Imputation under Missing at Random

How to Impute and How to Evaluate Imputations

Jeffrey Näf

March 14, 2024

Disclaimer

Contains ongoing research, please do not (yet) distribute.

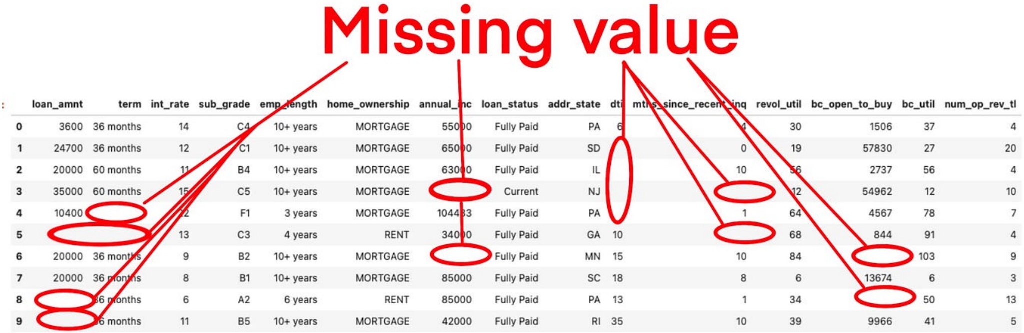
1. Background

2. MAR in the Pattern-Mixture Model

3. (Multiple) Imputation

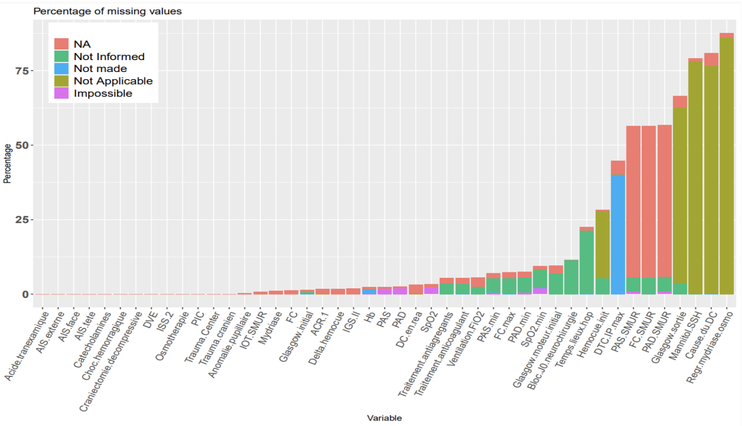
4. Imputation Scores

5. Conclusion



	loan_amnt	term	int_rate	sub_grade	emp_length	home_ownership	annual_inc	loan_status	addr_state	dti	mths_since_recent_inq	revol_util	bc_open_to_buy	bc_util	num_op_rev_tl
0	3600	36 months	14	C4	10+ years	MORTGAGE	55000	Fully Paid	PA	6	4	30	1506	37	4
1	24700	36 months	12	C1	10+ years	MORTGAGE	65000	Fully Paid	SD		0	19	57830	27	20
2	20000	60 months	11	B4	10+ years	MORTGAGE	63000	Fully Paid	IL		10	56	2737	56	4
3	35000	60 months	15	C5	10+ years	MORTGAGE		Current	NJ			12	54962	12	10
4	10400		12	F1	3 years	MORTGAGE	104433	Fully Paid	PA		1	64	4567	78	7
5			13	C3	4 years	RENT	34000	Fully Paid	GA	10		68	844	91	4
6	20000	36 months	9	B2	10+ years	MORTGAGE		Fully Paid	MN	15	10	84		103	9
7	20000	36 months	8	B1	10+ years	MORTGAGE	85000	Fully Paid	SC	18	8	6	13674	6	3
8		36 months	6	A2	6 years	RENT	85000	Fully Paid	PA	13	1	34		50	13
9		36 months	11	B5	10+ years	MORTGAGE	42000	Fully Paid	RI	35	10	39	9966	41	5

Figure: Source: Obtained from Medium


Percentage of missing values
NA
Not Informed
Not made
Not Applicable
Impossible
Percentage
75
50
25
0
Variable
Acide.tranexamique
AIS.externe
AIS.face
AIS.tete
Catecholamines
Choc.hemorragique
Craniectomie.decompressive
DVE
ISS.2
Osmotherapie
PIC
Trauma.Center
Trauma.cranien
Anomalie.pupillaire
IOT.SMUR
Mydriase
FC
Glasgow.initial
ACR.1
Delta.hemocue
IGS.II
Hb
PAS
PAD
DC.en.rea
SpO2
Traitement.antiagregants
Traitement.anticoagulant
Ventilation.FiO2
PAS.min
FC.max
PAD.min
SpO2.min
Glasgow.moteur.initial
Bloc.J0.neurochirurgie
Temps.lieux.hop
Hemocue.init
DTC.IP.max
PAS.SMUR
FC.SMUR
PAD.SMUR
Glasgow.sortie
Mannitol.SSH
Cause.du.DC
Regr.mydriase.osmo

Basic Ideas

- There are many potential ways how to deal with missing values, depending on the analysis
- A very natural idea: Replace the missing values with "reasonable" values
- This approach allows to do any further analysis (estimation/prediction) in a second step
- This is extremely common, especially also in machine learning
- Imputing multiple times, it is even possible to get some idea of the uncertainty coming from the missing values

Basic Ideas

- The imputation literature is somewhat messy; new imputation methods get developed left and right, seemingly without a common thread
- I will try here to develop a more systematic approach

Objectives of this Talk

- In this talk, the focus will lie on general-purpose (multiple) imputation of missing values
- While we will touch upon the more classical parametric ideas, the focus will be on more modern views of imputation

$$\mathbf{X}^* = \begin{pmatrix} x^*_{1,1} & x^*_{1,2} & x^*_{1,3} \\ x^*_{2,1} & x^*_{2,2} & x^*_{2,3} \\ x^*_{3,1} & x^*_{3,1} & x_{3,3} \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

Figure: Illustration: $\mathbf{X}^*$ is the assumed underlying full data, $\mathbf{M}$ is the vector of missing indicators and $\mathbf{X}$ arises when $\mathbf{M}$ is applied to $\mathbf{X}$.

