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"

December 16, 2024

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

Jeffrey Näf
(Postdoc Inria)

Erwan Scornet
(Prof. Sorbonne Univ.)

Traumabase: an observational French registry on trauma[2]

- 40000 patients
- 250 continuous and categorical variables
- 40 trauma centers, 4000 new patients/ year

Center	Accident	Age	Sex	Lactate	Blood Pres.	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292 000	
Pitie	gun	26	m	NA	131	no	323 000	
Beaujon	moto	63	m	3.9	NR	yes	318 000	
Pitie	moto	30	w	Imp	107	no	211 000	
⋮								⋱

[1] Zaffran, **J.**, Dieuleveut, Romano. Conformal Prediction with Missing Values. *ICML 2023.*

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

Traumabase: an observational French registry on trauma[2]

- 40000 patients
- 250 continuous and categorical variables
- 40 trauma centers, 4000 new patients/ year

Center	Accident	Age	Sex	Lactate	Blood Pres.	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292 000	
Pitie	gun	26	m	NA	131	no	323 000	
Beaujon	moto	63	m	3.9	NR	yes	318 000	
Pitie	moto	30	w	Imp	107	no	211 000	
⋮								⋱

⇒ **Explain and Predict** hemorrhagic shock, need for neurosurgery and need for a trauma center given pre-hospital features.
Ex: logistic regression/ random forests + **Quantify uncertainty**[1]

Clinical trial will be launched end 2024: real-time implementation of models in the ambulance via a mobile data collection application

[1] Zaffran, **J.**, Dieuleveut, Romano. Conformal Prediction with Missing Values. *ICML 2023.*
[2] `www.traumabase.eu - https://www.traumatrix.fr/`

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[3] (> 150 packages)

<u>Inferential aim</u>: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$ to get confidence intervals with the appropriate coverage

[3] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[4] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`
[5] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[6] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[3] (> 150 packages)

<u>Inferential aim</u>: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$ to get confidence intervals with the appropriate coverage

Modify the estimation process to deal with missing values

Maximum likelihood inference: Expectation Maximization algorithms[4]

[3] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[4] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`
[5] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[6] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[3] (> 150 packages)

Inferential aim: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$ to get confidence intervals with the appropriate coverage

Modify the estimation process to deal with missing values

Maximum likelihood inference: Expectation Maximization algorithms[4]

(Multiple) imputation to get a complete data set. Ex: (M)ICE

[3] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[4] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`
[5] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[6] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[3] (> 150 packages)

Inferential aim: **Estimate parameters & their variance, i.e. $\hat{\beta}$, $\hat{V}(\hat{\beta})$** to get confidence intervals with the appropriate coverage

Modify the estimation process to deal with missing values

Maximum likelihood inference: Expectation Maximization algorithms[4]

(Multiple) imputation to get a complete data set. Ex: (M)ICE

Matrix completion aim: **Predict the missing values** as well as possible.
Solutions: using low rank matrix approximation

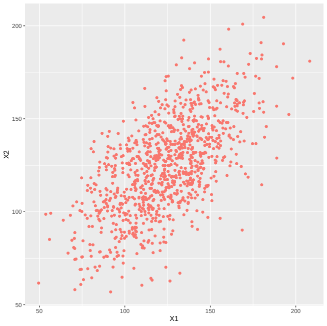
Predictive aim: **Predict an outcome** with missing values in covariates.[5,6]
Solutions: using deterministic (e.g. constant) imputation or Missing Incorporated in Attributes for trees based methods (`grf package`)

[3] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[4] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`
[5] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[6] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

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



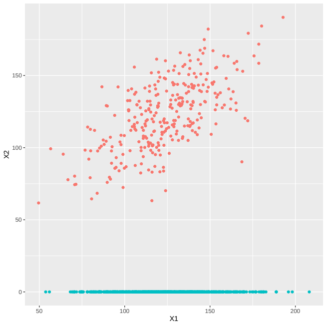
$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

