# What Is a Good Imputation Under MAR Missingness?

Jeffrey Näf, Julie Josse

# What Is a Good Imputation Under MAR Missingness?*

Jeffrey Näf, Julie Josse
Inria, PreMeDICaL Team, University of Montpellier

**Abstract**

Missing values pose a persistent challenge in modern data science. Consequently, there is an ever-growing number of publications introducing new imputation methods in various fields. The present paper attempts to take a step back and provide a more systematic analysis: Starting from an in-depth discussion of the Missing at Random (MAR) condition for nonparametric imputation, we first develop an identification result, showing that the widely used Multiple Imputation by Chained Equations (MICE) approach indeed identifies the right conditional distributions. This result, together with two illuminating examples, allows us to propose four essential properties a successful MICE imputation method should meet, thus enabling a more principled evaluation of existing methods and more targeted development of new methods. In particular, we introduce a new method that meets 3 out of the 4 criteria. We then discuss and refine ways to rank imputation methods, even in the challenging setting when the true underlying values are not available. The result is a powerful, easy-to-use scoring algorithm to rank missing value imputations under MAR missingness.

**Keywords:** imputation, missing at random, distributional prediction, proper scores

## 1 Introduction

In this paper, we study general-purpose (multiple) imputation of missing data sets. That is, instead of imputing for a specific estimation goal or target, we focus on imputations that can be used in a second step for a wide variety of analyses. Developing such imputation methods is still an area of active research, as is benchmarking imputations. To categorize the wealth of imputation methods, one usually differentiates between joint modeling methods that impute the data using one (implicit or explicit) model and the fully conditional specification (FCS) where a different model for each dimension is trained (van Buuren, 2007, 2018). Examples of joint modeling include using 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)). Another set of examples are methods that use a sequential approach to joint modeling, whereby, for a certain ordering of variables, the joint distribution is specified through a sequence of conditional distributions, see e.g., Ibrahim et al. (1999); Lee and Mitra (2016); Xu et al. (2016); Murray (2018) among others. The most prominent example of FCS is the Multiple Imputation by Chained Equations (MICE) methodology (van Buuren and Groothuis-Oudshoorn, 2011). While there has been recent progress providing results for MAR imputation in the case of GAN-based methods (Deng et al. (2022); Fang and Bao (2023)), such results appear to lack for the FCS approach. Indeed, while some papers claim that imputation is possible under MAR using a methodology such as MICE, without providing a source, Fang and Bao (2023) claims MICE can only be used to impute MCAR data.

*We thank Giulia Marchello for providing the code for GAIN and MIWAE.

This paper provides new insights into this research by, among other things, proving that the FCS approach identifies the right distributions under MAR in a population setting and providing a list of desirable properties a successful regression method should meet for FCS. We address three questions: First, is imputation under MAR possible with the FCS approach? Formally, we study whether the conditional distribution needed to impute a missing value is identifiable from the data. Since we do not specify a parametrization and in particular, do not assume that the parameters of the missingness mechanism and the distribution of the data are distinct, this is not clear in general as we will demonstrate using the so-called pattern-mixture model (PMM) representation of missingness (Little (1993)). We then show that it is nonetheless the case that the imputation distribution is identifiable, allowing for nonparametric imputations in MAR settings using the FCS approach. Our identification result, though simple, appears to be stronger than what exists already. It shows that imputation with the FCS approach is feasible in principle. In particular, we compare the MAR condition we use to stronger conditions used in the context of GAN-based imputation methods in Deng et al. (2022) and Fang and Bao (2023). Second, what properties should the ideal imputation method have? We first illustrate that, despite this identification result, MAR imputation can be extremely challenging. For instance, we consider a simple two-dimensional MAR example with two patterns with widely varying distributions of the observed variable. Based on these insights we develop four properties a successful imputation method should meet in a FCS/MICE framework. In short, a successful imputation method under MAR needs to be a distributional regression method that is able to deal with *covariate shifts*. We discuss existing methods that meet some of these criteria and introduce a new method, denoted "mice-DRF". Third, given MAR missingness how can one generally find the best imputation for a given data set? This question is independent of whether the FCS or generative approach has been used and has not been addressed at all until very recently. The first important contribution towards solving this problem was made in Näf et al. (2023) who define the concept of Imputation Scores (I-Scores) to rank imputations. These scores are called "proper" if their population versions rank the imputation methods highest that imputes from the correct conditional distributions. We follow their argument in this paper that imputation is a distributional prediction task and needs to be evaluated as such. In particular, when comparing imputation methods, even under purely academic scenarios where the true underlying values are available, one should refrain from using measures such as the Root Mean Squared Error (RMSE), as already pointed out previously (van Buuren, 2018; Hong and Lynn, 2020; Näf et al., 2023). Measures like RMSE favor methods that impute conditional means instead of draws from the conditional distribution. This artificially strengthens the dependence between variables and leads to severe biases in parameter estimates and uncertainty quantification. Instead, an imputation method should draw from the conditional distribution of missing given observed, which might include values in the tail of the distribution. Currently, imputation methods are largely benchmarked and evaluated based on measuring the RMSE between the imputed and the underlying true values, see e.g., Waljee et al. (2013); Anil Jadhav and Ramanathan (2019); Bertsimas et al. (2018); Stekhoven and Bühlmann (2011); Nazábal et al. (2020); Qiu et al. (2020); Jäger et al. (2021); Yoon et al. (2018); Dong et al. (2021) and many others.

Instead, we advocate to use a distributional metric or score (Gneiting and Raftery, 2007) between actual and imputed data sets when the true values are available. For instance, we propose to evaluate imputation methods by calculating the energy distance (Székely, 2003) between real and imputed datasets. In the more realistic scenario when true values are not available, we advocate using proper I-Scores, as in Näf et al. (2023). However, we show that the score developed in Näf et al. (2023) is only proper under a condition much stronger than MAR and instead define a score that is indeed proper under MAR while also more computationally efficient and easier to implement.

The remainder of the article is organized as follows. The remainder of this section introduces notation and related work. In Section 2, we discuss the MAR condition and imputation in more detail with two illuminating examples and present our identification result. We then use these insights to present recommendations for imputation methods, including four properties the ideal imputation

method should meet in Section 3. Section 4 then turns to the question of how to evaluate imputation methods and presents a new proper I-Score. Finally, we illustrate the main points of this paper in a three empirical examples in Section 5. Code to replicate the experiments and to use the new scoring methodology can be found in `https://github.com/JeffNaef/MARimputation`.

## 1.1 Notation

We assume an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which all random elements are defined. Throughout, we take $\mathcal{P}$ to be a collection of probability measures on $\mathbb{R}^d$, dominated by some $\sigma$-finite measure $\mu$. We denote the (unobserved) complete data distribution by $P^* \in \mathcal{P}$ and by $P$ the actually observed distribution with missing values. We assume that $P$ ($P^*$) has a density $p$ ($p^*$). We take $X$ ($X^*$) to be the random vector with distribution $P$ ($P^*$) and let $x_i$ ($x_i^*$), $i = 1, \dots, n$, be realizations of an i.i.d. copy of the random vector $X$ ($X^*$). Similarly, $M$ is the random vector in $\{0,1\}^d$, encoding the missingness pattern of $X$, with realization $m$, whereby for $j = 1, \dots, d$, $m_j = 0$ means that variable $j$ is observed, while $m_j = 1$ means it is missing. For instance, the observation $(\texttt{NA}, x_2, x_3)$ corresponds to the pattern $(1, 0, 0)$. We denote the support of $X$ as $\mathcal{X} \subset \mathbb{R}^d$ $M$ as $\mathcal{M} \subset \{0,1\}^d$.

To denote assumptions on the missingness mechanism, we use a notation along the lines of Seaman et al. (2013). For each realization $m$ of the missingness random vector $M$ we define with $o(X, m) := (X_j)_{j \in \{1,\dots,d\}: m_j = 0}$ the observed part of $X$ according to $m$ and with $o^c(X, m) := (X_j)_{j \in \{1,\dots,d\}: m_j = 1}$ the corresponding missing part. Note that this operation only filters the corresponding elements of $X$ according to $m$, regardless of whether or not these elements are actually missing or not. For instance, we might consider the unobserved part $o^c(X, m)$ according to $m$ for the fully observed $X$, that is $X \sim P | M = \mathbf{0}$, where $\mathbf{0}$ denotes the vector of zeros of length $d$.

As in Näf et al. (2023), we define $\mathcal{H}_P \subset \mathcal{P}$ to be the set of imputation distributions compatible with $P$, that is

$$\mathcal{H}_P := \{\mathcal{H} \in \mathcal{P} : H \text{ admits density } h \text{ and } h(o(x,m)|M = m) = p(o(x,m)|M = m) \text{ for all } m \in \mathcal{M}\}, \tag{1.1}$$

where as above for a pattern $m$, $o(x, m) = (x_j)_{j \in \{1,\dots,d\}: m_j = 0}$ 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[1]. Clearly, $P^* \in \mathcal{H}_P$, so that the true distribution $P^*$ can be seen as an imputation.

## 1.2 Contributions

Inspired by the discussion in Molenberghs et al. (2008); Näf et al. (2023), we study the MAR condition under the framework of pattern-mixture models (PMMs) introduced in Little (1993), which we argue is more natural for imputation. Overall, we present four main contributions: First, we thoroughly analyse different MAR conditions through the lens of imputation. Crucially, we do not follow the traditional assumption that the distribution of $X$ is parametrized by a vector $\theta$ and the distribution of $M \mid X$ by a distinct vector $\phi$. This removes the question of parameters of interest and allows to study general-purpose nonparametric imputation. Second, we provide an identification result for the FCS approach under the weakest MAR assumption. As the result concerns the identification of conditional distributions in a MAR setting, it can also be applied to the sequential approach of joint modeling as we show in a corollary. Third, based on the previous two contributions we discuss four essential properties a successful imputation method needs to meet in the FCS/MICE framework under MAR. We moreover discuss methods that approximately meet most of these criteria, including a new methodology which we refer to as mice-DRF. As an added benefit, this new methodology is

[1] Note that while $h$ and $p$ are densities on $\mathbb{R}^d$, notation is slightly abused by using expressions such as $h(o(x,m)|M = m)$ and $p(o(x,m)|M = m)$, which are densities on $\mathbb{R}^{|\{j: m_j = 0\}|}$.

able to impute a block of several variables at once, potentially reducing the heavy computational burden of MICE in high dimensions. We provide an implementation of this imputation method, based on the `mice` package, as part of our freely available code. Fourth, we discuss the evaluation of imputation methods and show that the Imputation Score developed in Näf et al. (2023) is not proper under MAR and build a new easy-to-use score with propriety under MAR. The new score is simple to implement and remarkably accurate, even in challenging examples, as we demonstrate empirically in Section 5. Though the score needs a set of fully observed variables to be provably proper, we also discuss an alternative version of the score that empirically works well even when all variables have missing values. Throughout, we provide a discussion of imputation under MAR that connects several threads of literature and also make the connection to classical ignorability in a likelihood framework.

### 1.3 Related Work

Though the literature on missingness is vast, the results and discussions presented in this paper are new to the best of our knowledge. Most papers discussing MAR add the additional assumption that the distribution of $X$ and $M \mid X$ are parametrized by two distinct sets of parameters as mentioned above, leading to the classical ignorability result of Rubin (1976). This simplifies the analysis and generally avoids the issues we discuss here. For instance, while the FCS and, in particular, the MICE approach has been studied theoretically (Little and Rubin, 1986; Liu et al., 2014; Zhu and Raghunathan, 2015) under this ignorability, the problems of identification in this general setting appear to not have been discussed before. Instead, these papers generally focus on the challenging problem of potential incompatibility of the conditional models and analyze the convergence and asymptotic properties of the FCS iterations. Our aim is in a sense much simpler, as we want to answer the question of whether the right conditional distributions are identifiable under MAR when no assumption on the parametrization is placed.

As the paper views missingness through the lens of pattern-mixture models of Little (1993), the conceptually closest papers to ours are those based on the Generative Adversarial Network (GAN) approach: Both Deng et al. (2022); Fang and Bao (2023) make use of the PMM view in their proofs, without explicitly mentioning this, as does the original GAIN paper of Yoon et al. (2018). We essentially provide a similar identification result for the FCS or sequential approach under MAR as Deng et al. (2022) provide for their GAN-based approach. Despite the simplicity of our identification result, it appears to be stronger than what exists in the literature. For instance, the identification results in Deng et al. (2022); Fang and Bao (2023) for GAN-based methods rely on stronger MAR conditions, as shown below. Similarly, Tian (2017) claims the full distribution is recoverable under MAR, but uses a conditional independence condition that is much stronger than the MAR condition we consider. Indeed, graph-based papers concerned with recoverability usually assume variables that are always observed and formulate MAR as conditional independence statements, see e.g Doretti et al. (2018). This is much stronger than the traditional MAR condition of Rubin (1976). To the best of our knowledge, we are also the first to propose a list of properties an imputation method in the FCS framework should have, based on a thorough analysis of the MAR condition. This list complements existing guidelines on general imputation methods with a different focus, see e.g., Murray (2018, Section 4). Finally, when considering the evaluation of imputation methods, we build upon the arguments in Näf et al. (2023) but heavily improve their score to develop a score that is truly proper under MAR, in the sense that it provably ranks the best imputation method highest in a population setting.