-1.39620134	0.392990827	-1.793903529
-0.03127511	-0.399754625	0.377495535
NA	-1.534761247	0.253225074
0.76128208	0.223539621	-0.226450819
NA	1.159951856	-1.440915214
0.38855277	-0.349869646	2.203688869
0.29811721	-0.341478180	-0.046631397
-1.92132971	-2.026330592	-2.992404026
-0.87455388	NA	-0.047272703
NA	NA	0.501245405

-1.39620134	0.3929908	-1.7939035
-0.03127511	-0.3997546	0.3774955
0.76128208	0.2235396	-0.2264508
0.38855277	-0.3498696	2.2036889
0.29811721	-0.3414782	-0.0466314
-1.92132971	-2.0263306	-2.9924040
NA	-1.5347612	0.2532251
NA	1.1599519	-1.4409152
NA	NA	0.5012454

Basic Framework

- We assume to observe an i.i.d. sample $(X_1, M_1), \dots, (X_n, M_n)$ of n observations.
- X_i : Data Row i of dimension d with NAs, M_i : vector in $\{0,1\}^d$

  $X_{i,j}$ observed: $M_{i,j} = 0$

  $X_{i,j} = \text{NA}$: $M_{i,j} = 1$
- Since it's i.i.d. we can often simply consider one generic observation (X, M).
- Conceptually we assume there is an X^* with distribution P^*, such that $X_{i,j} = X^*_{i,j}$, whenever $M_{i,j} = 0$.
- Thus X^* is the vector of true underlying values, and X is the observed vector of values when X^* gets masked by M.

$$\mathbf{X}^* = \begin{pmatrix} x^*_{1,1} & x^*_{1,2} & x^*_{1,3} \\ x^*_{2,1} & x^*_{2,2} & x^*_{2,3} \\ x^*_{3,1} & x^*_{3,1} & x_{3,3} \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

Figure: Illustration: $\mathbf{X}^*$ is the assumed underlying full data, $\mathbf{M}$ is the vector of missing indicators and $\mathbf{X}$ arises when $\mathbf{M}$ is applied to $\mathbf{X}$.

- P refers to the distribution of X with missing values with density p
- $P^* \in \mathcal{P}$ refers to the distribution of X^* without missing values, with density p^*
- We let $\tilde{X}$ be the imputed X with imputation distribution H, with density h.

Two Views

- From the above: We have two random vectors (X, M) with a joint distribution.
- There are two common ways to define/model this distribution: The **Selection Model (SM)** and the **Pattern Mixture Model (PMM)**:

$$\text{Selection Model: } p^*(M = m, x) = \mathbb{P}(M = m \mid x) \cdot p^*(x)$$
$$\text{PMM Model: } p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$$

- SM view is most used, but especially for imputation, I find PMM much more useful!

Notation

- Let $\mathcal{M}$ be the set of all possible missingness patterns m.
- For a missingness pattern $m \in \mathcal{M}$, $o(x, m) = (x_j)_{j \in \{1,\dots,d\} : m_j = 0}$ subsets the observed elements of x according to m, while $o^c(x, m) = (x_j)_{j \in \{1,\dots,d\} : m_j = 1}$, subsets the missing elements.

Notation

- Let $\mathcal{M}$ be the set of all possible missingness patterns m.
- For a missingness pattern $m \in \mathcal{M}$, $o(x, m) = (x_j)_{j \in \{1,\ldots,d\}: m_j = 0}$ subsets the observed elements of x according to m, while $o^c(x, m) = (x_j)_{j \in \{1,\ldots,d\}: m_j = 1}$, subsets the missing elements.

$$x = (x_1, x_2, x_3, x_4, x_5), \quad m = (1, 1, 0, 1, 0)$$
$$\Longrightarrow o(x, m) = (x_3, x_5)$$
$$\Longrightarrow o^c(x, m) = (x_1, x_2, x_4)$$

Inria

Missingness Mechanisms: SM

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

- **Missing Completely at Random (MCAR)**: The probability of an entry being missing is completely independent of the data
- **Missing at Random (MAR)**: The probability of an entry being missing only depends on the observed values of the data
- **Missing not at Random (MNAR)**: Everything goes

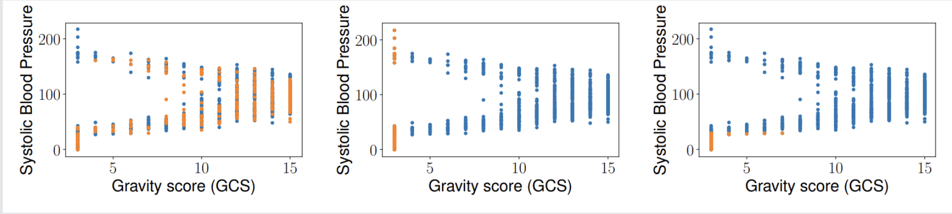


Figure: Gravity Score is always observed. From left to right: MCAR - MAR - MNAR

Formal MAR Definition Selection Model

Definition (SM-MAR)

The missingness mechanism is missing at random (MAR) if

$$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x,m)) \text{ for all } m \in \mathcal{M}, x. \tag{1}$$

MAR Example

- Consider an example with two variables: X_1 being the logarithm of **income**, and X_2 being **age**
- Assume a missing mechanism for the income X_1, whereby X_1 tends to be missing whenever age is "high"

$\Longrightarrow$ Thus the probability of income (X_1) being missing depends entirely on the value of age (X_2), which is always observed.