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

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



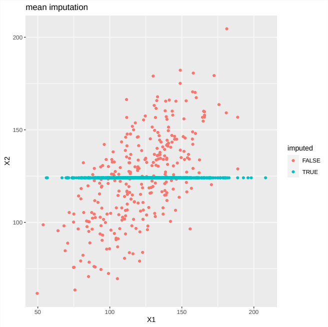
$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

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

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**



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

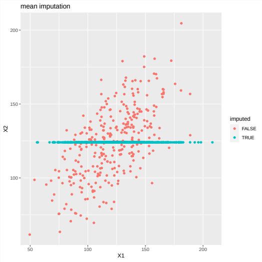
$\hat{\mu}_{x_2} = 0.01$
$\hat{\sigma}_{x_2} = 0.5$
$\hat{\rho} = 0.30$

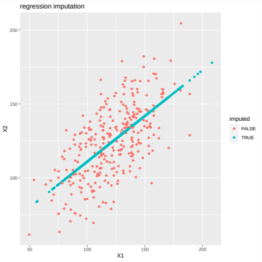
Mean imputation deforms joint and marginal distributions

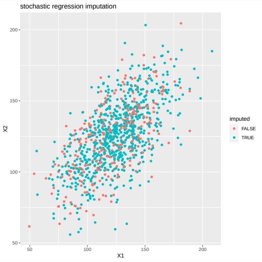
Objective: to impute while preserving distribution

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 β (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 β and σ - 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







	mean imputation	regression imputation	stochastic regression imputation
$\mu_{x_2} = 0$	0.01	0.01	0.01
$\sigma_{x_2} = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

Impute while preserving distribution. Multivariate case

Assuming a joint distribution

- Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
- Low rank : $X_{n\times d} = \mu_{n\times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank
  - $\Rightarrow$ Different regularization depending on noise regime[7]
  - $\Rightarrow$ Count data,[8] ordinal data, categorical data, blocks/multilevel data
- Optimal transport,[9] deep generative models: GAIN,[10] MIWAE,[11] etc.[12][13]

[7] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*. 2016.
[8] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[9] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[10] Yoon et al. GAIN: Missing data imputation using generative adversarial nets. *ICML*. 2018.
[11] Mattei & Frellsen. Miwae: Deep generative model. & imput. of incomplete data. *ICML*. 2018.
[12] Deng et al. Extended missing data imput. via gans. *Data Mining & Knowledge Discovery*. 2022.
[13] Fang Bao. Fragmgan gan for fragmentary data imputation. *Stat.theory & Related Fields*. 2023.
[14] van Buuren, S. Flexible Imputation of Missing Data. Chapman Hall/CRC Press. 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 μ of low rank
  - $\Rightarrow$ Different regularization depending on noise regime[7]
  - $\Rightarrow$ Count data,[8] ordinal data, categorical data, blocks/multilevel data
- Optimal transport,[9] deep generative models: GAIN,[10] MIWAE,[11] etc.[12][13]

Iterating conditional models (joint distribution implicitly defined)

- with parametric regression (M)ICE: (Multiple) Imput. by Chained Equations[14]
- iterative imputation of each variable by random forests[15]

---

[7] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*. 2016.
[8] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[9] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[10] Yoon et al. GAIN: Missing data imputation using generative adversarial nets. *ICML*. 2018.
[11] Mattei & Frellsen. Miwae: Deep generative model. & imput. of incomplete data. *ICML*. 2018.
[12] Deng et al. Extended missing data imput. via gans. *Data Mining & Knowledge Discovery*. 2022.
[13] Fang Bao. Fragmgan gan for fragmentary data imputation. *Stat.theory & Related Fields*. 2023.
[14] van Buuren, S. Flexible Imputation of Missing Data. Chapman Hall/CRC Press. 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, $o(x,m) = (x_j)_{j\in\{1,\ldots,d\}:m_j=0}$ the observed elements of x and while $o^c(x,m) = (x_j)_{j\in\{1,\ldots,d\}:m_j=1}$, the missing elements.

