What Is a Good Imputation Under MAR Missingness?

Jeffrey Näf, Erwan Scornet, Julie Josse

What Is a Good Imputation Under MAR Missingness?

Jeffrey Näf[1], Erwan Scornet[2], Julie Josse[1]
[1]Inria, PreMeDICaL Team, University of Montpellier
[2]Sorbonne Universite and Universite Paris Cite,
CNRS, Laboratoire de Probabilites,
Statistique et Modelisation, F-75005 Paris, France

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. Building on this analysis, we propose three essential properties a successful 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 imputation method, denoted mice-DRF, that meets two out of the three criteria. We then discuss and refine ways to rank imputation methods, developing a powerful, easy-to-use scoring algorithm to rank missing value imputations.

Keywords: Nonparametric imputation, missing at random, pattern-mixture models, distributional prediction, proper scores

1 Introduction

In this paper, we study general-purpose (multiple) imputation of missing data sets: instead of imputing for a specific estimation goal or target, we focus on imputations that can be used afterward 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). Joint modeling approaches may be based on parametric distributions (Schafer, 1997), or on neural networks, such as generative adversarial network (GAN) (Yoon et al. (2018); Deng et al. (2022); Fang and Bao (2023)) or variational autoencoder (VAE)(Mattei and Frellsen (2019); Nazábal et al. (2020); Qiu et al. (2020); Yuan et al. (2021)). In contrast, in the FCS approach, imputation is done one variable at a time, based on conditional distributions (see, e.g., Ibrahim et al., 1999; Lee and Mitra, 2016; Xu et al., 2016; Murray, 2018). The most prominent example of FCS is the multiple imputation by chained equations (MICE) methodology (van Buuren and Groothuis-Oudshoorn, 2011). In this paper, we address three questions.

First, is imputation under MAR possible with the FCS approach? Formally, using the so-called pattern-mixture model (PMM, Little (1993)) view of MAR, we prove that the conditional distribution needed to impute a missing value is identifiable. Thus, imputation with the FCS

approach is feasible in principle. As we will show, this result is non-trivial in the non-parametric missing data framework, as we do not assume that the parameters of the missingness mechanism and the distribution of the data are distinct. In this context, we compare the MAR condition we consider 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 an ideal imputation method have? Despite the previous positive identification result, MAR imputation can be extremely challenging, notably when distributions of observed variables differ across missing patterns. Thus, we list three properties a successful imputation method should meet. In short, it should be a distributional regression method, able to capture non-linear dependencies between covariates, while being robust to *covariate shifts*. We discuss existing methods that meet some of these criteria and introduce a new method based on the Distributional Random Forest of Ćevid et al. (2022), denoted "mice-DRF".

Third, how can one choose the best imputation for a given data set? While being of primary importance, this question 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 "proper" imputation scores (I-Scores) to rank imputations. Following their argument, imputation is a distributional prediction task and needs to be evaluated as such. In particular, when comparing imputation methods, one should refrain from using measures such as the root mean squared error (RMSE), as already pointed out (van Buuren, 2018; Hong and Lynn, 2020; Näf et al., 2023). Indeed, measures like RMSE favor methods that impute conditional means, instead of draws from the conditional distribution. Hence, using RMSE as a validation criterion artificially strengthens the dependence between variables and leads to severe biases in parameter estimates and uncertainty quantification. Currently, imputation methods are largely benchmarked and evaluated based on measuring the RMSE between the imputed and the underlying true values, see e.g., Stekhoven and Bühlmann (2011); Waljee et al. (2013); Yoon et al. (2018); Bertsimas et al. (2018); Kokla et al. (2019); Anil Jadhav and Ramanathan (2019); Nazábal et al. (2020); Qiu et al. (2020); Jäger et al. (2021); Dong et al. (2021) and many others. Instead, we advocate to use a distributional metric or score (Gneiting and Raftery, 2007; Székely, 2003) between actual and imputed data sets when the true values are available. In the more realistic scenario when true values are not available, we propose to use a new I-Score, which is proper under weaker conditions than that of Näf et al. (2023) while being more computationally efficient and easier to implement.