- This results in two patterns, one where both variables are fully observed (m_1) and a second (m_2), wherein X_1 is missing.
- If we assume that higher age is related to higher income, there is a clear shift in the distribution of income and age when moving from one pattern to the other.

Inria

MAR Example

We could model this with the following Gaussian mixture model for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$:

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

For both patterns, the conditional distribution of X_1 given X_2 is given as

$$p(x_1 \mid x_2, M = m_1) = p(x_1 \mid x_2, M = m_2) = N(x_2, 1)(x_1).$$

But the joint distribution of (X_1, X_2) is different in pattern m_1 than it is in m_2!

Historic MAR

- MAR was originally introduced in the seminal paper of Rubin [Rubin, 1976].
- There he proved an **ignorability result**: Under an important additional condition, a parameter of interest can be found with maximum likelihood, by only considering the observed part of the data
- Most lectures and books on missing values focus on this result, as it allows one to completely ignore missing values in a maximum likelihood context
- While it is an important result, it depends on strong parametric assumptions and I personally feel it is somewhat outdated

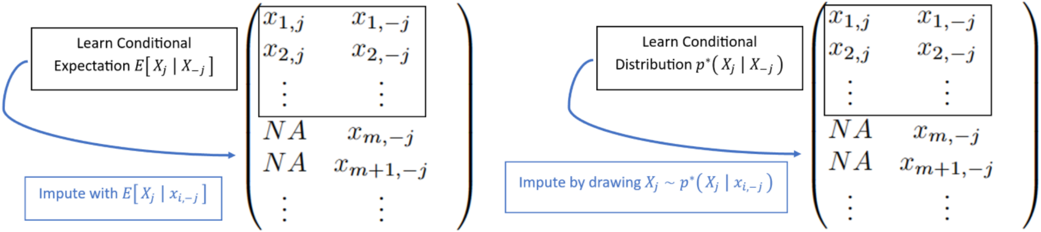
1. Background

2. MAR in the Pattern-Mixture Model

3. (Multiple) Imputation

4. Imputation Scores

5. Conclusion



- Need to make assumptions on X^*/P^* to make this possible
- In particular need assumptions on

$$p^*(o^c(x, m_2) \mid o(x, m_2), M = m') = p^*(x_1 \mid x_{-j}, M = m'),$$

for $m' = m_1$ and $m' = m_2$.

PMM Model : $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

$$\mathbf{X}^* = \begin{pmatrix} x^*_{1,1} & x^*_{1,2} & x^*_{1,3} \\ x^*_{2,1} & x^*_{2,2} & x^*_{2,3} \\ x^*_{3,1} & x^*_{3,1} & x_{3,3} \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

MAR Example

We could model this with the following Gaussian mixture model for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$:

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

For both patterns, the conditional distribution of X_1 given X_2 is given as

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

PMM Model : $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

$$\mathbf{X}^* = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,1} & x_{3,3} \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

Definition

The missingness mechanism is conditionally independent MAR (CIMAR) if

$$p^*(o^c(x,m) \mid o(x,m), M = m') = p^*(o^c(x,m) \mid o(x,m), M = m'')$$
for all $m, m', m'' \in \mathcal{M}, x.$ (CIMAR)

$$\text{PMM Model}: p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$$

$$\mathbf{X}^* = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,1} & x_{3,3} \end{pmatrix} \qquad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

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

$$\text{PMM Model} : p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$$

$$\mathbf{X}^* = \begin{pmatrix} x^*_{1,1} & x^*_{1,2} & x^*_{1,3} \\ x^*_{2,1} & x^*_{2,2} & x^*_{2,3} \\ x^*_{3,1} & x^*_{3,1} & x_{3,3} \end{pmatrix} \qquad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

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

PMM Model : $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

$$\mathbf{X}^* = \begin{pmatrix} x^*_{1,1} & x^*_{1,2} & x^*_{1,3} \\ x^*_{2,1} & x^*_{2,2} & x^*_{2,3} \\ x^*_{3,1} & x^*_{3,1} & x_{3,3} \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

Definition (PMM-MAR)

The missingness mechanism is missing at random (MAR) if

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

for all $m \in \mathcal{M}, x$. (PMM-MAR)

$$\text{PMM Model}: p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$$

$$\mathbf{X}^* = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,1} & x_{3,3} \end{pmatrix} \qquad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

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

PMM Model : $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

$$\mathbf{X}^* = \begin{pmatrix} x^*_{1,1} & x^*_{1,2} & x^*_{1,3} \\ x^*_{2,1} & x^*_{2,2} & x^*_{2,3} \\ x^*_{3,1} & x^*_{3,1} & x_{3,3} \end{pmatrix} \qquad \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}$$

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

A More Elaborate Example

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

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

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

SM-MAR:

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

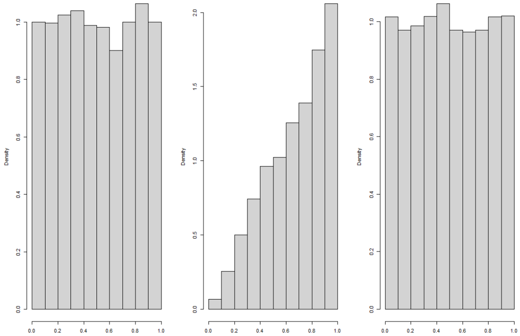


Figure: Left: Distribution we would like to impute $X_1 \mid M = m_3$. Middle: Distribution of X_1 in the fully observed pattern $(X_1 \mid M = m_1)$. Right: Distribution of all patterns for which X_1 is observed (Mixture of the distribution of X_1 in pattern 1 and 2).