## 2 Sequential Imputation under MAR

In the following, we first define MAR properly, following Rubin (1976); Seaman et al. (2013); Mealli and Rubin (2015), and analyze several different MAR conditions relevant to our discussion. The

| Selection Model: $\mathbb{P}(M = m \mid x)p^*(x)$ | Pattern Mixture Model: $p^*(x \mid M = m)\mathbb{P}(M = m)$ |
|---|---|
| $\mathbb{P}(M = m\|x) = \mathbb{P}(M = m\|\tilde{x})$ for all $m \in \mathcal{M}$ and $x, \tilde{x}$ such that $o(x, m) = o(\tilde{x}, m)$ (SM-MAR) | $p^*(o^c(x, m) \mid o(x, m), M = m) = p^*(o^c(x, m) \mid o(x, m))$ for all $m \in \mathcal{M}, x \in \mathcal{X}$ (PMM-MAR) |
| $\mathbb{P}(M = m\|x) = \mathbb{P}(M = m\|o(x, m))$ for all $m \in \mathcal{M}, x \in \mathcal{X}$ (SM-MAR II) | |
| | $p^*(o^c(x, m) \mid o(x, m), M = m') = p^*(o^c(x, m) \mid o(x, m))$ for $m' = m$ or $m' = 0$ and all $x \in \mathcal{X}$ (EMAR) |
| | $p^*(o^c(x, m) \mid o(x, m), M = m') = p^*(o^c(x, m) \mid o(x, m))$ for all $m \in \mathcal{M}, m' \in \mathcal{M}$, $x \in \mathcal{X}$ (CIMAR) |
| $\mathbb{P}(M = m\|x) = \mathbb{P}(M = m)$ for all $m \in \mathcal{M}$, $x \in \mathcal{X}$ (SM-MCAR) | $p^*(x \mid M = m) = p^*(x \mid M = m') = p^*(x)$ for all $m \in \mathcal{M}, m' \in \mathcal{M}$, $x \in \mathcal{X}$ (PMM-MCAR) |

**Table 1:** *Summary of the different MAR conditions discussed in this paper, when available both in the selection model and the pattern-mixture model. The conditions are ordered from weakest (top) to strongest (bottom). Conditions on the same level are equivalent.*

different definitions considered here are summarized in Table 1. Crucially, we do not assume anything about parametrization and instead purely focus on statements about conditional distributions. Using two examples illustrating these definitions, we show that this "nonparametric" view on MAR leads to nontrivial identification problems due to potential distribution shifts. We then present our identification result showing that identification is nonetheless possible in a population setting if one learns the conditional distribution using all available patterns. Finally, we return to the parametrized distribution case and contrast our findings with classical ignorability results in a likelihood framework.

## 2.1 MAR Definitions

We first properly define what we mean by MAR in the framework of the so-called *selection model* (SM, Little (1993)). In this framework, the joint distribution of $X$ and $M$ is factored as,

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

Through this view MAR is defined as:

**Definition 2.1.** *The missingness mechanism is missing at random (MAR) if*

$$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|\tilde{x}) \text{ for all } m \in \mathcal{M} \text{ and } x, \tilde{x} \text{ such that } o(x, m) = o(\tilde{x}, m). \quad \text{(SM-MAR)}$$

This is sometimes referred to as "Always Missing at Random", see e.g., Mealli and Rubin (2015); Deng et al. (2022). One can also weaken this requirement to be true only for the data and patterns that are actually observed, which is usually referred to as Realized MAR (RMAR). The arguments in this paper go through with slight modification, also in the case of RMAR, thus we focus on (SM-MAR) for simplicity. An alternative way to define MAR is

**Definition 2.2.** *The missingness mechanism is missing at random (MAR) if*

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

This is the definition used for instance in Molenberghs et al. (2008). Note that $o(x, m)$ is different for each $m$, and thus neither (SM-MAR) nor (SM-MAR II) are statements about conditional

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

**Figure 1:** *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* $\mathbf{X}^*$*. Thus each row of* $\mathbf{X}/\mathbf{X}^*$ *is an observation under a different pattern. Under condition* (CIMAR)*, the distribution of* $X_1, X_2 \mid X_3$ *is not allowed to change when moving from one pattern to another, though the marginal distribution of* $X_3$ *is allowed to change. In contrast, under MCAR* (PMM-MCAR)*, no change is allowed. Under MAR* (PMM-MAR) *the only constraint is that the distribution of* $X_1, X_2 \mid X_3$ *in the third pattern is the same as the unconditional one.*

independence as remarked in Mealli and Rubin (2015). Nonetheless, (SM-MAR II) is very intuitive: For any value of $m$, we assume that the probability of this value occurring only depends on the observed part of $x$. We show below that both definitions are indeed equivalent.

Considering instead the *pattern-mixture model* (PMM) framework (Little, 1993), we observe

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

This view emphasizes that the data we observe in $X$ are masked data from a vector $X^* \mid M$ and in particular, when learning quantities from one pattern, we have to be careful when changing to another, as distributions can change from pattern to pattern. A typical example is the Gaussian pattern-mixture model, whereby

$$X^* \mid M = m \sim N(\mu_m \mid \Sigma_m),$$

so that the distribution in each pattern might follow a different Gaussian distribution. It is well-known (Little, 1993), that the parameters of a pattern-mixture model are generally not identifiable without restrictions on how the distributions are allowed to change. Thus an immediate question becomes how the MAR condition constrains these distributions. This was answered in Molenberghs et al. (2008). We first give a new definition for better readability:

**Definition 2.3.** *The missingness mechanism is missing at random (MAR) if*

$$p^*(o^c(x,m) \mid o(x,m), M = m) = p^*(o^c(x,m) \mid o(x,m)) \\ \text{for all } m \in \mathcal{M}, x \in \mathcal{X}. \tag{PMM-MAR}$$

**Proposition 2.1** (Molenberghs et al. (2008))**.** *Condition* (SM-MAR II) *is equivalent to* (PMM-MAR)*.*

**Corollary 2.1.** *Condition* (SM-MAR) *is equivalent to* (SM-MAR II) *and both are equivalent to* (PMM-MAR)*.*

**Remark.** *The proof starting from* (SM-MAR) *is taken from Näf et al. (2023), though they wrongly thought to have proven equivalence to the stronger condition* (CIMAR) *below.*

To understand this condition, and how weak it in fact is, it makes sense to first consider a stronger, but more intuitive condition:

**Definition 2.4.** *The missingness mechanism is conditionally independent MAR (CIMAR) if*

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

This is a conditional independence statement, namely that $o^c(X, M) \mid o(X, M)$ is independent of $M'$. That is, no matter what pattern $m'$ is considered, the distribution of $o^c(X, M) \mid o(X, M)$ remains the same. As such, (CIMAR) allows to learn the distribution of $o^c(X, m) \mid o(X, m)$ from any pattern $m'$. It in turn is still weaker than MCAR however, which requires that

**Definition 2.5.** *The missingness mechanism is missing completely at random (MCAR), if*

$$p^*(x \mid M = m) = p^*(x \mid M = m') = p^*(x) \text{ for all } m \in \mathcal{M}, m' \in \mathcal{M}, x \in \mathcal{X}. \qquad \text{(PMM-MCAR)}$$

Figure 1 illustrates these different conditions in a small example.

Under (CIMAR), the observed variables can widely change their distribution from pattern to pattern, as shown in the following example:

**Example 1.** *Consider 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*

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

*where $N(x_2, 1)(x_1)$ is the univariate Gaussian density with mean $x_2$ and variance 1 evaluated at $x_1$. We first verify that the condition in* (CIMAR) *holds:*

$$\begin{aligned}
p^*(x_1 \mid x_2) &= \frac{P(M = m_1)p^*(x_1, x_2 \mid M = m_1) + P(M = m_2)p^*(x_1, x_2 \mid M = m_2)}{P(M = m_1)p^*(x_2 \mid M = m_1) + P(M = m_2)p^*(x_2 \mid M = m_2)} \\
&= \frac{\left(P(M = m_1)p^*(x_2 \mid M = m_1) + P(M = m_2)p^*(x_2 \mid M = m_2)\right)p^*(x_1 \mid x_2, M = m_2)}{P(M = m_1)p^*(x_2 \mid M = m_1) + P(M = m_2)p^*(x_2 \mid M = m_2)} \\
&= p^*(x_1 \mid x_2, M = m_2) \\
&= p^*(x_1 \mid x_2, M = m_1).
\end{aligned}$$

*However, the distribution of $X_2$ in pattern $m_1$ ($N(0, 1)$) is heavily shifted compared to pattern $m_2$ ($N(5, 1)$). Section 3 demonstrates how different imputation methods struggle to deal with this shift in distribution on simulated data.*

In the above example, we only have 2 patterns and thus (PMM-MAR) and (CIMAR) turn out to be equivalent and both hold in this example. However, an example with 3 patterns shows that (PMM-MAR) is strictly weaker than (CIMAR):

**Example 2.** *Consider*

$$\mathbf{X} = \begin{pmatrix}x_{1,1} & x_{1,2} & x_{1,3}\\ x_{2,1} & NA & x_{2,3}\\ NA & x_{3,2} & x_{3,3}\end{pmatrix}, \mathbf{M} = \begin{pmatrix}0 & 0 & 0\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix} = \begin{pmatrix}m_1\\ m_2\\ m_3\end{pmatrix}. \tag{2.1}$$

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

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

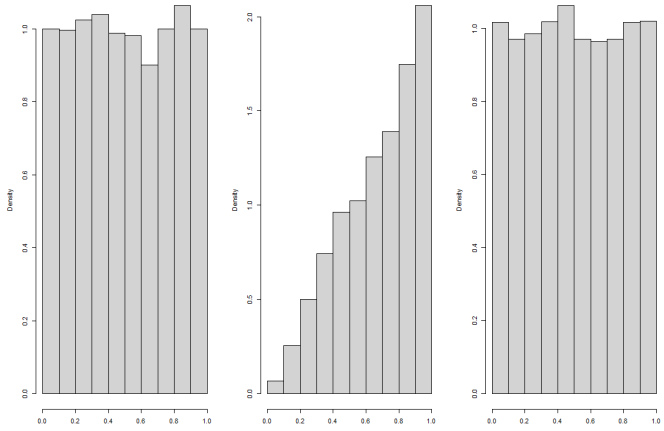

**Figure 2:** *Illustration of Example 2. Left: Distribution we would like to impute $X_1 \mid M = m_3$. Middle: Distribution of $X_1$ in the fully observed pattern $(X_1 \mid M = m_1)$. Right: Distribution of all patterns for which $X_1$ is observed (Mixture of the distribution of $X_1$ in patterns $m_1$ and $m_2$).*

*It can be checked that these are valid distributions, as in particular, $\sum_m \mathbb{P}(M = m) = 1$ and $\sum_m \mathbb{P}(M = m \mid x_j) = 1$ for $j = 1, \ldots, 3$. Moreover, $\mathbb{P}(M = m \mid x) = \mathbb{P}(M = m \mid o(x, m))$ and thus the MAR condition* (SM-MAR) *holds. In particular, for variable $x_1$ in pattern $m_3$, it holds that*

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

*However, if we consider $x_1$ given $(x_2, x_3)$ in the first pattern, we have:*

$$\begin{aligned} p^*(x_1 \mid x_2, x_3, M = m_1) &= \frac{\mathbb{P}(M = m_1 \mid x_1, x_2, x_3)}{\mathbb{P}(M = m_1 \mid x_2, x_3)} p^*(x_1 \mid x_2, x_3) \\ &= x_1 p^*(x_1 \mid x_2, x_3), \end{aligned}$$

*showing that* (CIMAR) *does not hold. Figure 2 illustrates this behavior: It shows the distribution of $X_1$ in different patterns. As the distribution of $(X_2, X_3)$ in the different patterns is always the same, this directly illustrates the change in the conditional distribution of $X_1 \mid X_2, X_3$ when changing from pattern $m_1$ to pattern $m_3$. The key is thus that* (PMM-MAR) *still allows for a change in the conditional distributions over different patterns. That is the distribution $X_1 \mid X_2, X_3$ in pattern $m_1$ is different in the above example from the distribution $X_1 \mid X_2, X_3$ in pattern $m_3$. All that is required is that the distribution $X_1 \mid X_2, X_3$ in pattern $m_3$ corresponds to the unconditional one.*

Thus we have just shown that

**Proposition 2.2.** *MCAR* (PMM-MCAR) *is strictly stronger than* (CIMAR) *which is strictly stronger than* (PMM-MAR).

Another important MAR condition is the extended MAR condition:

**Definition 2.6.** *The missingness mechanism is extended missing at random (EMAR), if*

$$\begin{aligned} &p^*(o^c(x, m) \mid o(x, m), M = m') = p^*(o^c(x, m) \mid o(x, m)) \\ &\text{for } m' = m \text{ or } m' = 0, \text{ for all } x \in \mathcal{X}. \end{aligned} \tag{EMAR}$$

This is clearly stronger than (PMM-MAR) and weaker than (CIMAR). Moreover, it is a useful condition as it allows to learn any conditional distribution of missing given observed from the fully observed pattern.