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

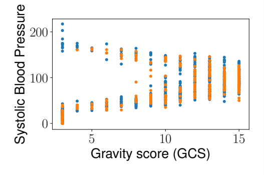
[16] Rubin. Inference and missing data. *Biometrika*. 1976.
[17] What Is Meant by "Missing at Random"? Seaman, et al. *Statistical Science*. 2013.

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.

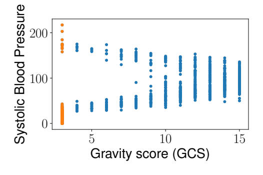
Ex: Simulated missing values according to the 3 mechanisms (Orange points will be missing) in Systolic Blood Pressure - GCS is always observed



Missing Completely at Random (MCAR)

$m \in \mathcal{M}, x \in \mathcal{X}$,

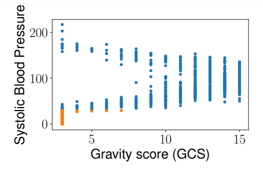
$\mathbb{P}(M = m|x) = \mathbb{P}(M = m)$



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 pattern m 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 pattern m 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

- <u>Gaussian PMM</u>: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. 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 restriction. How distributions can change?

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

MAR with shift in marginal distribution between patterns

- <u>Gaussian PMM</u>: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. 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 restriction. How distributions can change?

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

Definition (Conditional indep. MAR - CIMAR)

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

for all $m, m' \in \mathcal{M}, x \in \mathcal{X}$.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

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

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.

PMM-MAR

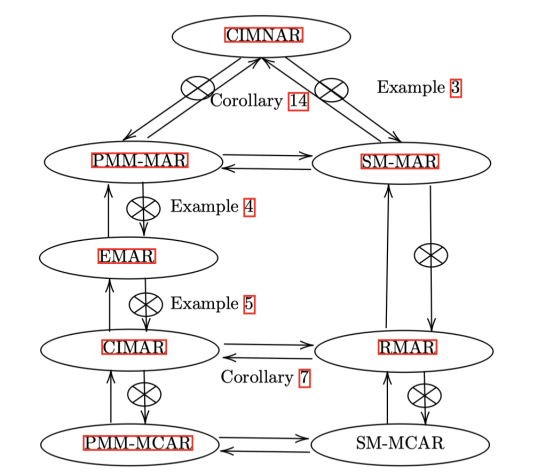
$p^*(x_1, x_2 \mid x_3, M = m_3) = p^*(x_1, x_2 \mid x_3)$

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

MCAR: No change allowed.

$m \in \mathcal{M}, m' \in \mathcal{M}, x \in \mathcal{X},\ p^*(x) = p^*(x \mid M = m) = p^*(x \mid M = m')$

Relationships between the M(N)AR conditions



(Non) Identifiability under non-parametric MAR

Definition: Imputing with a mixture of distribution

$p^*(o^c(x,m) \mid o(x,m))$ **is identifiable** from $\mathcal{M}_0 \subset \mathcal{M}$ if there exists some weights $w_{m'}(o(x,m))$ (summing to 1) such that the mixture

$$h^*(o^c(x,m) \mid o(x,m)) = \sum_{m' \in \mathcal{M}_0} w_{m'}(o(x,m)) p^*(o^c(x,m) \mid o(x,m), M = m')$$

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

Proposition: Identifiability under PMM-MAR is not trivial

Assume $|\mathcal{M}| > 3$. For any pattern $m \in \mathcal{M}$, $p^*(o^c(x,m) \mid o(x,m))$ is

- identifiable from any other pattern $m' \neq m$ under CIMAR,
- is not identifiable from any single pattern $m' \neq m$ under PMM-MAR.

If $\left|\sum_{j=1}^d m_j\right| > 1$, $p^*(o^c(x,m) \mid o(x,m))$ **is not identifiable from** L_m, **the set of patterns for which** $o^c(x,m)$ **is observed**.
$L_m = \{m' \in \mathcal{M} : m'_j = 0 \text{ for all } j \text{ such that } m_j = 1\}$.

