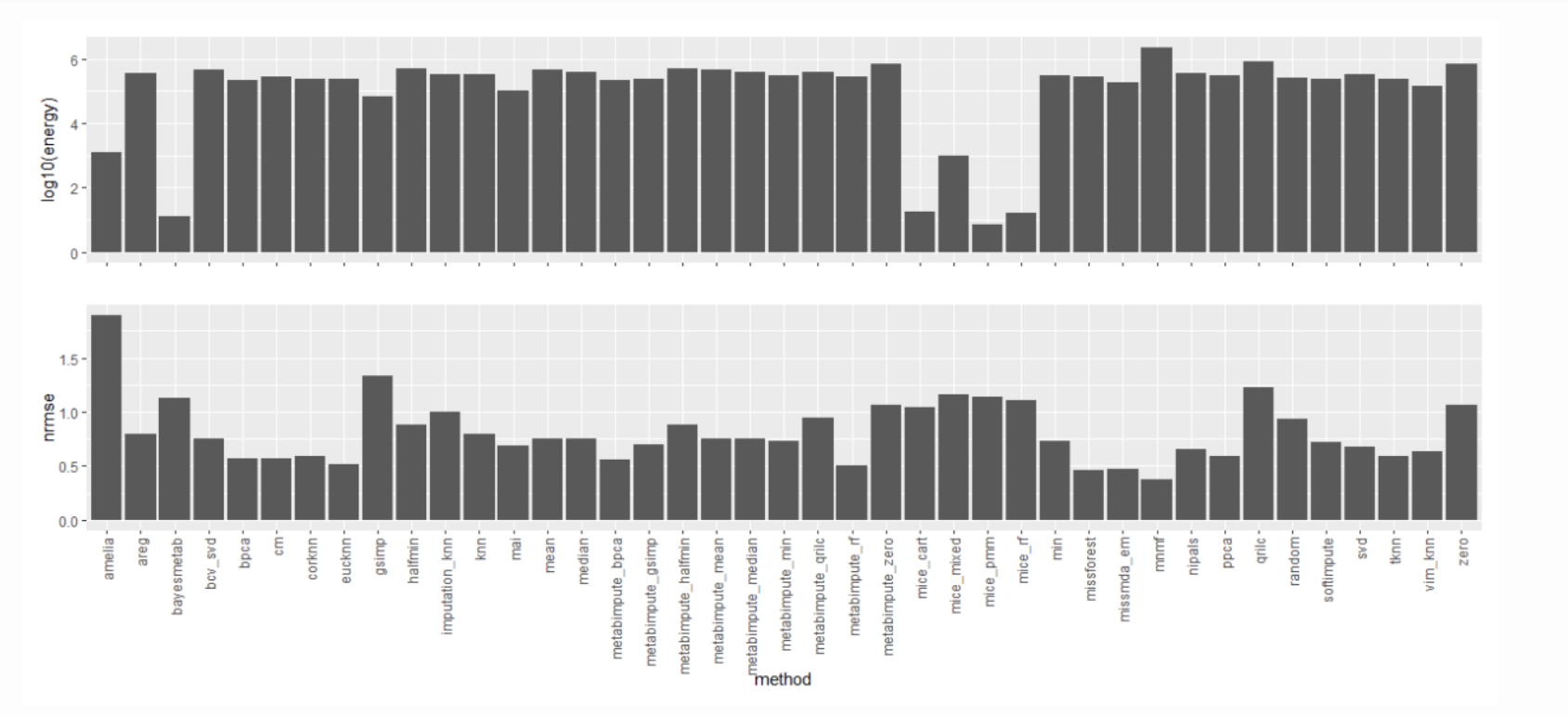
Ex with $d = 6$, $n = 1500$, 20% NA and CIMAR, $X_{O^c} = \mathbf{B}f(X_O) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$,

## Energy distance between imputed & real data

$$d(H, P^*) = 2\mathbb{E}[\|X - Y\|_{\mathbb{R}^d}] - \mathbb{E}[\|X - X'\|_{\mathbb{R}^d}] - \mathbb{E}[\|Y - Y'\|_{\mathbb{R}^d}],$$

where $\|\cdot\|_{\mathbb{R}^d}$ is the Euclidean metric on $\mathbb{R}^d$, $X \sim H$, $Y \sim P^*$ and $X', Y'$ are independent copies of $X$ and $Y$.