## 2.2 FCS in MAR

The previous discussion illustrates that from (PMM-MAR) alone, it is not clear whether learning a distribution in one pattern allows to impute values in the other pattern. However, in Example 2 while $P(M = m_1 \mid x_1)$ depends on $x_1$, $P(M = m_1 \mid x_1) + P(M = m_2 \mid x_1)$ does not. This is the key property in the proof of identification under MAR as it implies that $p^*(x_1 \mid x_2, x_3)$ needed for imputation can be identified if *all* patterns for which $x_1$ are observed are considered. We detail this now.

The goal of the FCS in general and the MICE approach in particular is to impute by iteratively drawing for all $j \in \{1, \dots, d\}$ and $t \geq 1$,

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

whereby $x_{-j}^{(t)} = \{x_l^{(t)}\}_{l \neq j}$ are the imputed and observed values of all other variables except $j$ at the $t$th iteration. Doing this repeatedly leads to a Gibbs sampler that converges under quite mild conditions (Little and Rubin (1986, Chapter 10.2.4.)). Naturally, if one does not have access to the true distribution $p^*$ and estimates the conditional model nonparametrically, this is a very complicated problem to analyze theoretically. Here we focus on a very simple question: If $x_{-j}$ has already been imputed by the correct distribution, that is we have access to the true underlying $(d-1)-$variate distribution $p^*(x_{-j})$, can we successfully impute $x_j$ by only looking at the patterns where $x_j$ is observed? This view connects to Example 2 and avoids any question of convergence of the Gibbs sampler to focus purely on identification.

Let in the following,

$$L_j = \{m \in \mathcal{M} : x_j \in o(x, m)\}, \tag{2.2}$$

be the set of patterns in which $x_j$ is observed. The best action one can do in this case is to draw from the distribution,

$$\begin{aligned} &h^*(x_j \mid x_{-j}) \\ &= \sum_{m \in L_j} \frac{\mathbb{P}(M = m)}{\sum_{m \in L_j} p^*(x_{-j} \mid M = m)\mathbb{P}(M = m)} p^*(x \mid M = m), \end{aligned} \tag{2.3}$$

which is the conditional distribution of $X_j \mid X_{-j}$ *learned from all patterns in which* $x_j$ *is observed.* Owing to the above example, the question is whether under MAR, $h^*(x_j \mid x_{-j})$ is indeed the same as $p^*(x_j \mid x_{-j})$;

**Proposition 2.3.** *Assume MAR in* (PMM-MAR) *holds. Then for* $h^*(x_j \mid x_{-j})$ *as in* (2.3)*,*

$$h^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j}), \tag{2.4}$$

*for all* $x_{-j}$ *with* $p^*(x_{-j}) > 0$*.*

This shows that the desired distribution is indeed recoverable in principle from all available patterns. Intuitively 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 $m \in \mathcal{M}$, it can be shown that the MAR condition implies that both aggregated patterns have the same conditional distribution $X_j \mid X_{-j}$, thus allowing to identify the right conditional distribution in the pattern where $X_j$ is observed.

Even with perfect estimation, conditioning on $X_{-j}$ would require iteration over several imputations, as mentioned above. To make the result more tangible, we can study the following simplified procedure that avoids iteration entirely: Assume in the following that one variable is fully observed. That is the possible pattern in $\mathcal{M}$ all share one zero, or

$$O = \{l : m_l = 0 \text{ for all } m \in \mathcal{M}\}, \tag{2.5}$$

is not empty. Without loss of generality, we assume that this fully observed variable is the $p$th one. Then for $j \in \{1, \dots, p-1\}$, let $L_j$ be defined as in (2.2). We then impute by drawing observations from

$$h^*(x_j \mid x_{j+1}, \dots, x_p) = \sum_{m \in L_j} \frac{\mathbb{P}(M = m)}{\sum_{m \in L_j} p^*(x_{j+1}, \dots, x_p \mid M = m)\mathbb{P}(M = m)} p^*(x_j, x_{j+1}, \dots, x_p \mid M = m),$$

which is the conditional distribution $X_j \mid X_{j+1}, \dots, X_p$ learned from all patterns $m_l$, $l \in L_j$. This in fact corresponds to the sequential approach to joint modeling, see e.g., Murray (2018) and the references therein. We denote the resulting distribution of the fully imputed data as $H^*$ with density $h^*$. For this simplified imputation approach it holds that:

**Corollary 2.2.** *Assume MAR in* (PMM-MAR) *holds and that $O$ in* (2.5) *is not empty. Then $H^* \in \mathcal{H}_P$ has*

$$h^*(x) = p^*(x), \text{ for all } x. \tag{2.6}$$

Proposition 2.3 and Corollary 2.2 show that sequential imputation with an algorithm that can perfectly learn the distribution under MAR is indeed identified, in the sense that we are able to learn the true conditional distribution needed to impute a missing value. The key for the proof is that (1) all available patterns are used to learn a distribution of $x_j \mid x_{j+1}, \dots, x_p$, (2) use of (SM-MAR II), which is equivalent to (PMM-MAR), and (3) that the conditional distributions $P(M = m \mid x)$ still need to sum to 1 over all values of $m$.

**Remark.** *In particular, Proposition 2.3 and Corollary 2.2 show that the FCS approach can identify the right conditional distributions under a weaker condition than GAN-based approaches. Deng et al. (2022) show that their GAN architecture is able to impute missingness under EMAR,* (EMAR). *This condition allows to learn a distribution from the fully observed pattern and is thus strictly stronger than* (PMM-MAR). *Similarly, Fang and Bao (2023) show that their GAN-based method can identify the conditional distribution of missing given observed data. However, while they claim this shows identification under MAR, the condition they present in Section 3.2. is actually stronger and more akin to* (CIMAR).

From the above, it can be seen that we can in general not just simply learn the conditional distributions from the fully observed data and then impute the missing variables. Instead, we need to consider *all* patterns wherein a variable $x_j$ is observed to be able to impute it. We now want to highlight why this discussion of distribution shifts under MAR may not be relevant for Maximum Likelihood Estimation (MLE).

## 2.3 Ignorability in Maximum Likelihood Estimation

In the context of MLE, it has long been established (Rubin, 1976) that the missing mechanism is ignorable under MAR and an additional condition. This additional condition is critical for our discussion. To formalize this assume $p^*$ is parametrized by a vector $\theta$. Moreover, assume the conditional distribution of $M \mid x$ is parametrized by $\phi$. Then we can rewrite the MAR definition in (SM-MAR) slightly, as in Rubin (1976); Mealli and Rubin (2015):

$$\begin{aligned} &\mathbb{P}_\phi(M = m|x) = \mathbb{P}_\phi(M = m|\tilde{x}) \text{ for all } m \in \mathcal{M} \\ &\text{and } x, \tilde{x} \text{ such that } o(x, m) = o(\tilde{x}, m). \end{aligned} \tag{2.7}$$

As so far, $\phi$ and $\theta$ are not restricted to be finite-dimensional, this could in principle be assumed without loss of generality, such that (2.7) is indeed the same as condition (SM-MAR). In the

following, we will assume for simplicity that $\theta$ is finite-dimensional. Let $\Omega_\theta$ be the space of $\theta$, $\Omega_\phi$ the space of possible $\phi$ and $\Omega_{\theta,\phi}$ the joint space of the parameters. The crucial additional condition is that:

$$\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi. \tag{2.8}$$

This just means that $\phi$ is distinct from $\theta$, so that $\mathbb{P}_\phi(M = m|x)$ does not depend on $\theta$ (Rubin, 1976; Seaman et al., 2013; Mealli and Rubin, 2015). In this case, we can rederive the classical ignorability result for MAR in a likelihood context: Consider the likelihood for a pattern $m$,

$$\mathcal{L}(\theta; o(x,m)) = p^*_{\theta,\phi}(o(x,m), M = m) = \int p^*_{\theta,\phi}(x, M = m) do^c(x,m).$$

That is, $\mathcal{L}(\theta; o(x,m))$ is the joint density of the observed values with respect to pattern $m$, and $M = m$, seen as a function of $\theta$. Under (2.7) it can be checked that

$$\begin{aligned}\int p^*_{\theta,\phi}(x, M = m) do^c(x,m) &= \mathbb{P}_\phi(M = m \mid o(x,m)) p^*_\theta(o(x,m)) \\ &= c(o(x,m)) p^*_\theta(o(x,m)),\end{aligned}$$

with $c(o(x,m))$ not depending on $\theta$. Consequently, it is possible to ignore the missingness mechanism (and potential distribution shifts) in a likelihood setting due to (a) the assumption of distinct parameters $\theta, \phi$ (2.8) and (b) the nature of maximum likelihood. In particular, even though the distribution $p^*_{\theta,\phi}(o(x,m), M = m)$ is not the same as the $p^*_\theta(o(x,m))$, it is *essentially* the same from an MLE perspective: We can therefore simply maximize $p^*_\theta(o(x,m))$ over $\theta$ to get the MLE. Whether this ignorability holds under MAR is a question of *parametrization*, as we illustrate in Example 2:

**Example 3** (Example 2 Continued)**.** *Consider again the setting of Example 2, that is:*

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

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

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

*Now assume that the parameter of interest is the upper boundary of $x_1$, such that $X_1$ is uniform on $[0,\theta]$. As $\mathbb{P}(M = m_i \mid x)$ does not change, it follows that:*

$$p^*_{\theta,\phi}(x_1, x_2, x_3, M = m_1) = \mathbb{P}(M = m_1 \mid x_1) p_\theta(x_1, x_2, x_3) = \frac{x_1}{3} p^*_\theta(x_1, x_2, x_3). \tag{2.9}$$

*Thus for optimization purposes, maximizing $p^*_{\theta,\phi}(x_1, x_2, x_3, M = m_1)$ over $\theta$ is equivalent to maximizing $p^*_\theta(x_1, x_2, x_3)$ over $\theta$. In particular, being able to identify $\theta$ allows to identify $p^*_\theta(x_1 \mid x_2, x_3)$ and thus to impute $x_1$. This, despite the fact that*

$$p^*_\theta(x_1, x_2, x_3, M = m_1) = \frac{x_1}{3} p^*_\theta(x_1, x_2, x_3) \neq p^*_\theta(x_1, x_2, x_3).$$

*Having obtained $\theta$, it is then possible to impute $X_1$ in the third patterns by drawing from $p^*_\theta(x_1 \mid x_2, x_3)$. However, notice that this is not the same as saying that $\theta$ can be recovered from only looking at the first pattern $m_1$. Indeed in this case:*

$$p^*_{\theta,\phi}(x_1, x_2, x_3 \mid M = m_1) = \frac{\mathbb{P}(M = m_1 \mid x_1)}{\mathbb{P}(M = m_1)} p_\theta(x_1, x_2, x_3) = \frac{x_1}{\theta} p_\theta(x_1, x_2, x_3), \tag{2.10}$$

*as $P(M = m_1) = \theta/3$. Thus maximizing $p^*_{\theta,\phi}(x_1, x_2, x_3 \mid M = m_1)$ is not equivalent to maximizing $p^*_\theta(x_1, x_2, x_3)$. On the flipside, if one changes $\mathbb{P}(M = m_1 \mid x_1)$ to $x_1/3\theta$, violating (2.8), maximizing $p_{\theta,\phi}(x_1, x_2, x_3, M = m_1)$ will not recover $\theta$.*

# 3 Requirements for Imputation Methods

We have seen that both conditional as well as marginal distribution shifts can occur for different patterns under MAR. However, conditional shifts can be disregarded when using a sequential approach (i.e. MICE), as, for a variable $X_j$, considering all patterns $m$ in which $X_j$ is missing identifies the right conditional distribution. Nonetheless, marginal distribution shifts, such as in Example 1, can still occur. In particular, a successful imputation method needs to be able to deal with distributional shifts in the observed variables. Moreover, in practice, an estimation method in the FCS framework should be able to estimate the potentially complex distribution of $X_j \mid X_{-j}$ as accurately as possible.