Identifiability under MAR considering one variable at a time

- Consider the following mixture of distribution

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

with $L_j = \{m \in \mathcal{M} : m_j = 0\}$, the patterns where x_j is observed

Theorem: Identifiability of the right conditional distribution

Assume **PMM-MAR** holds,

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

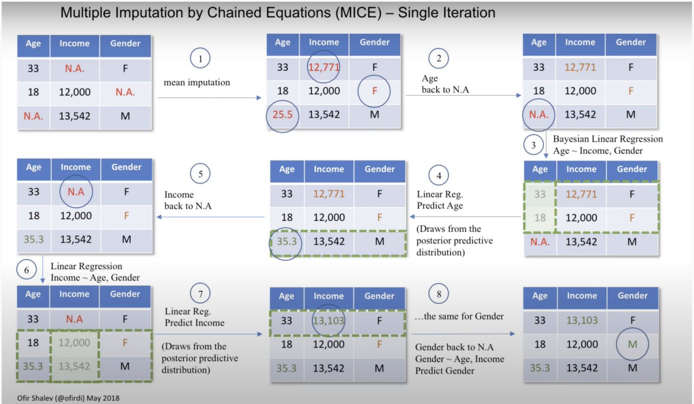
At X_j, one can reduce the $|\mathcal{M}|$ patterns to two, one where X_j is missing, and one where it is observed. Though these two aggregated patterns are mixtures of several patterns $m \in \mathcal{M}$, MAR implies that both aggregated patterns have the same conditional distribution $X_j^* \mid X_{-j}^*$

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 $x_j^{(t+1)} \sim p^t(x_j \mid x_{-j}^{(t)})$, where $x_{-j}^{(t)} = \{x_l^{(t)}\}_{l \neq j}$ the imputed & observed values of other variables except j at the tth iteration.

3. Iterate until convergence



Fully conditional specification - FCS, (M)ICE

1. Fill NA with plausible values to get an initial completed dataset
2. For $j \in \{1, \dots, d\}$, $t \geq 1$ use a univariate imputation to sample new imputed values $x_j^{(t+1)} \sim p^t(x_j \mid x_{-j}^{(t)})$, where $x_{-j}^{(t)} = \{x_l^{(t)}\}_{l \neq j}$ the imputed & observed values of other variables except j at the tth iteration.
3. Iterate until convergence

Theorem shows that if we assume to have access to the true distribution $p^*(x_{-j})$ (assume x_{-j} is well imputed), we can impute according to the true distribution $p^*(x_j \mid x_{-j})$ by drawing from the conditional distrib. of $X_j \mid X_{-j}$ **learned from all patterns in which** x_j **is observed**

FCS approach can identify the right conditional distributions under PMM MAR

What is a good imputation method under MAR?

- 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,

1-3 are crucial for imputation under MAR
4 is only relevant to reduce the computational burden.

Rk: Block-wise FCS (multi-output methods to impute variables as blocks) should not be used: do not recover the correct distribution

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,

Method	(1)	(2)	(3)
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 (Regression)			✓

[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,

Method	(1)	(2)	(3)
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 (Regression)			✓

- 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

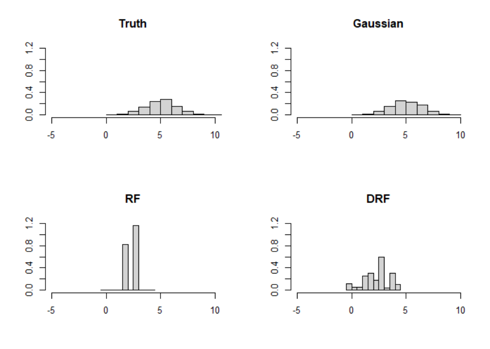
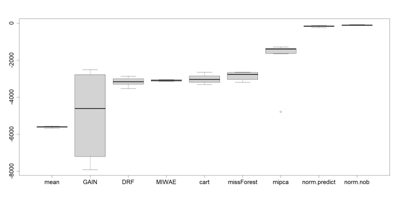


Figure 1: True distribution against a draw from different imputation methods.