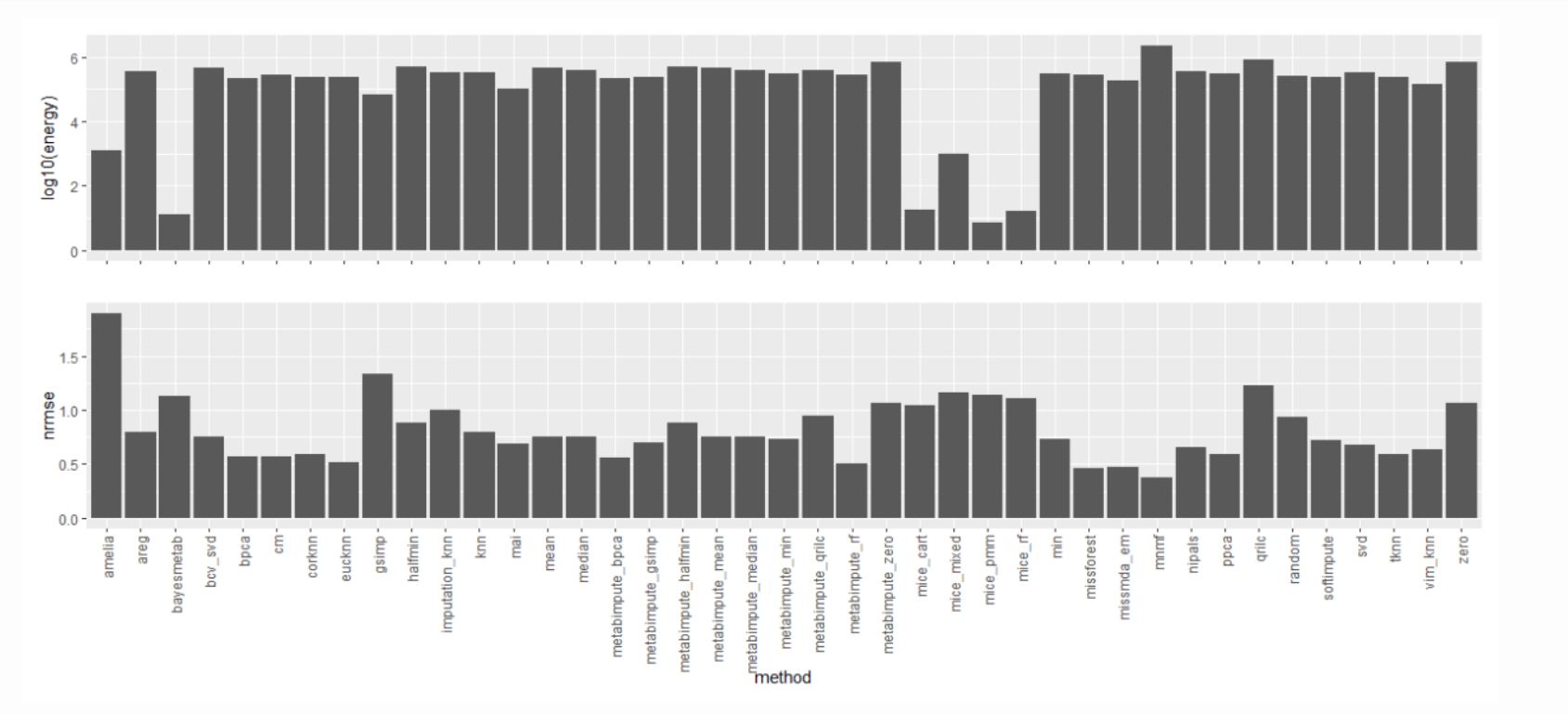
# Empirical study: ranking with energy scores and not RMSE




credit: Krystyna Grzesiak, Michal Burdukiewicz[22] 230 scenarios (10 missing values patterns 23 different-sized datasets)

[22]imputomics: web server and R package for missing values imputation in metabolomics data. *Bioinformatics 2024*.

# Empirical study: ranking with energy scores and not RMSE




credit: Krystyna Grzesiak, Michal Burdukiewicz[22] 230 scenarios (10 missing values patterns 23 different-sized datasets)

[22] imputomics: web server and R package for missing values imputation in metabolomics data. *Bioinformatics 2024*.