The above considerations thus suggest desirable properties an imputation method should meet in an FCS framework: It should

(1) be a distributional regression method,

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

(3) be fast to fit,

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

Moreover, a helpful property to allow to use the FCS approach in high dimensions is if

(5) the method is able to deal with multivariate responses.

missForest which was shown to be extremely successful in terms of RMSE in various benchmarking analyses, only meets (2) and (3). In particular, random forest imputation such as missForest was deemed more successful than GAN-based methods in Jäger et al. (2021). However, a more appropriate scoring will very likely reverse these insights, ranking GAIN and similar methods higher than missForest.[2] On the other hand, Wang et al. (2022) which provides a careful benchmarking of imputation methods more in line with this paper, finds that mice-cart and mice-RF (Burgette and Reiter, 2010; Doove et al., 2014) are more successful than GAIN. These methods use one or several trees respectively, but sample from the leaves to obtain the imputation, approximating draws from the conditional distribution to approximate (1). Similarly, Näf et al. (2023) find mice-cart/RF to be extremely successful imputation methods. As such, they could be combining the best of both worlds; inheriting the accuracy of missForest, while providing draws from the conditional distribution. However, they are ultimately not designed for the task of distributional regression. Thus a forest-based distributional method such as DRF of Ćevid et al. (2022) might even attain better results and indeed meets (1)–(3). Moreover, DRF is designed to handle multivariate outputs and thus also meets (5). This makes the method accessible to high-dimensional datasets, as MICE can be used in blocks as described in van Buuren (2018, Chapter 4.7). We implemented this option in our new mice-DRF R function. Thus if $d = 1000$, one might define blocks of size 100 and in each pass, train DRF by regression a 100 variables on the remaining 900. This would reduce the number of passes in each iteration from 1000 to 10. We thus implement the following routine in mice: For each $j$, fit a DRF regressing the observed $x_{i,j}$ onto $x_{i,-j}$ to obtain an estimate of the conditional distribution, given by forest-induced weights. For each unobserved $x_{i,j}$, we predict the weights based on $x_{i,-j}$ and draw from the observed set according to those weights. This is essentially the mice-RF implementation described in Doove et al. (2014), with the traditional Random Forest exchanged by the Distributional Random Forest.

However, as a forest-based method, DRF still generalizes poorly outside of the training set, i.e. Requirement (4) is not met. Figure 3 illustrates the behavior of different imputation strategies for Example 1. First, the Gaussian imputation simply fits a regression in pattern $m_1$ and then draws from a conditional Gaussian distribution given the estimated parameters. As such it is the ideal

[2] Though our experiments in Appendix B suggest otherwise.

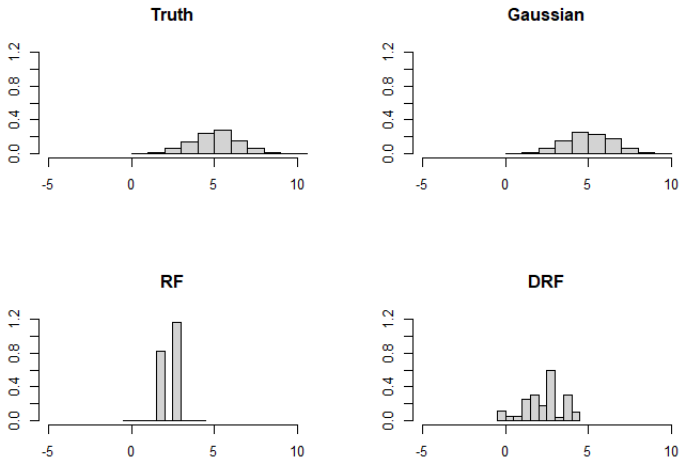


**Figure 3:** *The true distribution against a draw from different imputation procedures for imputing $X_1$ in Example 1.*

method in this setting and serves as an illustration that the data can be correctly imputed. For the nonparametric methods, DRF, as a distributional method, performs better than mice-RF. However, it still fails to deal with the covariate shift, centering around 2, when it should center around 5.

Thus, while previous analysis indicates that forest-based methods such as mice-cart, mice-RF, and likely also mice-DRF might be some of the most successful methods currently available, and in particular will likely beat GAN-based methods such as GAIN, finding an imputation method that (approximately) meets (1)–(5) is still an open problem.

Finally, the above list overlaps with and complements the three points mentioned in Murray (2018, Section 4) for general imputation methods:

(1') Imputations should reflect uncertainty about missing values and about the imputation model.

(2') Imputation models should generally include as many variables as possible.

(3') Imputation models should be as flexible as possible.

The first part of (1') corresponds to (1) of our list; instead of providing the *best* value for imputation, one should draw from the right conditional distribution to impute, such that the underlying distribution is replicated. To reiterate this, Figure 4 shows a small example. However, as we note in Section 6, the second part, that the uncertainty of the imputation model should be considered as well, is not met by the imputation methods we present here and is an open problem for nonparametric imputation. While this gets less consequential in large samples, this additional uncertainty is needed for reliable uncertainty quantification with multiple imputation. Point (2') is not relevant to our discussion, while (3') coincides with (2) above.

# 4 Assessing Imputation Methods

We now turn to the question of how to find the best out of several imputations. First, the above discussion suggests that in academic scenarios, where the true underlying values are available, distributional distances or scores should be used to evaluate imputation methods. We will in the following

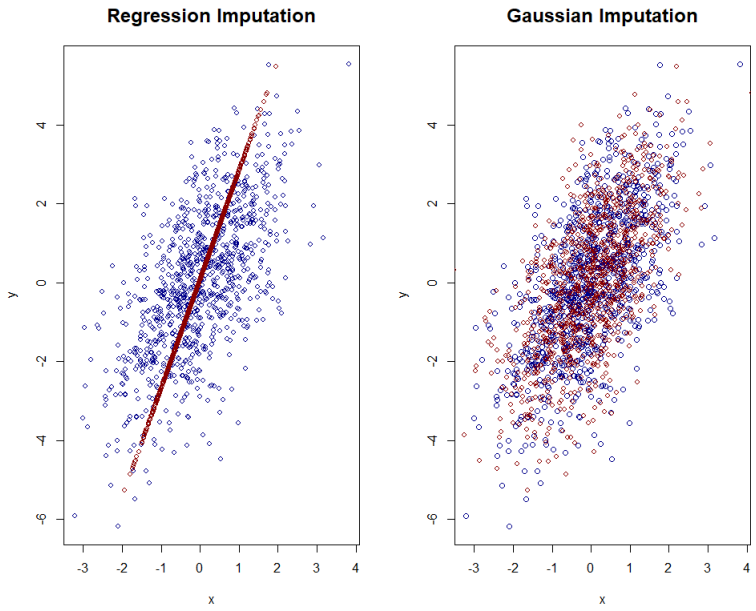


**Figure 4:** *5000 observations of a bivariate Gaussian Example with around 50% MCAR missing values in $X_1$. Left: Imputation by fitting a regression model and imputing the prediction, Right: Imputation by fitting a regression model and imputing by drawing from a conditional Gaussian distribution. Parameters calculated with the regression imputation tend to have a large bias, more so than if only complete-case analysis is used.*

use the (negative) energy distance between imputed and 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 $Y$. The energy distance is directly related to the energy score (Gneiting and Raftery, 2007; Gneiting et al., 2008):

$$es(H, y) = \frac{1}{2}\mathbb{E}[\|X - X'\|_{\mathbb{R}^d}] - \mathbb{E}[\|X - y\|_{\mathbb{R}^d}], \tag{4.1}$$

where $X \sim H$ and $X' \sim H$ is an independent copy. Let $S(H, P^*) = \mathbb{E}[es(H, Y)]$, where the expectation is taken over $Y \sim P^*$. Gneiting and Raftery (2007) showed that

$$S(H, P^*) \leq S(P^*, P^*), \tag{4.2}$$

i.e. $S$ is proper in the traditional sense. That is, if we predict the distribution $P^*$, and the "test data" $y$ are indeed drawn from $P^*$, taking the average over a "large" number of $y$ will lead to the maximal value. We will make use of the energy score to create a reliable ranking method when the underlying data are not available. To this end, we consider the I-Scores framework of Näf et al. (2023):

**Definition 4.1** (Definition 4.1 in Näf et al. (2023))**.** *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 $H \in \mathcal{H}_P$. It is strictly proper iff the inequality is strict for $H \neq P^*$.*

The key is that we would like to score $P^*$, when only samples from $P$ are available.

Näf et al. (2023) developed a first I-Score using Density Ratios and random projections $A \subset \{1, \dots, d\}$. The score was shown to be proper under (CIMAR). In Appendix C.1 we show that it is however not proper under MAR. Following the arguments in this paper, we now develop a score that is not only easier to use but also proper under MAR, without any projections. However, it necessitates that there is at least one variable that is always observed, or that $O$ defined in (2.5) is not empty. Adapting the proof of Proposition 2.3, the perfect imputation method learns the distribution,

$$\begin{aligned} &h^*(x_j \mid x_O) \\ &= \sum_{m \in L_j} \frac{\mathbb{P}(M = m)}{\sum_{m \in L_j} p^*(x_O \mid M = m)\mathbb{P}(M = m)} p^*(x_j, x_O \mid M = m), \end{aligned}$$

which is simply the conditional distribution of $x_j \mid x_O$ learned from all patterns in which $x_j$ is not missing. Consequently, $\mathbb{E}[es(H_{X_j|x_O}, Y)]$, with the integration taken over $Y \sim H^*_{X_j|x_O}$, is maximal when $h(x_j \mid x_O) = h^*(x_j \mid x_O)$ by propriety of the energy score. We then define the score of variable $j$ as

$$S^j_{NA}(H, P) = \mathbb{E}[\mathbb{E}[es(H_{X_j|X_O}, Y)]], \tag{4.3}$$

where the outer expectation is taken over $X_O \sim P^*_O$, the distribution of all fully observed variables. Usually, in the scoring literature, one only considers the inner expectation, even though in practice "scores are reported as averages over comparable sets of probabilistic forecasts" (Gneiting et al., 2008, page 222). We thus also consider the outer expectation to model the different test points. Finally, the full score is given as

$$S^{es}_{NA}(H, P) = \frac{1}{|O^c|} \sum_{j \in O^c} S^j_{NA}(H, P),$$

whereby $O^c$ is the complement of $O$, i.e. the set of all variables with at least one missing element. Since by Proposition 2.3, $h^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j})$, for all $x_{-j}$ with $p^*_{-j}(x_{-j}) > 0$, we obtain:

**Proposition 4.1.** *Assume MAR in* (PMM-MAR) *holds and that $O$ is not empty. Then $S^{es}_{NA}(H, P)$ is a proper I-Score.*

In practice, we propose the following two approximations to this approach: Consider a dimension $j \notin O$ and recall that $L_j$ collects all patterns $m$, such that $m_j = 0$. For each observed $(x_{i,j})_{m_i \in L_j}$, we assume to have a sample of $N$ points, say $(\tilde{X}^{(i)}_l)$, $l = 1, \dots, N$, approximately generated from $H_{X_j|x_{i,O}}$. This can be used to estimate $S^j_{NA}(H, P)$, as

$$\hat{S}^j_{NA}(H, P) = \frac{1}{|i : m_i \in L_j|} \sum_{i : m_i \in L_j} \left( \frac{1}{2N^2} \sum_{l=1}^{N} \sum_{\ell=1}^{N} \|\tilde{X}^{(i)}_l - \tilde{X}^{(i)}_\ell\|_{\mathbb{R}} - \frac{1}{N} \sum_{l=1}^{N} \|\tilde{X}^{(i)}_l - x_{i,j}\|_{\mathbb{R}} \right), \tag{4.4}$$

as in Gneiting et al. (2008, Equation (7)). Thus the observed points of $X_j$ act as the "test points" for the predicted distribution $H_{X_j|X_O}$. The final score is then given as

$$\hat{S}^{es}_{NA}(H, P) = \frac{1}{|O^c|} \sum_{j \in O^c} \hat{S}^j_{NA}(H, P). \tag{4.5}$$