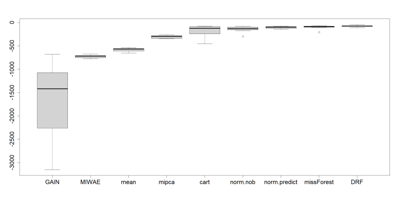
DRF, a distributional method $>$ mice-RF but fails to deal with the covariate shift (centering ≈ 2 instead of 5).

Finding an imputation method that meets (1) - (4) 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, $X_{O^c} = \mathbf{B} f(X_O) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$

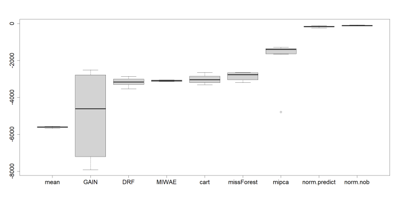
Energy distance[22] between imputed & real data

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

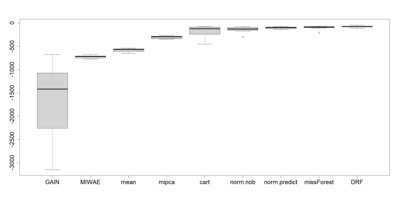
where $\|\cdot\|_{\mathbb{R}^d}$ is the Euclidean metric on $\mathbb{R}^d$, $X \sim H$, $Y \sim P^*$ and X', Y' are independent copies of X and Y.

[22] Székely & Rizzo. Energy statistics *Journal of stat. planning & inference*. 2013

Empirical study: ranking with energy scores and not RMSE



Gaussian relation with shifts



Non linear relation with shifts

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

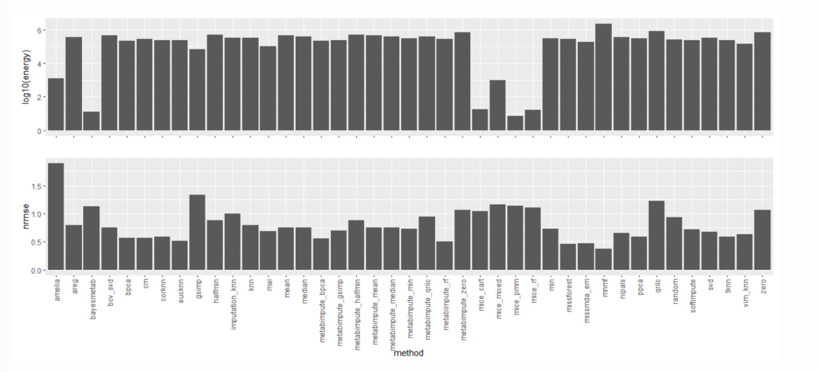
Energy distance[22] between imputed & real data

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

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.

[22] Székely & Rizzo. Energy statistics *Journal of stat. planning & inference*. 2013

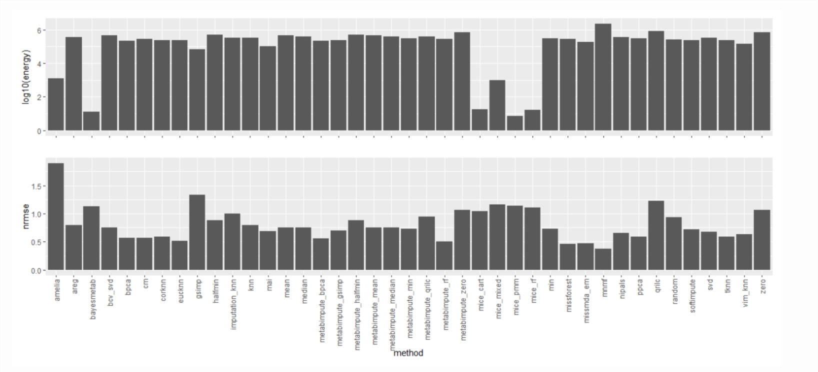
Empirical study: ranking with energy scores and not RMSE