A More Elaborate Example

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

Figure: Even conditional distributions can change under MAR

- Under MAR, not only the distribution of observed variables can change from pattern to pattern, but even $o^c(X, m) \mid o(X, m)$.
- Nonetheless, if imputation is done **iteratively**, it recovers the correct distributions under perfect estimation.

$\Longrightarrow$ **FCS Approach!**

1. Background

2. MAR in the Pattern-Mixture Model

3. (Multiple) Imputation

4. Imputation Scores

5. Conclusion

Imputation Approaches

- First, there are two broad classes of imputation approaches;
  - **Joint Modeling (JM)** methods that impute the data using one model: Examples include parametric distributions [Schafer, 1997], and more recently, Generative Adversarial Network (GAN)-based ([Yoon et al., 2018, Deng et al., 2022, Fang and Bao, 2023]) and Variational Autoencoder (VAE)-based methods ([Mattei and Frellsen, 2019, Nazábal et al., 2020, Qiu et al., 2020, Yuan et al., 2021])
  - **Fully Conditional Specification (FCS)** where a different model for each dimension is trained [van Buuren, 2007, van Buuren, 2018]: Most Prominent Example: Multiple Imputation by Chained Equations (MICE) methodology [van Buuren and Groothuis-Oudshoorn, 2011]
- Here we focus on the FCS approach

FCS Imputation

- Let in the following for $j \in \{1, \dots, d\}$,

$$X_{-j} = (X_l)_{l \neq j} \, .$$

- In the classical Fully Conditional Specification, we specify a probability distribution p_j for each $X_j \mid X_{-j}$.
- For several iterations we draw

$$x_j^{(t+1)} \sim p_j^{(t)}(x_j \mid x_{-j}^{(t)}),$$

where $p_j^{(t)}$ is updated/estimated in each iteration t.

X1	X2
NA	-1.879658573
NA	-2.534835620
-0.835628612	1.454974147
NA	2.329639344
0.329507772	0.250524041
NA	0.164414845
NA	0.563111651
NA	-1.114695987
NA	-2.426687462
-0.305388387	-0.599655950

X1	X2
0.845467412	0.664501159
0.467247396	-0.364692729
-0.402055064	0.542906157
-0.008055641	-0.209162216
-0.799126982	-0.830104755
1.004233021	0.629847025
-0.311973356	-1.603593030
NA	-1.879658573
NA	-2.534835620
NA	2.329639344

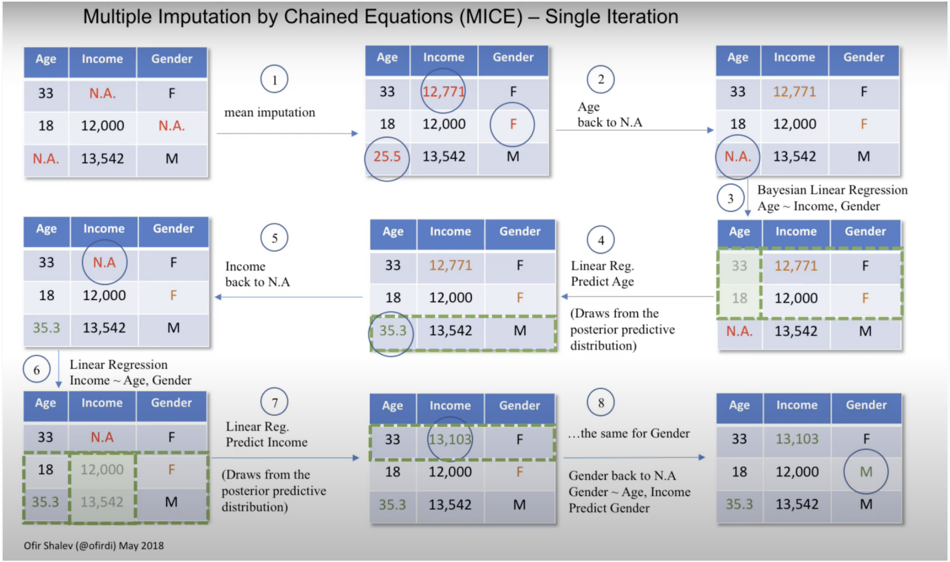


Figure: Source: [van Buuren, 2018]

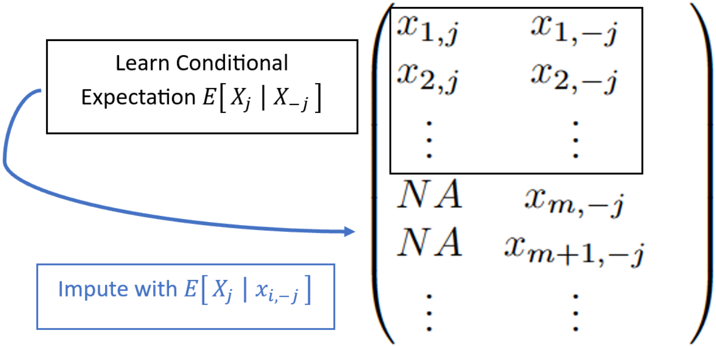
3 Lessons

- Lesson I: Imputation is a Generative Approach
- Lesson II: FCS might just work, but it is hard
- Lesson III: Imputation should be evaluated as a Generative Approach