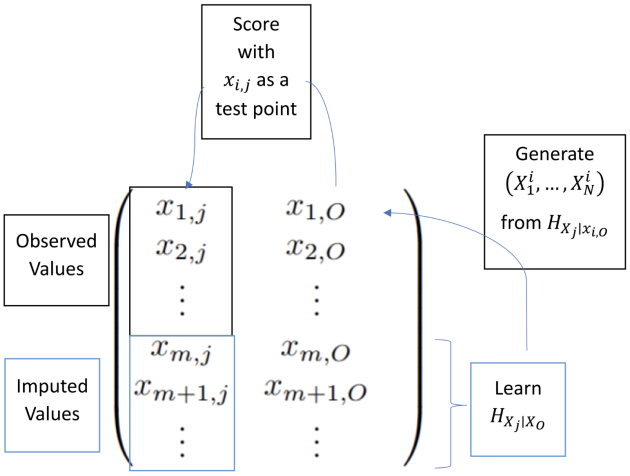


**Figure 5:** *Conceptual illustration of the score approximation.*

**Remark.** *Formally, always observed observations are needed to ensure the test points $x_{i,j}$ are truly sampled from $h^*(x_j \mid x_O)$, which in turn is equal to $p^*(x_j \mid x_O)$. While these points are observed, their marginal distribution is fixed to $p^*(x_j)$, but since $x_{-j}$ is imputed and thus drawn from $H_{-j}$, relative to the imputed point $x_{i,-j}$, the test point $x_{i,j}$ might not be sampled from the right distribution $p^*(x_j \mid x_{-j})$. Appendix A presents an informal argument, indicating that in general, the score might not be proper if all variables $x_{i,-j}$ instead of $x_{i,O}$ are used. Nonetheless, another version of the score is presented there, based on $X_{-j}$ instead of $X_O$. Though it remains an open problem under which conditions this score can be proven to be proper, it works remarkably well empirically.*

We now detail how we obtain $(\tilde{X}_l^{(i)})$, $l = 1, \ldots, N$. Given an imputed data set and the imputation function itself, we subset and concatenate the imputed points and observed points $x_{i,j}$ of $j$ and the fully observed points $x_{i,O}$, $i = 1, \ldots, n$. In this new data set, we keep the imputed points, that is all $X_{i,j}$ with $m_i \in L_j^c$ are still drawn from $H$, while we set the *observed* observations of $X_j$ to missing, i.e. $X_{i,j} = \texttt{NA}$ for $i$ with $m_i \in L_j$:

$$\begin{pmatrix} \texttt{NA} & (x_{i,O})_{m_i \in L_j} \\ (x_{i,j})_{m_i \in L_j^c} & (x_{i,O})_{m_i \in L_j^c} \end{pmatrix} \tag{4.6}$$

Then we approximate the sampling from $H_{X_j|x_{i,O}}$ in two ways:

(1) Regress $(x_{i,j})_{m_i \in L_j^c}$ onto $(x_{i,O})_{m_i \in L_j^c}$ in (4.6) using DRF. Then for each test point $x_{i,O}$, $m_i \in L_j$, sample $N$ times from the estimated conditional distribution obtained from DRF.

(2) Impute the NA values in (4.6) with $H$, $N$ times.

We refer to the first approach as drf-I-Score, and to the second as $m$-I-Score. The idea in both cases is to use $X_O$ and the imputation of $X_j$ to generate a sample from the distribution $H_{X_j|X_O}$ for *points that are already observed*. Note that while the drf-I-Score does this by utilizing the sampling of DRF, the $m$-I-Score uses the ability of the imputation method itself to generate samples. Thus, while the drf-I-Score can be averaged over several imputations to score multiple imputations, the $m$-I-Score scores multiple imputation naturally.

As a downside, the $m$-I-Score can be computationally demanding, as $N$ should be chosen high, say at least 50 to give an accurate score. This would be infeasible for realistic dimensions if the

full data set had to be imputed. However note that in step (2), by construction only one variable has missing values, while all the others are observed. This means that only one pass is needed to impute, which essentially corresponds to fitting the chosen model (e.g., RF) once.

# 5 Empirical Study

The goal of this section is to illustrate the concepts discussed in this paper on both simulated and real data, including the performance of the new score. We employ the FCS methods discussed above, namely mice-cart and the new mice-DRF, missForest, as well as regression and Gaussian imputations used in the previous section. Both fit a regression to the observed data to obtain the regression parameters. The regression imputation then simply imputes by predicting from the linear regression model, while Gaussian imputation uses the prediction as the mean of a Gaussian distribution from which it draws imputed values. However, in the following, we will follow the naming guideline of the R-package `mice` (van Buuren and Groothuis-Oudshoorn, 2011) and refer to the regression imputation as mice-norm.predict and to the Gaussian imputation as mice-norm.nob. If a method requires the specification of parameters, we use the default values. To evaluate the imputation methods we calculate the (negative) energy distance between the true and imputed data sets, using the `energy` R-package (Rizzo and Szekely, 2022). As this "score" is able to access the true underlying values, we will refer to it as the full information score. We compare the orderings of the full information score with the drf- and $m$-I-Score, which do not have access to the values underlying the missing values. The only hyperparameter to choose in this case is the number of samples $N$, which we will set to $N = 100$. Finally, though we focus on FCS imputation methods here, we also add a comparison to GAIN (Yoon et al., 2018) and MIWAE (Mattei and Frellsen, 2019), in terms of the full information score in the Appendix B.

The three examples considered in this section, as well as the analysis in Appendices A, B and further tests that are not shown here, indicate that:

(I) For the methods and data sets considered here mice-DRF and mice-cart are the most promising methods. This aligns with the findings in Wang et al. (2022); Näf et al. (2023). In particular, they tend to perform stronger than missForest, GAIN, and MIWAE.

(II) However, none of the methods is able to reliably deal with distributional shifts and nonlinearity, showing once again that better imputation methods need to be found.

(III) Contrary to our speculation in Section 3, Appendix B indicates that GAIN and MIWAE do not beat missForest, even in terms of negative energy distance. However, it remains to be seen whether this changes for higher dimensional data sets.

(IV) The ordering of the $m$-I-Score is quite sensible and similar to the one of the full information score, even in the first challenging distributional shift example in Section 5.2. If differences arise, it is often because the $m$-I-Score penalizes methods that cannot produce multiple imputations. Given the discussion in this paper, this might be desirable. An exception is the third example in Section 5.3 where none of the methods perform well. Here the scores disagree quite heavily.

(V) Remarkably, the score using the full data $X_{-j}$ in Appendix A appears to work as well as the one using $X_O$.

## 5.1 Air Quality Data

We start with the air quality data set obtained from `https://github.com/lorismichel/drf/tree/master/applications/air_data/data/datasets/air_data_benchmark2.Rdata`. This is a preprocessed version of the data set that was originally obtained from the website of the Environmental

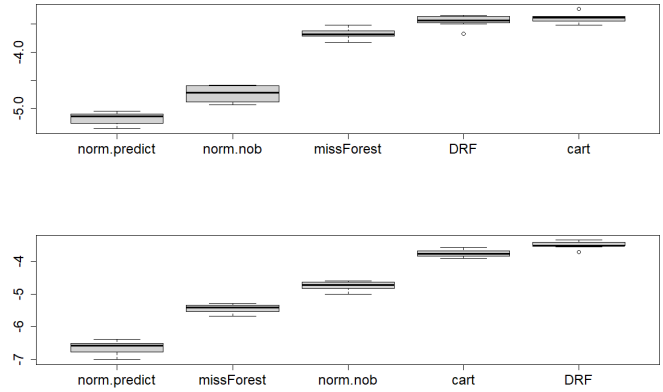


**Figure 6:** *Scores for the air quality data example. Top: DRF-Score over 10 iterations. Bottom: m-I-Score over 10 iterations.*

Protection Agency website (`https://aqs.epa.gov/aqsweb/airdata/download_files.html`). For a detailed description of the data set, we refer to Ćevid et al. (2022, Appendix C.1). The data set contains a total of 50'000 observations with 11 dimensions.

The goal of this example is to consider a real dataset with MAR missing values generated with an established procedure. We use the "ampute" function of the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011) to introduce MAR missingness into the first four numerical variables. The ampute function presents a flexible way of introducing missingness according to a desired mechanism, based on Rianne Margaretha Schouten and Vink (2018). We specify the 4 patterns

$$\begin{aligned} m_1 &= (1, 0, 0, 0, \ldots, 0) \\ m_2 &= (0, 1, 0, 0, \ldots, 0) \\ m_3 &= (0, 0, 1, 0, \ldots, 0) \\ m_4 &= (0, 0, 0, 1, \ldots, 0), \end{aligned}$$

and the ampute function to generate missingness according to these patterns.

The wealth of data allows us to redraw a data set of 2'000 observations $B = 10$ times to get an idea of the variation of our scores. That is, we redraw the data randomly $B$ times and generate the missingness mechanism using the ampute function. Figure 6 shows the drf- and $m$-I-Scores (obtained without using the true underlying values), while Figure 7 shows the negative energy distance between imputed and true data set. The ordering of the scores is remarkably similar, showing mice-cart and mice-DRF first and mice-norm.predict last. This makes sense as mice-norm.predict neither draws from the conditional distribution nor is it able to deal with the apparent nonlinearities in the data. In contrast, missForest scores higher, though interestingly the scores are not in complete agreement. While both the full information score and drf-I-Score put it in third place, the $m$-I-Score puts it just above mice-norm.predict. This might be due to the fact that missForest, while predicting instead of drawing from a conditional distribution, still models the nonlinearities in the data relatively well, a feat the Gaussian-based norm.nob cannot achieve. However, the $m$-I-Score punishes the inability of missForest to draw samples more severely and thus puts it lower than the other two scores. Given the discussion in this paper, one might argue that the low ordering of missForest of the $m$-I-Score is more accurate in this example.

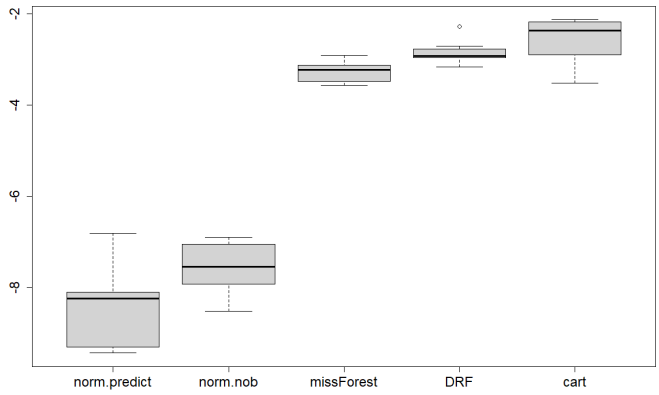


**Figure 7:** *Negative Energy Distance for the air quality data example, calculated with full data.*

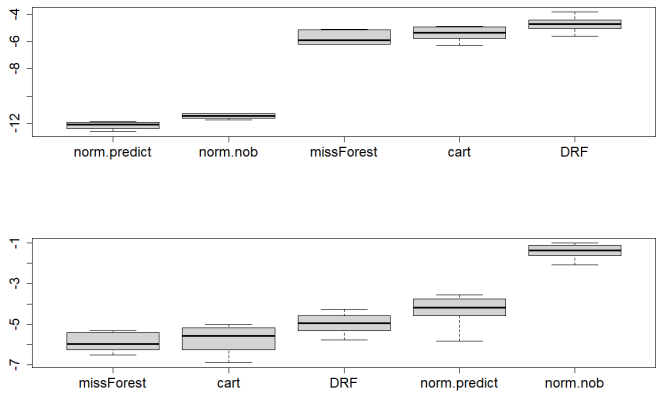


**Figure 8:** *Scores for the Gaussian mixture model with distribution shift. Top: DRF-Score over 10 iterations. Bottom: m-I-Score over 10 iterations.*

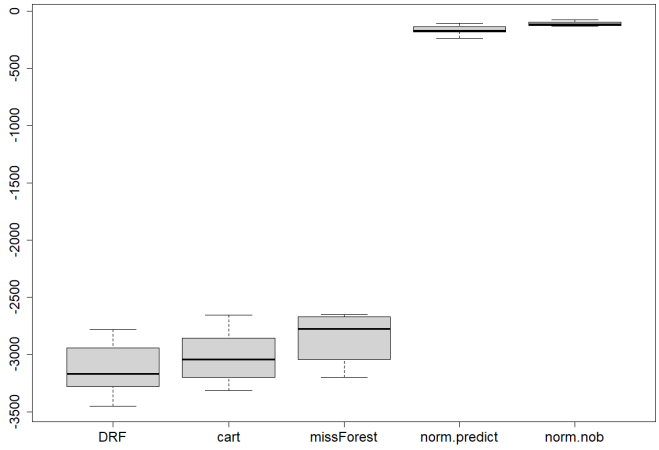


**Figure 9:** *Negative Energy Distance for the Gaussian mixture model with distribution shift, calculated with full data.*

## 5.2 Gaussian Mixture Model

We next turn to a Gaussian Mixture model to be able to put more emphasis on distribution shifts under MAR. In particular, we will simulate the distribution shift of Example 1 in a larger setting. We take $d = 6$ and 3 patterns,

$$m_1 = (1, 0, 0, 0, 0, 0)$$
$$m_2 = (0, 1, 0, 0, 0, 0)$$
$$m_3 = (0, 0, 1, 0, 0, 0)$$

The last three columns of fully observed variables are all drawn from three-dimensional Gaussians with randomly generated mean and covariance. For instance, for the first pattern, the mean (rounded) is given as $(3, 3, 4)$, while for the second it is given as $(-4, -3, -5)$. Thus each pattern can have quite different parameters. To preserve MAR, the (potentially unobserved) first three columns are built as

$$X_{O^c} = \mathbf{B} X_O + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix},$$

where $\mathbf{B}$ is a $3 \times 3$ matrix of coefficients, $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are independent standard Gaussian random errors and $O = \{4, 5, 6\}$ is again the index of fully observed values. This is a somewhat different example than the one before. Now the data is Gaussian with linear relationships, but there is a strong distribution shift between the different patterns. However, this distributional shift only stems from the observed variables, leaving the conditional distributions of missing given observed unchanged, as in Example 1. Consequently, it can be shown that the missingness mechanism meets (CIMAR) and is thus MAR.