# Conclusion

- Non-parametric PMM view of missing (different environments) helps understand non-parametric imputation under MAR
- Identification result for FCS: the right conditional distributions are identifiable under MAR with no parametric assumption
- Identification under the weakest MAR assumption.[23] Link between all MAR (MAR is broad): CIMAR, Extended MAR (EMAR), PMM-MAR

[23] Deng et al., (2022) and Fang (2023) showed identifiability for GAN imputation under CIMAR

# Conclusion

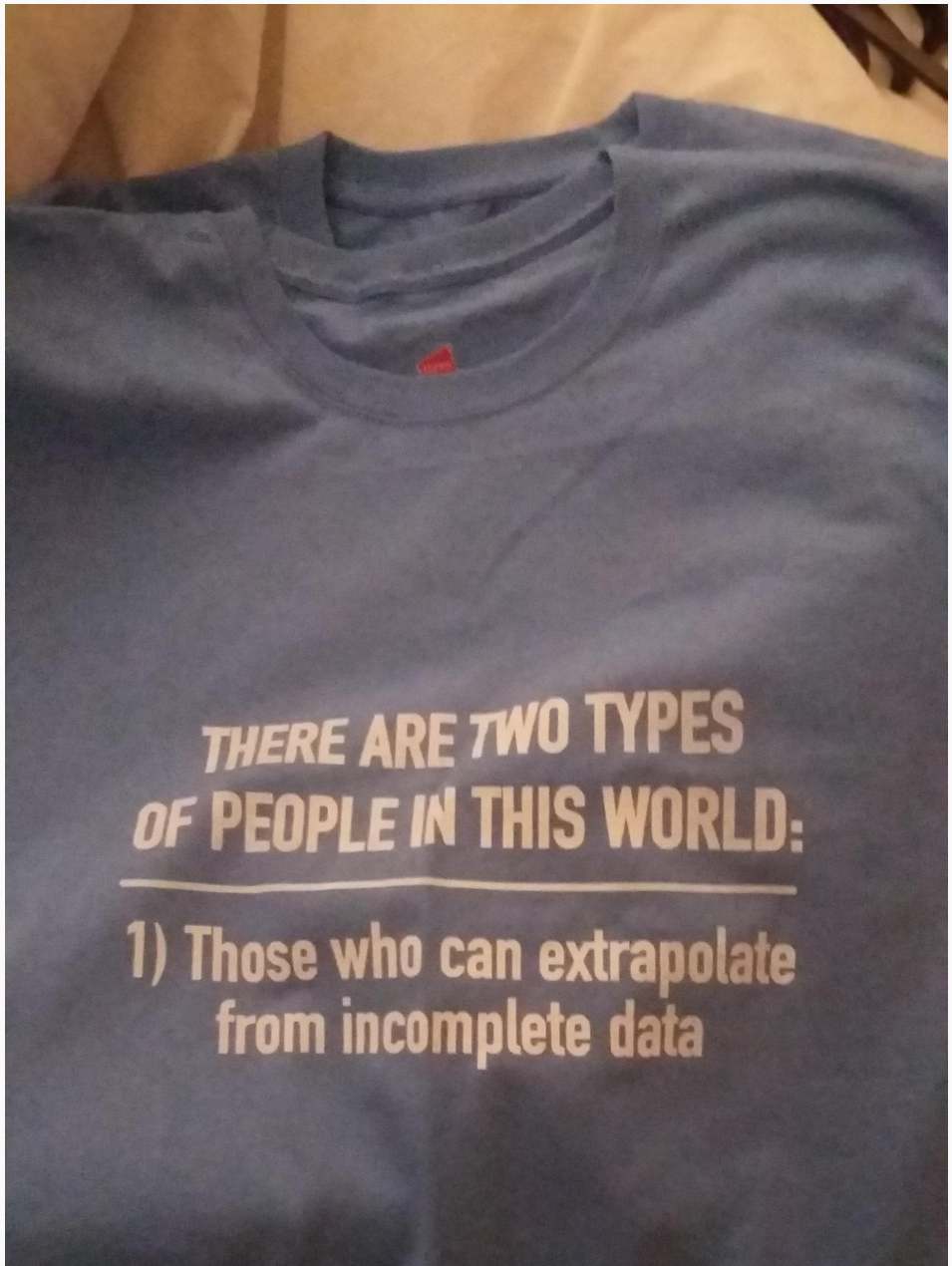
- ▷ Non-parametric PMM view of missing (different environments) helps understand non-parametric imputation under MAR
- ▷ Identification result for FCS: the right conditional distributions are identifiable under MAR with no parametric assumption
- ▷ Identification under the weakest MAR assumption.[23] Link between all MAR (MAR is broad): CIMAR, Extended MAR (EMAR), PMM-MAR