Lesson I: Imputation is a Generative Approach

- The question of what is a “reasonable” value for the missing value is the question of **what kind of imputation to use**.
- In the FCS approach this corresponds to specifying p_j
- Often p_j is specified as a *point measure*
- Example: Methods that estimate $\mathbb{E}[X_j \mid x_{-j}^{(t)}]$ on observed data points and “draw”:

$$x_j^{(t+1)} \sim \delta_{\mathbb{E}[X_j|x_{-j}^{(t)}]}.$$


Learn Conditional Expectation $E[X_j \mid X_{-j}]$
Impute with $E[X_j \mid x_{i,-j}]$
$\begin{pmatrix} x_{1,j} & x_{1,-j} \\ x_{2,j} & x_{2,-j} \\ \vdots & \vdots \\ NA & x_{m,-j} \\ NA & x_{m+1,-j} \\ \vdots & \vdots \end{pmatrix}$

Lesson I: Imputation is a Generative Approach

- Example: Methods that estimate $\mathbb{E}[X_j \mid x_{-j}^{(t)}]$ on observed data points and "draw":

$$x_j^{(t+1)} \sim \delta_{\mathbb{E}[X_j | x_{-j}^{(t)}]}.$$

- While this can be good enough for certain applications, such as prediction, here we aim higher.

$\Longrightarrow$ The ideal imputation should draw samples from the conditional distribution of missing given observed: $p^*(o^c(x, m) \mid o(x, m))$.

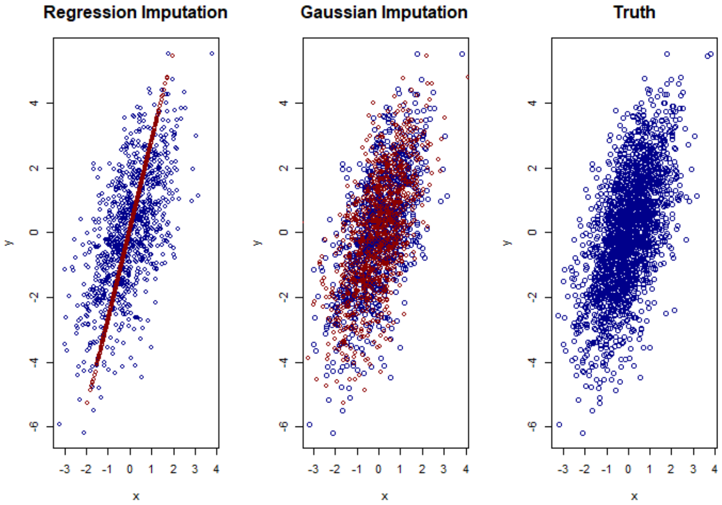


Figure: 5000 observations of the bivariate Gaussian Example with around 50% MCAR missing values in X_1.

Lesson I: Imputation is a Generative Approach

$\Longrightarrow$ The ideal imputation should draw samples from the conditional distribution of missing given observed: $p^*(o^c(x, m) \mid o(x, m))$.

- In particular: We should not look for the best value to impute
- In other words: **Imputation is not prediction**.
- p_j should not be a point distribution, but as close as possible to the true conditional distribution of $X_j \mid X_{-j}$.

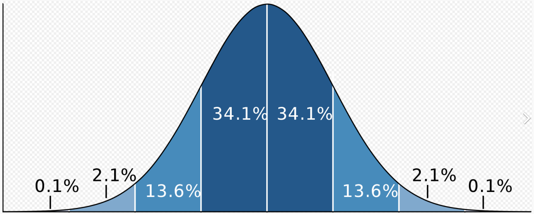


Figure: Source: Wikipedia

Learn Conditional Distribution $p^*(X_j \mid X_{-j})$

Impute by drawing $X_j \sim p^*(X_j \mid x_{i,-j})$

$$\begin{pmatrix} x_{1,j} & x_{1,-j} \\ x_{2,j} & x_{2,-j} \\ \vdots & \vdots \\ NA & x_{m,-j} \\ NA & x_{m+1,-j} \\ \vdots & \vdots \end{pmatrix}$$

Example: $p_1(x_1 \mid x_2) = N(\hat{\beta} x_2, \hat{\sigma}^2)(x_1)$

Multiple Imputation

- Another advantage of being able to draw from the conditional distribution, is the ability to generate **multiple imputations**.
- This allows to factor in the additional uncertainty of the missing values.

Lesson II: FCS might just work, but it is hard

- We have seen that distribution shifts are possible under MAR.
- In one example (age/income) only the marginal distributions shifted, but in the second example, even the conditional distribution could shift!
- Nonetheless one can show that **FCS identifies the right distributions**.

Lesson II: FCS might just work, but it is hard

Theorem

In a population setting (perfect estimation), FCS identifies the right distributions under MAR.