In this example, the ability to generalize is important, while the ability to model nonlinear relationships is not. Indeed, we note that $P^*$ corresponds to the Gaussian imputation (mice-norm.nob) with the (unknown) true parameters. As such, a proper score should rank mice-norm.nob highest. In contrast, the forest-based scores should have the worst performance here, as they are not able to deal with the distribution shift. On the other hand, they might still be deemed better than

mice-norm.predict, which only imputes the regression prediction. Results for the drf- and $m$-I-Score are given in Figure 8, while Figure 9 shows the full information score. While the full information and $m$-I-Score behave as expected, with mice-norm.nob and mice-norm.predict in first and second place, and the forest-based methods last, the inability of DRF to meaningfully extrapolate beyond the sample points severely biases the drf-I-Score. Thus, it wrongly scores the forest-based methods highest. In contrast, despite the challenging setting, the $m$-I-Score still provides a very sensible ordering. An interesting difference between the $m$-I-Score and the full information score is that DRF and missForest are reversed in the two. However, this again makes sense as missForest gets more severely punished when it creates $N$ imputations with very limited variation. In fact, in this sense, the score, without having access to the true data, might actually give a more accurate picture of the correct ordering.

### 5.3 Mixture Model with Nonlinear Relationships

We now turn to a more complex version of the model in Section 5.2 to add nonlinear relationships to the distributional shifts. This final example should indicate that the search for successful imputation methods is by no means completed.

Using the same missingness pattern, and Gaussian variables $X_O$ we use a nonlinear function $f$ for the conditional distribution:

$$X_{O^c} = \mathbf{B} f(X_O) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix},$$

with

$$f(x_1, x_2, x_3) = (x_3 \sin(x_1 x_2), x_2 \cdot \mathbf{1}\{x_2 > 0\}, \arctan(x_1) \arctan(x_2)).$$

This introduces highly nonlinear relationships between the elements of $X_{O^c}$ and $X_O$, though the conditional distribution of $X_{O^c} \mid X_O$ is still Gaussian and the missingness mechanism is CIMAR. In this example, the ability to generalize is important, and so is the ability to model nonlinear relationships. Accordingly, this is a very difficult example and the ordering of the scores is quite different. In particular, they do not agree on the best two methods, though they all rank mice-DRF high. This serves to illustrate, that while at least the $m$-I-Score should be able to identify the "ideal" imputation, there is no guarantee for what happens when all imputations are bad. The disagreement of the scores should thus be seen as more of a testament that none of the methods perform well than a sign that the scores themselves are flawed.

## 6 Discussion

This paper attempted to give a more systematic discussion of MAR imputation. We analyse the MAR condition in detail for imputation and, based on this analysis, propose four essential properties an ideal imputation method should meet, as well as a principled way of ranking imputation methods.

An important message of the paper is that RMSE is not a sensible way of evaluating imputations. Dropping RMSE as an evaluation method likely has important implications. For instance, the recommendation of papers to use single imputation methods such as k-NN imputation (Anil Jadhav and Ramanathan, 2019) or missForest (Waljee et al., 2013; Tang and Ishwaran, 2017) appears to rest entirely on the use of RMSE. Even well-designed paper benchmarking imputation methods such as Jäger et al. (2021) use RMSE. Nonetheless, there appear to be only a handful of recent papers that at least consider different evaluation methods, for instance, Muzellec et al. (2020); Hong and Lynn (2020); Wang et al. (2022). Indeed, the problems of RMSE and its recommendations appear to be being rediscovered in different fields. For instance, recently Hong and Lynn (2020) again demonstrated empirically that, while missForest achieves the smallest RMSE, parameters attained

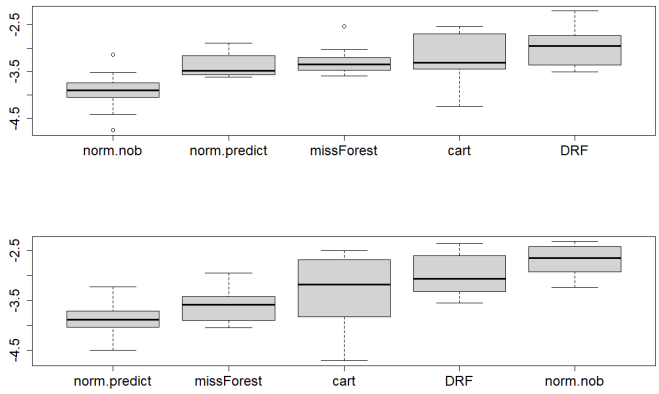


**Figure 10:** *Scores for the nonlinear mixture model with distribution shift. Top: DRF-Score over 10 iterations. Bottom: m-I-Score over 10 iterations.*

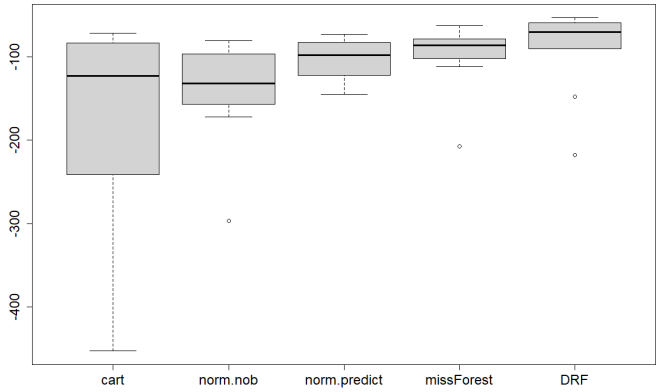


**Figure 11:** *Negative Energy Distance for the nonlinear mixture model with distribution shift, calculated with full data.*

from linear regression are severely biased. Similarly, Wang et al. (2022) discusses some problems with using RMSE in the machine learning literature. In contrast, GAN-based approaches recognize the objective of drawing imputations from the respective conditional distributions and naturally use the pattern-mixture modeling approach. However, despite having the right objective, these papers again use RMSE to compare the imputation quality of their method to competitors.

A second important message is that the problem of imputation is by no means solved. Though there is a set of promising imputation methods with mice-cart, mice-RF, and mice-DRF that will likely work well in a wide range of settings, there is room for improvement, especially concerning the ability to deal with covariate shifts. In particular, Section 5.3 shows an example with distribution shifts and nonlinear relationships for which all methods fail. The $m$-I-Score developed here can help to identify the right distribution, though this might not be helpful if all imputations are sufficiently bad. Appendix B demonstrates that modern joint modeling approaches do not fare better in this example. In fact, contrary to what we theorized in Section 3, on these low-dimensional data sets GAIN and MIWAE are outperformed by missForest, even in terms of energy distance.

We believe the paper touches on a few crucial issues that have not been discussed before. However, it also has several shortcomings. The $m$-I-Score, while promising, needs a set of fully observed variables, at least theoretically. In addition, the performance of mice-cart stands out, even when compared to mice-DRF. It remains an open question why the performance of mice-cart is so strong and whether a systematic benchmarking of imputation methods over a wider array of data sets can confirm the empirical findings in this paper. In general, a much more comprehensive empirical evaluation of both the new score and the forest-based imputation methods is needed. Finally, when talking about multiple imputation, we note that none of the studied nonparametric methods is able to include *model uncertainty*. However this would technically be needed for correct uncertainty quantification with multiple imputation, see e.g., Murray (2018). Though both mice-rf of Doove et al. (2014) and the new mice-DRF attempt to account for model uncertainty using several trees, this is only a heuristic solution. Moreover, the scores developed in this paper are unable to account for this and will instead likely place methods that include model uncertainty lower than those that do not, which in turn could explain the success of mice-cart in terms of these scores.

Finally, we discussed some challenging MAR conditions, particularly using the Gaussian Mixture Model. However, we did not discuss how *likely* such MAR settings may be. Intuitively, it appears that distributional shifts under MAR should be quite common. Consider an example with two variables, $X_1$ being income, and $X_2$ being age. Moreover, assume a missing mechanism for the income $X_1$, whereby $X_1$ tends to be missing whenever age is "high". Thus the probability of income ($X_1$) being missing depends entirely on the value of age ($X_2$), which is always observed. This is a textbook MAR example with two patterns, one where both variables are fully observed ($m_1$) and a second ($m_2$), wherein $X_1$ is missing. Despite the simplicity of this example, 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. In pattern $m_2$, where income is missing, values of both the observed age and the (unobserved) income tend to be higher. It thus appears intuitive that the combination of distributional shifts and nonlinear relationships is widespread in real data. At the same time, the success of forest-based methods such as missForest and mice-cart in benchmark papers suggests that current ways of introducing MAR might not produce enough distribution shifts in general. For instance, Näf et al. (2023) analyzed a range of data sets using the standard MAR mechanism of the ampute function implementing the procedure of Rianne Margaretha Schouten and Vink (2018), as we did in Section 5.1. Though their score is not proper under MAR, as shown in Appendix C.1, their analysis also showed mice-cart consistently in first place. Thus, tweaking the approach of Rianne Margaretha Schouten and Vink (2018) to produce MAR data with distribution shifts, might be an avenue for further research.

# A Score Version without Fully Observed Data

We first informally discuss the problems that arise when instead of the set of fully observed variables, we use all remaining variables $X_{-j} \sim H_{-j}$ in the score defined in Section 4. We note that, while the distribution of the observed test points $X_j$ is fixed to $p^*(x_j)$, it holds that

$$\begin{aligned} p^*(x_j) &= \int p^*(x_j \mid x_{-j}) p^*(x_{-j}) dx_{-j} \\ &= \int h(x_j \mid x_{-j}) h(x_{-j}) dx_{-j}. \end{aligned}$$

If $H_{-j}$ is different from $P^*_{-j}$, then the two conditional distribution will in general be different as well. This means with $X_{-j} \sim H_{-j}$ it is not clear, whether $X_j \mid X_{-j}$ has the desired distribution $(p^*(x_j \mid x_{-j}))$. This problem is numerically evident when the original imputation is used for $H_{-j}$. For instance, simply adapting the $m$-I-Score by using $X_{-j}$ instead of $X_O$ tends to score the regression imputation higher than the Gaussian imputation in the example in Section 5.2. To alleviate this, we instead generate $X_{-j} \sim H_{-j}$ independently, i.e., we impute the $d-1$ dimensional dataset without $X_j$ and keep the original imputation of $X_j$. That is, the only difference to the score of Section 4 is that, given an imputed data set and the imputation function itself, we first generate a new draw $X_{1,-j}, \ldots, X_{n,-j}$ from $H_{-j}$ by imputing all variables except $X_j$. Then we proceed as before: For $X_j$ we keep the imputed points, that is all $x_{i,j}$ with $m_i \in L_j^c$ are still drawn from the original imputation, while we set the *observed* observations of $X_j$ to missing, i.e. $x_{i,j} = \texttt{NA}$ for $i$ with $m_i \in L_j$. Then we concatenate

$$\begin{pmatrix} \texttt{NA} & (x_{i,-j})_{m_i \in L_j} \\ (x_{i,j})_{m_i \in L_j^c} & (x_{i,-j})_{m_i \in L_j^c}, \end{pmatrix}$$

and approximate the sampling from $H_{X_j \mid x_{i,-j}}$ in two ways:

(1) Regress $(x_{i,j})_{m_i \in L_j^c}$ onto $(x_{i,-j})_{m_i \in L_j^c}$ using DRF. Then for each test point $x_{i,-j}$, $m_i \in L_j$, sample $N$ times from the estimated conditional distribution obtained from DRF.