The remainder of the article is organized as follows. In Section 2, we study different MAR conditions and imputations in more detail and present our identification results. We then use these insights to present recommendations for imputation methods, including three properties any ideal imputation method should meet in Section 3.1. Section 3.2 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 four empirical examples in Section 4. Code to replicate the experiments and to use the new scoring methodology can be found in `https://github.com/JeffNaef/MARimputation`.

1.1 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^* (complete observations) and $M \mid X^*$ (distribution of the missing pattern conditional on complete observations) 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 have been largely ignored. 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. An important exception is given in the recent work of Ren et al. (2023). They appear to be the first to discuss the same identification result under MAR. Our result was derived independently and in the context of a more general discussion of the MAR condition for imputation. Moreover, we also study a weaker condition than MAR. In contrast Ren et al. (2023) focus on binary data and the No-Self-Censoring condition, the latter being an MNAR situation that neither implies nor is implied by MAR (Ren et al., 2023).

As the paper views missingness through the lens of pattern-mixture models of Little (1993), other conceptually close papers are those based on the generative adversarial network approach: Both Deng et al. (2022) and 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. However, the identification results in Deng et al. (2022) and 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 approach to develop a score that is proper, in the sense that it provably ranks the best imputation method highest in a population setting, under a much weaker condition.

1.2 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 σ-finite measure μ. 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.

Finally, when talking about scoring imputations, we will take $\|\cdot\|_2$ to be the Euclidean distance on $\mathbb{R}^d$ and write expectations as $\mathbb{E}_{\substack{X\sim H\\ Y\sim P^*}}[\|X-Y\|_{\mathbb{R}^d}]$, to clarify over which distributions the expectation is taken.

2 Sequential Imputation under MAR

In this section, we present different definitions of MAR used in the literature. While the original definition of MAR (Rubin, 1976) is established under parametric assumptions, we consider here the general case of non-parametric distribution, as done in more recent literature (see references below). We first study the exact relations between all these definitions, as summarized in Figure 2. We then show that if the number of missing values of a pattern m is larger than one, learning the imputation distribution directly from other patterns is generally not possible under the common MAR definition (PMM-MAR). However, in Section 2.2, we prove that learning this imputation distribution is theoretically possible for PMM-MAR if one variable at a time is imputed, as is done in the FCS approach. We then also consider a specific MNAR setting.

2.1 MAR Definitions

In this section, we analyze several different MAR conditions. We present two different settings, the selection model (SM) and the pattern-mixture model (PMM), each one leading to a different set of MAR assumptions.

Selection Model In the *selection model* framework (Little, 1993), the joint distribution of X^* and M is factored as $p^*(x, M=m) = \mathbb{P}(M=m \mid x)p^*(x)$. In this setting, MAR is defined as follows.

Definition 2.1 (SM-MAR)**.** *The missingness mechanism is missing at random (MAR) if, for all $m \in \mathcal{M}$, and all $x, \tilde{x}$ such that $o(x,m) = o(\tilde{x},m)$, we have*

$$\mathbb{P}(M=m|x) = \mathbb{P}(M=m|\tilde{x}). \tag{SM-MAR}$$

SM-MAR is sometimes referred to as "Always Missing at Random" (see, e.g., Mealli and Rubin, 2015; Deng et al., 2022). A widely used alternative definition of MAR (see e.g., Molenberghs et al. (2008); Little et al. (2019)) is the following.

Definition 2.2 (SM-MAR II)**.** *The missingness mechanism is missing at random (MAR) if, for all $m \in \mathcal{M}, x \in \mathcal{X}$, we have*

$$\mathbb{P}(M=m|x) = \mathbb{P}(M=m|o(x,m)). \tag{SM-MAR II}$$

[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 $\mathbb{R}^{|\{j:m_j=0\}|}$.