- ▷ 5 points the ideal sequential imputation method should meet
- ▷ The quest for an imputation method meeting all 5 points is still open
- ▷ mice-DRF promising (code available)
- ▷ Imputation scores with missing values that are proper under MAR: ranking imputation methods

Impact for causal inference

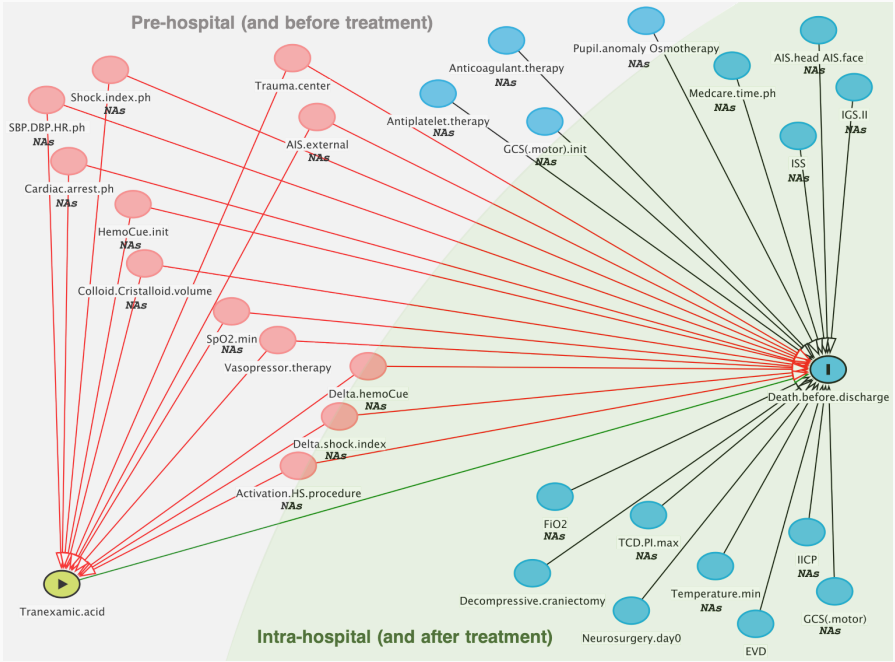
[23] Deng et al., (2022) and Fang (2023) showed identifiability for GAN imputation under CIMAR

# Thank you


THERE ARE TWO TYPES
OF PEOPLE IN THIS WORLD:
1) Those who can extrapolate
from incomplete data

# Causal identifiability assumptions adapted to missing values

http://www.dagitty.net/




| Covariates | | | Treatment | Outcome(s) | |
|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | W | Y(0) | Y(1) |
| NA | 20 | F | 1 | ? | 200 |
| -6 | 45 | NA | 0 | 10 | ? |
| 0 | NA | M | 1 | ? | 150 |
| NA | 32 | F | 1 | ? | 100 |
| 1 | 63 | M | 1 | 15 | ? |
| -2 | NA | M | 0 | 20 | ? |

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

$\Rightarrow$ Doctors give us the DAG (do not ask for the complete graph only for a sufficient adjustment set), obtained by a Delphi method

**Unconfoundedness with missing values**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X$

$X = (1 - M) \odot X^\star + M \odot NA$; with $M_{ij} = 1$ 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

# 2. Augmented IPW under unconfoundeness with missing values

## Augmented IPW[24] with missing values

$$\hat{\tau} = \frac{1}{n}\sum_i \left( \widehat{\mu_{(1)}}(X_i) - \widehat{\mu_{(0)}}(X_i) + W_i \frac{Y_i - \widehat{\mu_{(1)}}(X_i)}{\widehat{e}(X_i)} - (1 - W_i)\frac{Y_i - \widehat{\mu_{(0)}}(X_i)}{1 - \widehat{e}(X_i)} \right)$$

## Generalized propensity score[25]

$$e(x) = \mathbb{P}(W = 1 \,|\, X = x)$$

One model per pattern: $\sum_{m \in \{0,1\}^d} \mathbb{E}\left[W | X_{obs(m)}, M = m\right] \; \mathbb{1}_{M=m}$

$\Rightarrow$ Supervised learning with missing values[26,27]

- Learning with a universally consistent learner on (Mean) imputed data is Bayes consistent for **all missing data mechanism**
- Missing incorporate in attributes (MIA) for tree methods (grf package)

[24] Mayer, Wager, J. Doubly robust treat. effect estim. with incomplete confounders *AOAS*. 2020.
[25] Rosenbaum & Rubin. Reducing bias in observational studies *JASA*. (1984).
[26] **J.** et al. Consistency of supervised learning with missing values. *Stats Papers*. 2028-24.
[27] Le Morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips 2021*.