(2) Impute the NA values with $H$, $N$ times.

The $j$th score is then given as in (4.4), but now with the $N$ points generated approximately from $H_{X_j \mid X_{-j}}$. The idea in both cases is to use an imputation of $X_{-j}$ and the initial imputation of $X_j$ to generate a sample from the distribution $H_{X_j \mid X_{-j}}$ for points that are already observed. This approximation is clearly not perfect to obtain a sample from $H_{X_j \mid X_{-j}}$ but repeating the experiments in Section 5, Figures 12–14 show that the two scores closely follow their counterparts in Section 5.

**Remark.** *We note that compared to the score in Section 4, we now also need to obtain a sample from $H_{-j}$. Thus a $(d-1)$-variate data set has to be imputed for each $j$, which can be computationally challenging when $d$ is large. This could be solved by using random or predefined* projections *$A$, as in Näf et al. (2023), thus reducing the dimensionality. This will not hurt propriety but might diminish the power to detect differences between the methods. In fact, the score in Section 4 might be seen as an example of this with $A = O$.*

# B Comparison of MICE to GAIN and MIWAE

Here we use the negative energy distance advocated in the main text (i.e. the "full information score") to compare the performance of the MICE methods used in Section 5 to the joint modeling methods GAIN and MIWAE. The code for GAIN was taken from the original Github repository

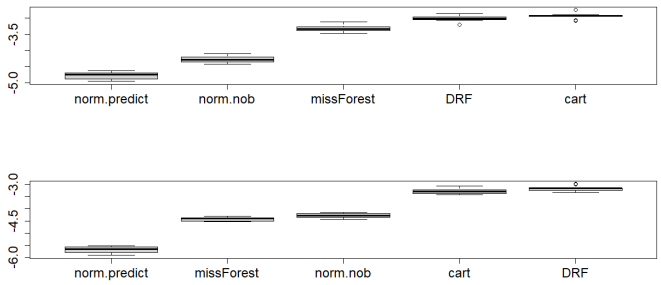


**Figure 12:** *Scores for the air quality data example. Top: DRF-Score over 10 iterations. Bottom: m-I-Score over 10 iterations.*

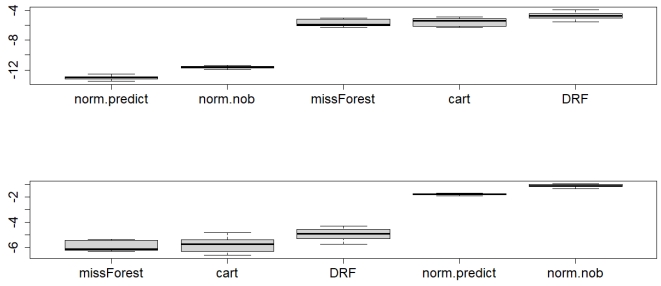


**Figure 13:** *Scores for the Gaussian mixture model with distribution shift. Top: DRF-Score over 10 iterations. Bottom: m-I-Score over 10 iterations.*

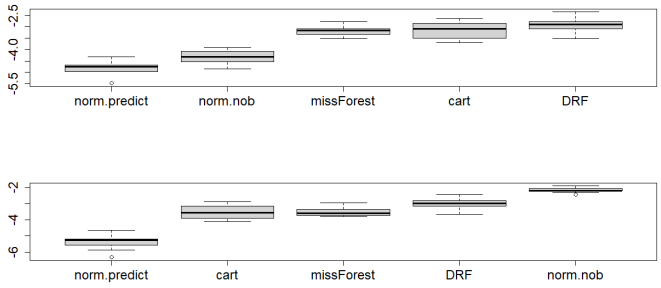


**Figure 14:** *Scores for the nonlinear mixture model with distribution shift. Top: DRF-Score over 10 iterations. Bottom: m-I-Score over 10 iterations.*

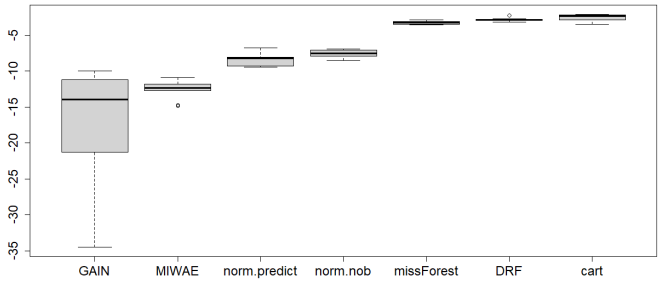


**Figure 15:** *Negative Energy Distance for the air quality data example with GAIN and MIWAE, calculated with full data.*

`https://github.com/jsyoon0823/GAIN`, while the implementation of MIWAE was obtained from `https://github.com/nbip/notMIWAE/blob/master/MIWAE.py`. As both were coded in Python, the R package `reticulate` (Ushey et al., 2024) was used to embed the code into R.

Figures 15 – 17 show the results. Overall these two methods cannot compete with MICE and usually are scored last, except in the Gaussian example with distribution shift (Figure 16) where MIWAE performs about the same as mice-cart and mice-DRF. However, we gave MIWAE a somewhat unfair advantage: We standardized the data in Application 1, as otherwise the implementation broke down, but did not do this for Applications 2 and 3. In practice, one would likely always standardize the data, given the numerical problems one faces otherwise, and this would have led to a lower ranking of MIWAE. Interestingly, this experiment does not confirm our suspicion in Section 3; GAIN and MIWAE tend to perform worse than missForest, even in terms of the energy distance. To analyze this further, we additionally consider a larger data set, the spambase data set of Lichman (2013), with more missing values. Specifically, we consider a simple MCAR mechanism whereby each variable is missing randomly such that we have around 20% of missingness in total. This dataset has dimension $d = 57$ and $n = 4601$ observations and was used to show that GAIN performs better than other imputation methods in Yoon et al. (2018). The combination of high frequency of missing values, and relatively high dimension and number of observations, means imputation with MICE takes considerably longer than in the two examples before. In particular, on a desktop computer, imputation times ranged from three minutes for mice-cart up to 29 minutes for missForest obtained from the `missForest` R-package (Stekhoven, 2022), mice-DRF needed 9 minutes, a computation time that can however be halved with around the same accuracy when defining blocks of size 2 and imputing once per block as described above. The result, shown in Figure 18, remains the same however, GAIN and MIWAE perform worse than even missForest in terms of energy distance, while missForest in turn is largely outperformed by mice-DRF and mice-cart.

All in all this small analysis provides a further hint that, at least for data sets of small or moderate dimensions, modern joint modeling methods such as GAIN and MIWAE cannot compete with FCS.

# C Proofs and Additional Results

In this section, we provide additional results and collect the proofs of the results not shown in the main paper. We start by showing that the score developed in Näf et al. (2023) is not proper under (PMM-MAR).

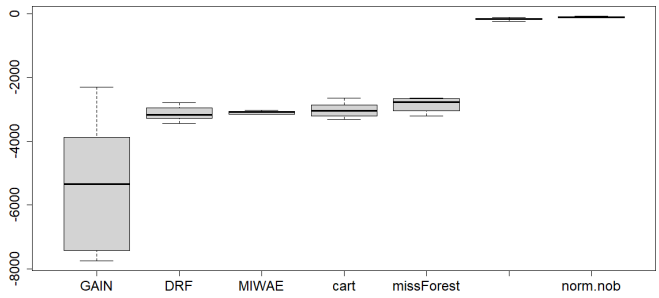


**Figure 16:** *Negative Energy Distance for the Gaussian mixture model with distribution shift with GAIN and MIWAE, calculated with full data.*

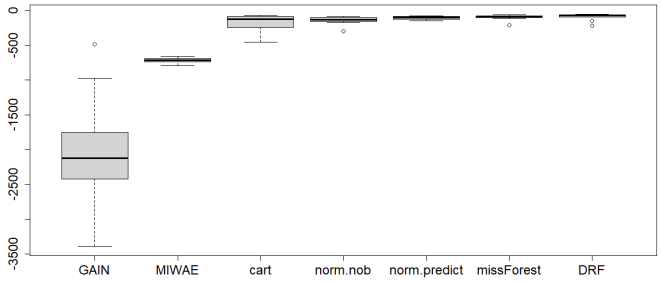


**Figure 17:** *Negative Energy Distance for the nonlinear mixture model with distribution shift with GAIN and MIWAE, calculated with full data.*

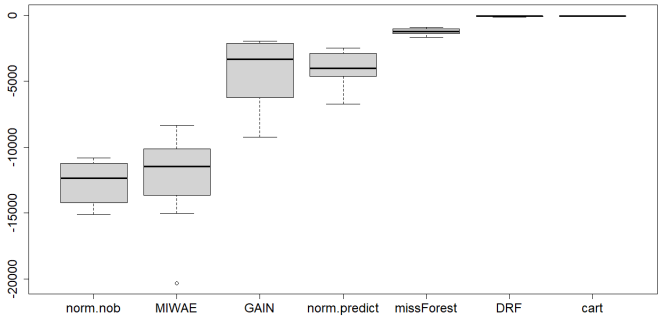


**Figure 18:** *Negative Energy Distance for the spam data example with GAIN and MIWAE, calculated with full data.*

## C.1 DR-I-Score is not proper under MAR

Here we show that the Density Ratio I-Score of Näf et al. (2023) is not proper under MAR. Define the Kullback-Leibler divergence (KL divergence) between two distributions $P, Q \in \mathcal{P}$ on $\mathbb{R}^d$ with densities $p$, $q$

$$D_{KL}(p \,||\, q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) d\mu(x).$$

Näf et al. (2023) developed a proper I-Score using the KL divergence estimated by a classifier in conjunction with random projections $A \subset \{1, \ldots, p\}$. The projections were done as a way to obtain more observations of each pattern. They proved that the population version of their score is a proper I-Score if condition (CIMAR) holds *each projection* $A$. Even without considering any projections, i.e. $A = \{1, \ldots, d\}$, this is a stronger condition than (PMM-MAR), as was shown above. In particular, in Example 2, their score will not be proper. Since the score is defined using a pattern-by-pattern comparison, when $H = P^*$ it will compare $p^*(x_1 \mid x_2, x_3)p^*(x_2, x_3)$ (third pattern) to

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

in the second pattern. Thus, while we would like to score the imputation $p^*(x_1 \mid x_2, x_3)$ highest, imputing by $h(x_1 \mid x_2, x_3) = x_1 p^*(x_1 \mid x_2, x_3)$ will lead to a score value of exactly zero, while

$$D_{KL}(p^* \,||\, p^*) = \int p^*(x_1, x_2, x_3) \log\left(\frac{1}{x_1}\right) d\mu(x_1, x_2, x_3) > 0.$$

Thus we have just shown that

**Proposition C.1.** *The I-Score defined in Näf et al. (2023) is not proper if* (PMM-MAR) *holds, but not* (CIMAR).

## C.2 Proofs

**Corollary 2.1.** *Condition* (SM-MAR) *is equivalent to* (SM-MAR II) *and both are equivalent to* (PMM-MAR).

*Proof.* We start by reformulating (SM-MAR), for any $x, \tilde{x}$ such that $o(x, m) = o(\tilde{x}, m)$,

$$\begin{aligned}
&\mathbb{P}(M = m|x) = \mathbb{P}(M = m|\tilde{x}) \Leftrightarrow \\
&\frac{p^*(x|M = m)\mathbb{P}(M = m)}{p^*(x)} = \frac{p^*(\tilde{x}|M = m)\mathbb{P}(M = m)}{p^*(\tilde{x})} \Leftrightarrow \\
&\frac{p^*(o(x, m), o^c(x, m) \mid M = m)}{p^*(o(\tilde{x}, m), o^c(\tilde{x}, m) \mid M = m)} = \frac{p^*(o(x, m), o^c(x, m))}{p^*(o(\tilde{x}, m), o^c(\tilde{x}, m))} \Leftrightarrow \\
&\frac{p^*(o^c(x, m) \mid o(x, m), M = m)}{p^*(o^c(x, m) \mid o(x, m))} = \frac{p^*(o^c(\tilde{x}, m) \mid o(x, m), M = m)}{p^*(o^c(\tilde{x}, m) \mid o(x, m))} \Leftrightarrow \\
&p^*(o^c(x, m) \mid o(x, m), M = m) = \frac{p^*(o^c(\tilde{x}, m) \mid o(x, m), M = m)}{p^*(o^c(\tilde{x}, m) \mid o(x, m))} p^*(o^c(x, m) \mid o(x, m)) \qquad \text{(C.1)}
\end{aligned}$$

Integrating (C.1) with respect to the missing part of $x$, $o^c(x, m)$, only shows that

$$\frac{p^*(o^c(\tilde{x}, m) \mid o(x, m), M = m)}{p^*(o^c(\tilde{x}, m) \mid o(x, m))} = 1,$$

and thus also (PMM-MAR). This shows that (SM-MAR) and (PMM-MAR) are equivalent. Molenberghs et al. (2008) show that (SM-MAR II) is also equivalent to (PMM-MAR), proving the result. □

**Proposition 2.3.** *Assume MAR in* (PMM-MAR) *holds. Then for* $h^*(x_j \mid x_{-j})$ *as in* (2.3),

$$h^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j}), \tag{2.4}$$

*for all* $x_{-j}$ *with* $p^*(x_{-j}) > 0$.