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

[23]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[23] 230 scenarios (10 missing values patterns 23 different-size datasets)

[23] 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.[24] Beyond MAR. $\forall$j $\in \{1, \ldots, d\}$, $\forall$x $\in \mathcal{X}$, CIMNAR: $\mathbb{P}(M_j = 1 | x) = \mathbb{P}(M_j = 1 | x_{-j})$

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

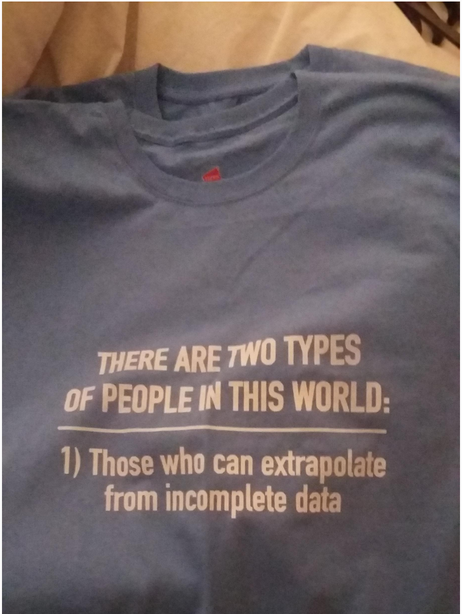
Conclusion

- ▷ 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.[24] Beyond MAR. $\forall$j $\in \{1, \dots, d\}$, $\forall$x $\in \mathcal{X}$, CIMNAR: $\mathbb{P}(M_j = 1|x) = \mathbb{P}(M_j = 1|x_{-j})$

- ▷ The quest for an FCS imputation method meeting all 3 points is open
- ▷ mice-DRF promising (code available)
- ▷ Imputation scores with missing values that are proper under MAR: ranking imputation methods
- ▷ Simulations MAR for benchmarks

[24] 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

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)

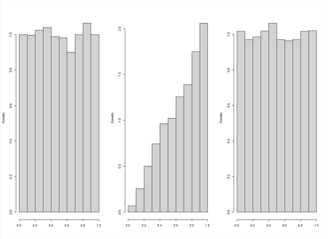
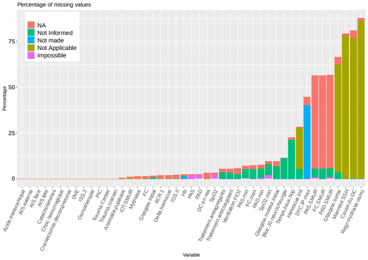


Figure 2: Distrib. of X_1 in different patterns. Left: Distrib. of $X_1 \mid M = m_3$. Middle: $(X_1 \mid M = m_1)$. Right: Distribution of all patterns for which X_1 is observed (Mixture of the distribution of X_1 in pattern 1 and 2).

• 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.

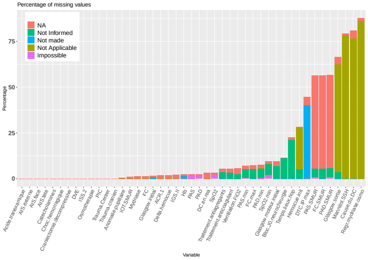
Missing data: important bottleneck in statistical practice



"One of the ironies of Big Data is that missing data play an ever more significant role"[25]

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

Missing data: important bottleneck in statistical practice



"One of the ironies of Big Data is that missing data play an ever more significant role"[25]

Complete case analysis: delete incomplete samples

- **Bias**: Resulting sample not representative of the target population
- **Information loss**: Take a matrix with d features where each entry is missing with probability $1/100$, remove a row (of length d) when one entry is missing

$$d = 5 \implies \approx 95\% \text{ of rows kept}$$
$$d = 300 \implies \approx 5\% \text{ of rows kept}$$

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