- However, with finite sample we don't have perfect estimation, and different imputation methods will perform differently.
- **How do we even evaluate an imputation method?**

Lesson III: Imputation should be evaluated as a Generative Approach

- A natural question is now, how we measure what is a “good” imputation method.
- How do we rank imputation methods in practice?
- Imagine an academic setting where the true underlying values are available
- In this scenario, researchers often use RMSE:

$$\sum_i \sqrt{\sum_j (\text{imputed}_{i,j} - \text{true}_{i,j})^2}$$

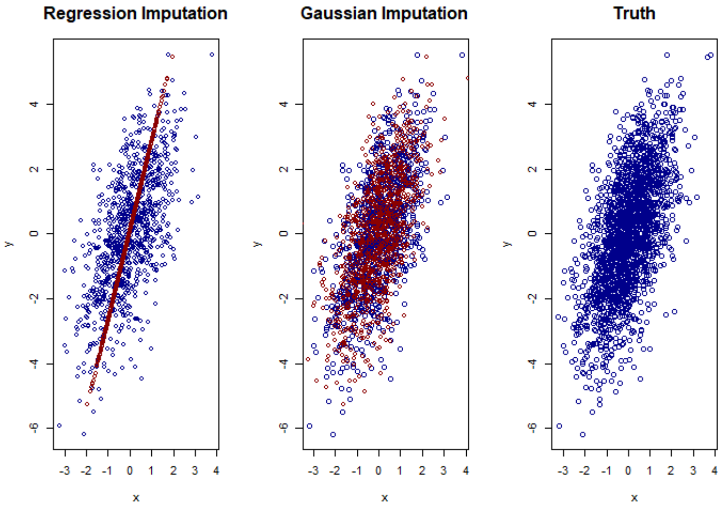


Figure: The imputation on the left has a lower RMSE than the imputation on the right

Lesson III: Imputation should be evaluated as a Generative Approach

- RMSE is minimized when we impute by the conditional expecation
- Instead we want a measure that is minimized when we draw from the right conditional distributions
- If the true values are available, this can be achieved by a **distributional metric**
- For instance we can estimate the **energy distance** between true and imputed data set:

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

for $X \sim P^*$, $Y \sim H$ and X', Y' an independent copy of X, Y.

Lesson III: Imputation should be evaluated as a Generative Approach

- For an imputation $\tilde{X}_1, \ldots, \tilde{X}_n$, obtained from an imputation distribution H, the energy distance $\mathit{energy}(H, P^*)$ can easily be estimated with the true values $X_1^*, \ldots, X_n^*$ from P^*
- In R: Package `energy`

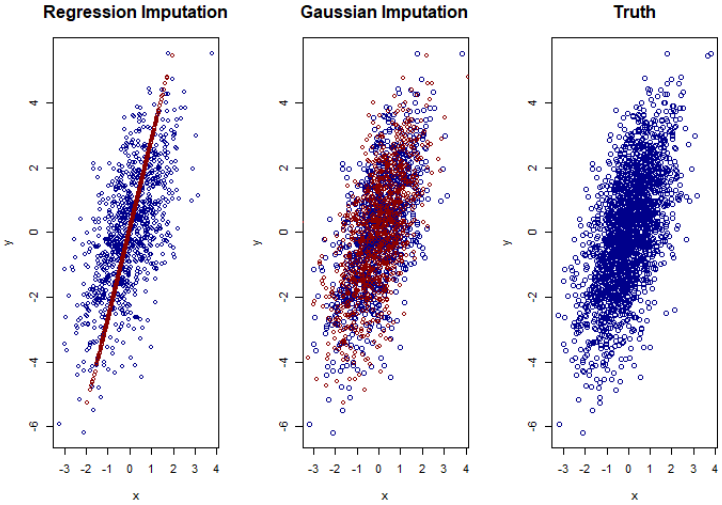
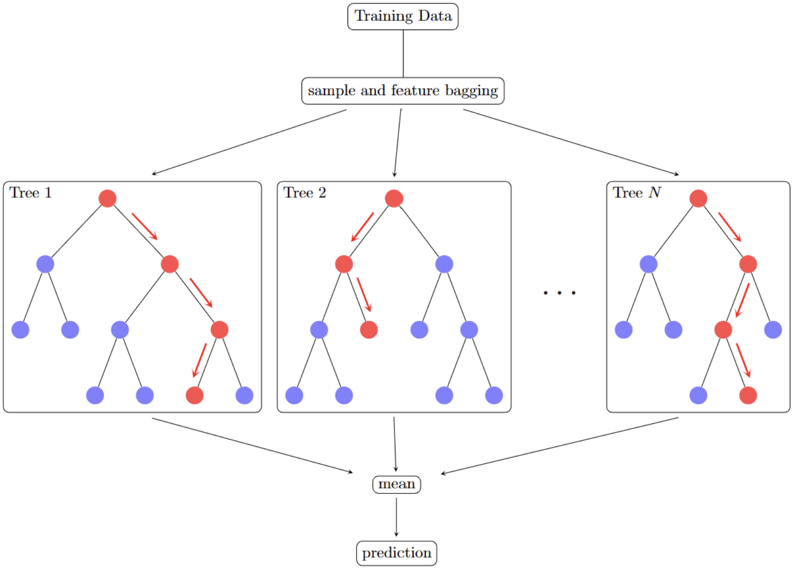


Figure: The imputation on the right has a lower energy distance than the imputation on the left

What is a good Imputation Method?