# 2. Augmented IPW under unconfoundeness with missing values

## Augmented IPW[24] with missing values

$$\hat{\tau} = \frac{1}{n}\sum_i \left( \widehat{\mu_{(1)}}(X_i) - \widehat{\mu_{(0)}}(X_i) + W_i \frac{Y_i - \widehat{\mu_{(1)}}(X_i)}{\widehat{e}(X_i)} - (1 - W_i)\frac{Y_i - \widehat{\mu_{(0)}}(X_i)}{1 - \widehat{e}(X_i)} \right)$$

## Generalized propensity score

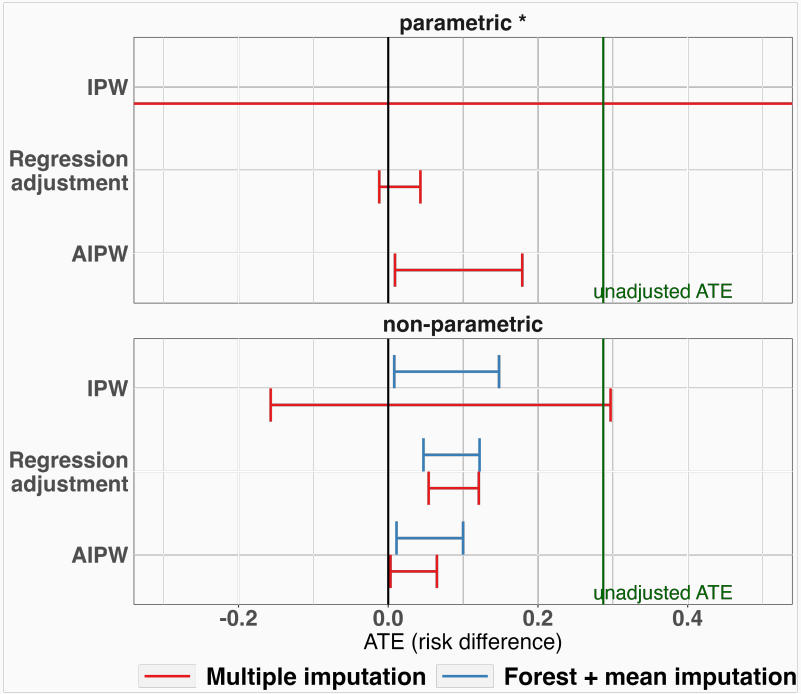
$$e(x) = \mathbb{P}(W = 1 \,|\, X = x)$$

One model per pattern: $\sum_{m \in \{0,1\}^d} \mathbb{E}\left[W | X_{obs(m)}, M = m\right] \; \mathbb{1}_{M=m}$

[24] Mayer, Wager, J. Doubly robust treat. effect estim. with incomplete confounders *AOAS*. 2020.

# ATE estimations: effect of tranexamic acid on in-ICU mortality

- 40 covariates, 18 confounders (categorical and quantitative). 8248 patients
- Multiple imputation assumes **MAR & classical unconfoundeness** while other **unconfoundeness with missing & (no) assumptions on missing mechanism**