*Proof.* Let in the following $L_j$ be defined as in (2.2). We assume that $L_j$ is not empty. As all previous variables have been imputed and $x_j$ is observed, it is thus possible to identify the full distribution $p^*(x \mid M = m)$ for all $m \in L_j$. Thus, we learn the mixture of joint distributions

$$\begin{aligned} h^*(x_j, x_{-j}) &= \frac{1}{C} \sum_{m \in L_j} \mathbb{P}(M = m) \cdot p^*(x \mid M = m) \\ &= \frac{1}{C} \sum_{m \in L_j} \mathbb{P}(M = m \mid x) \cdot p^*(x), \end{aligned}$$

where $C$ is a constant such that $h^*(x_j, x_{-j})$ integrates to 1. Integrating $h^*(x_j, x_{-j})$ over $x_j$, we obtain similarly

$$h^*(x_{-j}) = \frac{1}{C} \sum_{m \in L_j} \mathbb{P}(M = m \mid x_{-j}) \cdot p^*(x_{-j})$$

Thus in fact:

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

It only remains to show that

$$\frac{\sum_{m \in L_j} \mathbb{P}(M = m \mid x)}{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_{-j})} = 1. \tag{C.2}$$

Indeed, we note that for any $m \in L_j^c$,

$$\mathbb{P}(M = m \mid x) = \mathbb{P}(M = m \mid x_{-j}),$$

by (SM-MAR). Consequently,

$$1 = \sum_{m \in L_j} \mathbb{P}(M = m \mid x) + \sum_{m \in L_j^c} \mathbb{P}(M = m \mid x_{-j}),$$

so that

$$\begin{aligned} \sum_{m \in L_j} \mathbb{P}(M = m \mid x) &= 1 - \sum_{m \in L_j^c} \mathbb{P}(M = m \mid x_{-j}) \\ &= \sum_{m \in L_j} \mathbb{P}(M = m \mid x_{-j}), \end{aligned}$$

and thus (C.2) indeed holds. □

**Corollary 2.2.** *Assume MAR in* (PMM-MAR) *holds and that $O$ in* (2.5) *is not empty. Then $H^* \in \mathcal{H}_P$ has*

$$h^*(x) = p^*(x), \text{ for all } x. \tag{2.6}$$

*Proof.* By construction, $H^* \in \mathcal{H}_P$. Assume that we are at step $j \in \{1, \dots, p\}$ of our imputation. That is, all variables $x_{i,l}$, $i = 1, \dots, n$, $l > j$ have successfully be imputed with a draw from $p^*(x_l \mid x_{l+1}, \dots, x_p, M = m)$. Let in the following $L_j$ be defined as in (2.2). We first note that any pattern $m \in L_j^c$ ( where $x_j \in o^c(x, m)$) has

$$p^*(o^c(x,m) \mid o(x,m), M = m) = p^*(o^c(x,m) \mid o(x,m)), \tag{C.3}$$

by (PMM-MAR). Integrating both sides, this means that for any $A \subset \{x_1, \dots, x_p\}$,

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

Thus the correct imputation distribution for this pattern $m$ is given by

$$\begin{aligned}
&p^*(x_j \mid x_{j+1}, \dots, x_p, M = m) \\
&= p^*(x_j \mid (\{x_{j+1}, \dots, x_p\} \cap o^c(x,m)) \cup (\{x_{j+1}, \dots, x_p\} \cap o(x,m)), M = m) \\
&= \frac{p^*(\{x_j, \dots, x_p\} \cap o^c(x,m) \mid \{x_{j+1}, \dots, x_p\} \cap o(x,m), M = m)}{p^*(\{x_{j+1}, \dots, x_p\} \cap o^c(x,m) \mid \{x_{j+1}, \dots, x_p\} \cap o(x,m), M = m)} \\
&= \frac{p^*(\{x_j, \dots, x_p\} \cap o^c(x,m) \mid \{x_{j+1}, \dots, x_p\} \cap o(x,m))}{p^*(\{x_{j+1}, \dots, x_p\} \cap o^c(x,m) \mid \{x_{j+1}, \dots, x_p\} \cap o(x,m))} \\
&= p^*(x_j \mid x_{j+1}, \dots, x_p).
\end{aligned}$$

Thus we need to learn $p^*(x_j \mid x_{j+1}, \dots, x_p)$ to successfully impute all patterns $m$ where $x_j$ is not observed. We assume that $L_j$ is not empty for any $j$. As all previous variables have been imputed and $x_j$ is observed, it is thus possible to learn the full distribution $p^*(x_j, x_{j+1}, \dots, x_p \mid M = m)$ for all $m \in L_j$. With the same arguments as in the proof of Proposition 2.3, we then obtain that

$$h^*(x_j \mid x_{j+1}, \dots, x_p) = p^*(x_j \mid x_{j+1}, \dots, x_p)$$

Thus we have shown that the learned (imputation) distribution is indeed the correct one. It then also holds that

$$\begin{aligned}
h^*(x) &= \sum_{m \in \mathcal{M}} \mathbb{P}(M = m) h^*(o^c(x,m) \mid o(x,m), M = m) h^*(o(x,m) \mid M = m) \\
&= \sum_{m \in \mathcal{M}} \mathbb{P}(M = m) p^*(o^c(x,m) \mid o(x,m), M = m) p^*(o(x,m) \mid M = m) \\
&= p^*(x),
\end{aligned}$$

whereby $h^*(o(x,m) \mid M = m) = p^*(o(x,m) \mid M = m)$ by assumption and $h^*(o^c(x,m) \mid o(x,m), M = m) = p^*(o^c(x,m) \mid o(x,m), M = m)$ as shown above. □

**Proposition 4.1.** *Assume MAR in* (PMM-MAR) *holds and that $O$ is not empty. Then $S_{NA}^{es}(H, P)$ is a proper I-Score.*

*Proof.* We show that for each $j$,

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

holds. Indeed, by propriety of the energy score $\mathbb{E}[es(H_{X_j|x_O}, Y)] \leq \mathbb{E}[es(H^*_{X_j|x_O}, Y)]$, when $Y \sim H^*_{X_j|x_O}$. Taking expectations on both sides shows that

$$S^{es}_{NA}(H, P) = \mathbb{E}[\mathbb{E}[es(H_{X_j|X_O}, Y)]] \leq \mathbb{E}[\mathbb{E}[es(H^*_{X_j|X_O}, Y)]]. \tag{C.4}$$

Moreover, similar to Proposition 2.3, it can be shown that $\mathbb{E}[es(H^*_{X_j|x_O}, Y)] = \mathbb{E}[es(P^*_{X_j|x_O}, Y)]$. We repeat the argument here for completeness: First

$$\begin{aligned} h^*(x_j \mid x_O) &= \frac{h^*(x_j, x_O)}{h^*(x_O)} \\ &= \frac{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_j, x_O) \cdot p^*(x_j, x_O)}{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_O) \cdot p^*(x_O)} \\ &= p^*(x_j \mid x_O) \frac{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_j, x_O)}{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_O)}. \end{aligned}$$

It only remains to show that

$$\frac{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_j, x_O)}{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_O)} = 1. \tag{C.5}$$

Indeed, we note that for any $m \in L^c_j$,

$$\mathbb{P}(M = m \mid x_j, x_O) = \mathbb{P}(M = m \mid x_O),$$

by (SM-MAR). Consequently,

$$1 = \sum_{m \in L_j} \mathbb{P}(M = m \mid x_j, x_O) + \sum_{m \in L^c_j} \mathbb{P}(M = m \mid x_O),$$

so that

$$\begin{aligned} \sum_{m \in L_j} \mathbb{P}(M = m \mid x_j, x_O) &= 1 - \sum_{m \in L^c_j} \mathbb{P}(M = m \mid x_O) \\ &= \sum_{m \in L_j} \mathbb{P}(M = m \mid x_O). \end{aligned}$$

It follows that,

$$\mathbb{E}[\mathbb{E}[es(H^*_{X_j|X_O}, Y)]] = \mathbb{E}[\mathbb{E}[es(P^*_{X_j|X_O}, Y)]] = S^{es}_{NA}(P^*, P). \tag{C.6}$$

Combining (C.4) and (C.6) gives the result. □

# References


Anil Jadhav, D. P. and Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933.

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2018). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18(196):1–39.

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.

Dong, W., Fong, D. Y. T., Yoon, J.-s., Wan, E. Y. F., Bedford, L. E., Tang, E. H. M., and Lam, C. L. K. (2021). Generative adversarial networks for imputing missing data for big data clinical research. *BMC Medical Research Methodology*, 21(1):78.

Doove, L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

Doretti, M., Geneletti, S., and Stanghellini, E. (2018). Missing data: A unified taxonomy guided by conditional independence. *International Statistical Review*, 86(2):189–204.

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.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235.

Hong, S. and Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, 20(1):199.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):173–190.

Jäger, S., Allhorn, A., and Bießmann, F. (2021). A benchmark for data imputation methods. *Frontiers in Big Data*, 4.

Lee, M. C. and Mitra, R. (2016). Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. *Computational Statistics & Data Analysis*, 95:24–38.

Lichman, M. (2013). UCI machine learning repository.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.

Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, (1):155–173.

Mattei, P.-A. and Frellsen, J. (2019). MIWAE: Deep generative modelling and imputation of incomplete data sets. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4413–4423.

Mealli, F. and Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102(4):995–1000.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(2):371–388.

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142 – 159.

Muzellec, B., Josse, J., Boyer, C., and Cuturi, M. (2020). Missing data imputation using optimal transport. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7130–7140.

Nazábal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2020). Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107:107501.

Näf, J., Spohn, M.-L., Michel, L., and Meinshausen, N. (2023). Imputation scores. *The Annals of Applied Statistics*, 17(3):2452 – 2472.

Qiu, Y. L., Zheng, H., and Gevaert, O. (2020). Genomic data imputation with variational autoencoders. *GigaScience*, 9(8):giaa082.

Rianne Margaretha Schouten, P. L. and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930.

Rizzo, M. and Szekely, G. (2022). *energy: E-Statistics: Multivariate Inference via the Energy of Data*. R package version 1.7-11.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "Missing at Random"? *Statistical Science*, 28(2):257–268.

Stekhoven, D. J. (2022). *missForest: Nonparametric Missing Value Imputation using Random Forest*. R package version 1.5.

Stekhoven, D. J. and Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Székely, G. J. (2003). E-statistics: the energy of statistical samples. Technical Report 05, Bowling Green State University, Department of Mathematics and Statistics.

Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. *Stat Anal Data Min*, 10(6):363–377.

Tian, J. (2017). Recovering probability distributions from missing data. In *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 574–589.

Ushey, K., Allaire, J., and Tang, Y. (2024). *reticulate: Interface to 'Python'*. R package version 1.35.0, https://github.com/rstudio/reticulate.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*, 16(3):219–242.

van Buuren, S. (2018). *Flexible Imputation of Missing Data. Second Edition.* Chapman & Hall/CRC Press.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8):e002847.

Wang, Z., Akande, O., Poulos, J., and Li, F. (2022). Are deep learning models superior for missing data imputation in surveys? Evidence from an empirical comparison. *Survey Methodology*, 48(2).

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17(3):589–602.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698.

Yuan, Y., Shen, Y., Wang, J., Liu, Y., and Zhang, L. (2021). VAEM: a deep generative model for heterogeneous mixed type data. In *Advances in Neural Information Processing Systems 34*, pages 4044–4054.

Zhu, J. and Raghunathan, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511):1112–1124.