- So what should we take for p_j?
- In the age of machine learning, we can specify p_j as a method to estimate $p^*(x_j \mid x_{-j}^{(t)})$ **nonparametrically**
- For instance, we can specify that for each j we estimate $p^*(x_j \mid x_{-j}^{(t)})$ using an adaptation of Random Forest, the Distributional Random Forest (DRF) of [Ćevid et al., 2022] ("**mice-DRF**")
- This was also approximated earlier [Burgette and Reiter, 2010], using one regression tree + sampling from the leaves ("**mice-cart**")

Inria


Training Data
sample and feature bagging
Tree 1
Tree 2
. . .
Tree N
mean
prediction

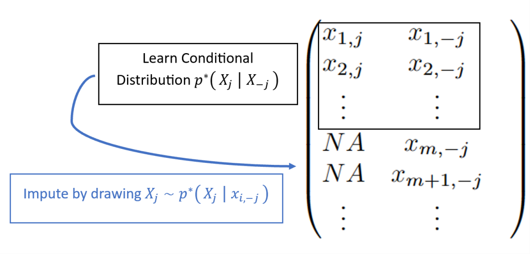
What is a good Imputation Method?

- So what should we take for p_j?
- For instance, we can specify that for each j we estimate $p^*(x_j \mid x_{-j}^{(t)})$ using an adaptation of Random Forest, the Distributional Random Forest (DRF) of [Ćevid et al., 2022] ("**mice-DRF**")
- This was also approximated earlier [Burgette and Reiter, 2010], using one regression tree + sampling from the leaves ("**mice-cart**")
- Though a proper study has yet to be done, prior analysis for tabular data indicates that both methods are extremely hard to beat and in particular outperform neural-net-based approaches!

What is a good Imputation Method?

An ideal imputation method should

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



What is a good Imputation Method?

An ideal imputation method should

(1) be a distributional regression method,

(2) be able to capture nonlinearities in the data,

(3) be fast to fit,

(4) be able to deal with distributional shifts in the observed variables.

- Though there is some indication that **mice-DRF** and **mice-cart** perform extremely well, they only meet (1)-(3).
- **missForest** of [Stekhoven and Bühlmann, 2011], which was touted as an extremely strong imputation method by several benchmarking studies, only meets (2) and (3).

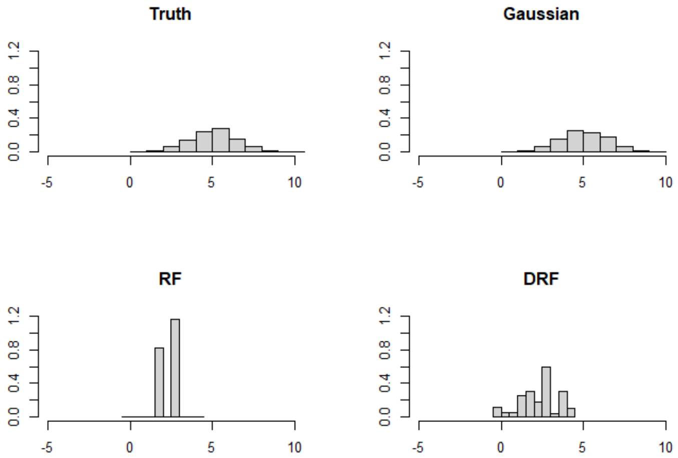


Figure: The true distribution against a draw from different imputation procedures for imputing X_1 in the income/age example.

1. Background

2. MAR in the Pattern-Mixture Model

3. (Multiple) Imputation

4. Imputation Scores

5. Conclusion

What if the underlying values are not available?

- The question of how to evaluate imputation methods becomes much harder when the true underlying values are not available
- The energy distance is directly related to the energy score [Gneiting and Raftery, 2007, Gneiting et al., 2008]:

$$es(H, y) = \mathbb{E}[\|X - y\|_{\mathbb{R}^d}] - \frac{1}{2}\mathbb{E}[\|X - X'\|_{\mathbb{R}^d}], \tag{3}$$

where $X \sim H$ and $X' \sim H$ is an independent copy.

General Idea of Scores

- Proper scores have been an active area of research in the last decade
- The idea is as important as it is simple: **A proper score is minimized in expectation (in a population setting) when one inserts the quantity of interest**

  RMSE: $\mathbb{E}_{Y \sim P^*}[||c - Y||^2_{\mathbb{R}^d}]$ is minimized when c is the expectation of Y.
  MAE: $\mathbb{E}_{Y \sim P^*}[|c - Y|]$ is minimized when c is the median of Y.
  Energy Score: $\mathbb{E}_{Y \sim P^*}[es(H, y)]$ is minimized when $H = P^*$

General Idea of Scores

- The **Energy** score can be used to score **distributional prediction**
- Assume we have learned a distribution H based on n samples, from which we can sample (for instance using DRF)
- We would like to test this distribution against a new test point y drawn from P^*
- Can use the Energy score:

$$S(y,H) = \mathbb{E}_{X\sim H}[\|X-y\|_{\mathbb{R}^d}] - \frac{1}{2}\mathbb{E}_{X\sim H}[\|X-X'\|_{\mathbb{R}^d}]$$

Theorem

In expectation, we score the true distribution lowest, i.e. :

$$S(P^*,H) := \mathbb{E}_{Y\sim P^*}[S(Y,H)] \geq \mathbb{E}_{Y\sim P^*}[S(Y,P^*)] := S(P^*,P^*)$$

General Idea of Scores

- The **Energy** score can be used to score **distributional prediction**
- Assume we have learned a distribution H based on n samples, from which we can sample (for instance using DRF)
- We would like to test this distribution against a new test point y
- Can use the Energy score:

$$S(y, H) = \mathbb{E}_{X \sim H}[\|X - y\|_{\mathbb{R}^d}] - \frac{1}{2}\mathbb{E}_{X \sim H}[\|X - X'\|_{\mathbb{R}^d}]$$

- **If we can sample from H, $S(y, H)$ can be easily approximated!**

Inria

Imputation Scores

- P refers to the distribution of X with missing values
- $P^* \in \mathcal{P}$ refers to the distribution of X^* without missing values.
- H refers to an imputation distribution.