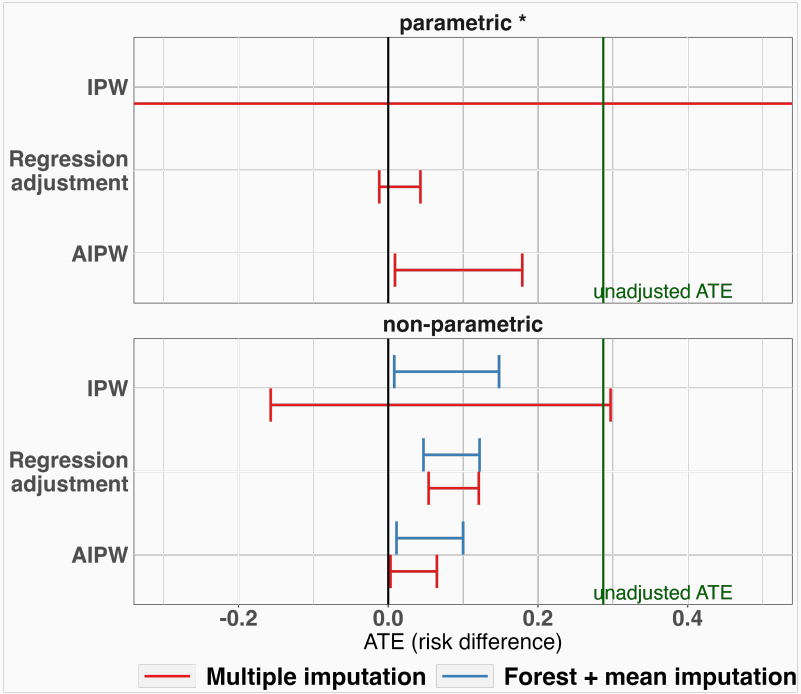



$x$-axis: Estimat. of the ATE ($\times 100$), bootstrap CI, $y$-axis: Methods with logistic regression or forests for nuisances. Missing values handled with multiple imputation or MIA[25]

[25] Other estimators (latent confounding, Kallus 2018 or parametric models with EM algorithms Jiang, J. 2019) are available bur not displayed for clarity (all tend to a slightly detrimental effect)

# ATE estimations: effect of tranexamic acid on in-ICU mortality

- 40 covariates, 18 confounders (categorical and quantitative). 8248 patients
- Multiple imputation assumes **MAR & classical unconfoundeness** while other **unconfoundeness with missing & (no) assumptions on missing mechanism**




$x$-axis: Estimat. of the ATE ($\times 100$), bootstrap CI, $y$-axis: Methods with logistic regression or forests for nuisances. Missing values handled with multiple imputation or MIA[25]

[25] Other estimators (latent confounding, Kallus 2018 or parametric models with EM algorithms Jiang, J. 2019) are available bur not displayed for clarity (all tend to a slightly detrimental effect)

# Imputing with a mixture of patterns

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & NA & x_{2,3} \\ NA & x_{3,2} & x_{3,3} \end{pmatrix}, \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}.$$

whereby $(X_1, X_2, X_3)$ are independently uniformly distributed on $[0, 1]$.

$$\begin{aligned} \mathbb{P}(M = m_1 \mid x) &= \mathbb{P}(M = m_1 \mid x_1) = x_1/3 \\ \mathbb{P}(M = m_2 \mid x) &= \mathbb{P}(M = m_2 \mid x_1) = 2/3 - x_1/3 \\ \mathbb{P}(M = m_3 \mid x) &= \mathbb{P}(M = m_3) = 1/3. \end{aligned}$$

# Imputing with a mixture of patterns

We want to impute $X_1$ in the third pattern (with $X_2$ and $X_3$ observed)

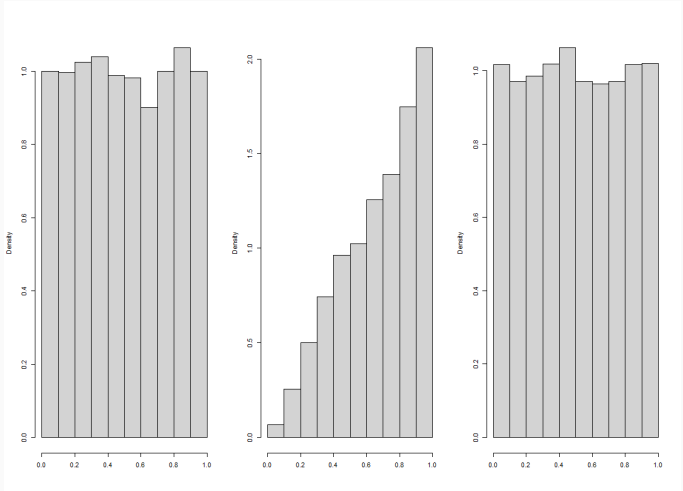



**Figure 3:** Distrib. of $X_1$ in different patterns. Left: Distrib. of $X_1 \mid M = m_3$. Middle: $(X_1 \mid M = m_1)$. Right: Distribution of all patterns for which $X_1$ is observed (Mixture of the distribution of $X_1$ in pattern 1 and 2).

• As the distrib. of $(X_2, X_3)$ in each patterns is the same, this shows the change of $X_1 \mid X_2, X_3$ from $m_3$ to $m_1$: PMM-MAR allows change in the conditional distrib. over patterns.

• Note that the distrib. $X_1 \mid X_2, X_3$ in $m_3$ corresponds to the mixture of distribution of $X_1 \mid X_2, X_3$ in the patterns where $X_1$ is observed.