Definition (Proper Imputation Score (I-Score))

A real-valued function $S_{NA}(H, P)$ is a proper I-Score iff

$$S_{NA}(H, P) \leq S_{NA}(P^*, P),$$

for any imputation distribution H.

Imputation Scores

- For this to work under the challenging MAR setting we unfortunately need to have a set of variables that is **always observed**.
- Lets call this set O, i.e. for all $j \in O$, $m_j = 0$ for all $m \in \mathcal{M}$
- A score that does not need that is also available and seems to work exceedingly well, but without theoretical guarantees.

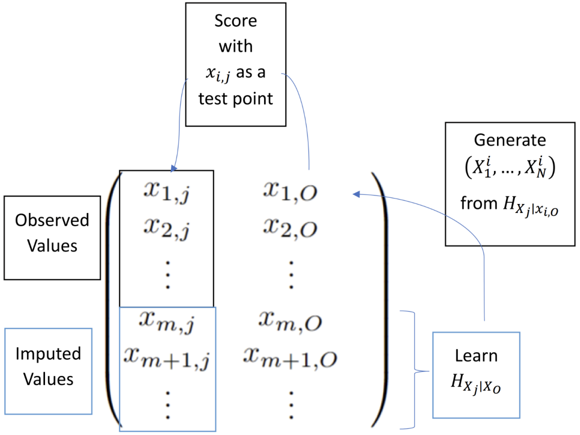


Figure: Illustration of the new scoring method. The PMM view shows that only certain conditional distributions can be compared under MAR. This is what we utilize here.

Score Estimation

- L_j: All patterns in $\mathcal{M}$ with $m_j = 1$ (i.e. all possible pattern in which X_j is observed)
- $(\tilde{X}_l^{(i)})$, $l = 1, \dots, N$ sample generated from the conditional imputation distribution $H_{X_j|x_{i,-j}}$

$$\hat{S}_{NA}^j(H,P) = \frac{1}{|i : m_i \in L_j|} \sum_{i:m_i \in L_j} \underbrace{\left(\frac{1}{2N^2} \sum_{l=1}^{N} \sum_{\ell=1}^{N} \|\tilde{X}_l^{(i)} - \tilde{X}_\ell^{(i)}\|_{\mathbb{R}} - \frac{1}{N} \sum_{l=1}^{N} \|\tilde{X}_l^{(i)} - x_{i,j}\|_{\mathbb{R}} \right)}_{\substack{\text{Estimated Energy Score with predictive distribution} \\ \text{represented by } (\tilde{X}_l^{(i)})_l \text{ and test point } x_{i,j}}},$$

Final score, $S_{NA}^{es}(H,P)$, is the average of $\hat{S}_{NA}^j(H,P)$ over j.

Propriety

Theorem

Assume MAR in (PMM-MAR) *holds and that O is not empty. Then the population version* $S_{NA}^{es}(H, P)$ *is a proper I-Score.*

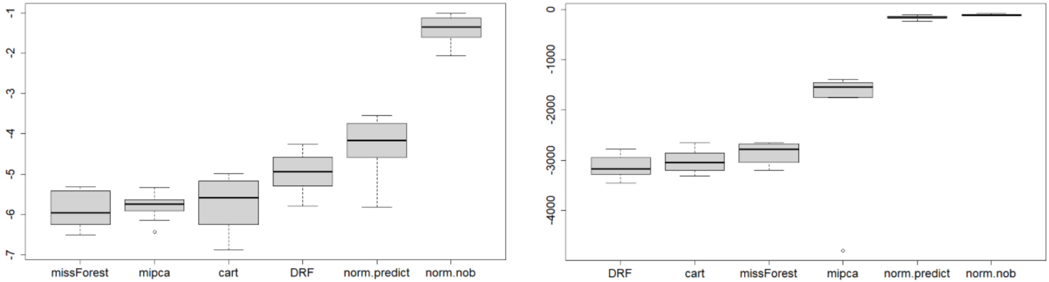


Figure: Left: Ordering of the I-score, Right: Ordering of the (negative) energy distance. The latter uses the true underlying values.

1. Background

2. MAR in the Pattern-Mixture Model

3. (Multiple) Imputation

4. Imputation Scores

5. Conclusion

Conclusion

- This talk discussed the PMM view of missingness that helps understand imputation under MAR
- We discussed 4 points the ideal imputation method should meet and potential ways to evaluate imputation methods
- Despite intensive research, the quest for an imputation method meeting all 4 points is still open
- We discussed the Imputation Scores and looked at a new score that is proper under MAR

Bibliography

Burgette, L. F. and Reiter, J. P. (2010).
Multiple Imputation for Missing Data via Sequential Regression Trees.
American Journal of Epidemiology, 172(9):1070–1076.

Ćevid, D., Michel, L., Näf, J., Meinshausen, N., and Bühlmann, P. (2022).
Distributional random forests: Heterogeneity adjustment and multivariate distributional regression.
Journal of Machine Learning Research, 23(333):1–79.

Deng, G., Han, C., and Matteson, D. S. (2022).
Extended missing data imputation via gans for ranking applications.
Data Mining and Knowledge Discovery, 36(4):1498–1520.

Fang, F. and Bao, S. (2023).
Fragmgan: generative adversarial nets for fragmentary data imputation and prediction.
Statistical Theory and Related Fields, 0(0):1–14.