Note that $o(x, m)$ is different for each m, and thus neither SM-MAR nor SM-MAR II are statements about conditional independence, as remarked in Mealli and Rubin (2015). Nonetheless, SM-MAR II is intuitive: for any value of m, the probability of occurrence of missing pattern m only depends on the observed part of x. We can verify that these two definitions are equivalent.

Corollary 2.1. *Condition SM-MAR is equivalent to SM-MAR II.*

Pattern Mixture Model We now turn to the *pattern-mixture model* (PMM) framework (Little, 1993), which is based on the following decomposition $p^*(x, M = m) = p^*(x \mid M = m)\mathbb{P}(M = m)$. The PMM view emphasizes that the conditional distribution of the complete vector $X^* \mid M = m$ may vary across $m \in \mathcal{M}$. Consequently, learning a distribution of a given pattern m based on another pattern m' may be challenging, as the two distributions may differ drastically. A typical example is the Gaussian pattern-mixture model, where $X^* \mid M = m \sim N(\mu_m, \Sigma_m)$, so that the distribution in each pattern might follow a different Gaussian distribution. In this setting, Molenberghs et al. (2008) proposed the following definition.

Definition 2.3. *The missingness mechanism is missing at random (MAR) if, for all $m \in \mathcal{M}, x \in \mathcal{X}$,*

$$p^*(o^c(x, m) \mid o(x, m), M = m) = p^*(o^c(x, m) \mid o(x, m)). \tag{PMM-MAR}$$

Thus the conditional distribution of the missing part given the observed part $o^c(X^*, m) \mid o(X^*, m)$ in pattern m is equal to the conditional distribution, when no information about the pattern is available.

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

A stronger, but more interpretable condition, than PMM-MAR is the conditionally independent MAR (CIMAR), introduced in Deng et al. (2022).

Definition 2.4. *The missingness mechanism is conditionally independent MAR (CIMAR) if, for all $m \in \mathcal{M}, m' \in \mathcal{M}, x \in \mathcal{X}$, we have*

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

Contrary to all previous assumptions, CIMAR is a conditional independence statement, namely that $o^c(X^*, m) \mid o(X^*, m)$ is independent of M: the distribution of $o^c(X^*, m) \mid o(X^*, m)$ remains the same, for all missing patterns $M = m'$. Thus, CIMAR allows learning the distribution of $o^c(X^*, m) \mid o(X^*, m)$ from any pattern m' (see Näf et al., 2023, for an application). It in turn is still weaker than MCAR however, which requires that for all $m \in \mathcal{M}, m' \in \mathcal{M}, x \in \mathcal{X}$,

$$p^*(x) = p^*(x \mid M = m) = p^*(x \mid M = m'). \tag{PMM-MCAR}$$

Clearly, PMM-MCAR implies CIMAR, which implies PMM-MAR. Example 1 below and Example 4 in Appendix B.4 also show respectively an example that is CIMAR but not PMM-MCAR, and one example that is PMM-MAR but not CIMAR. Thus it holds that:

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

Figure 1 illustrates these different conditions in a small example. Another important MAR condition is the extended MAR condition, introduced in Deng et al. (2022).

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

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.*

Definition 2.5. *The missingness mechanism is Extended Missing At Random (EMAR), if, for all* $x \in \mathcal{X}$*, for all* $m \in \mathcal{M}$*, for* $m' = m$ *and* $m' = 0$*, we have*

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

EMAR can be rewritten as $p^*(o^c(x,m) \mid o(x,m), M = m) = p^*(o^c(x,m) \mid o(x,m)), M = 0)$. Clearly, CIMAR implies EMAR and Example 4 in Appendix B.4 demonstrates that EMAR is stronger than PMM-MAR, while it is clearly weaker than CIMAR. Figure 2 summarizes the different implications between the above MAR assumptions.

In the presence of missing data, one may resort to imputation, to approximately recover a sample from P^*. To impute correctly, one needs to determine the distribution of $o^c(X^*, m) \mid o(X^*, m)$, which can be used to impute the missing components $o^c(X^*, m) \mid M = m$. Clearly under CIMAR $o^c(X^*, m) \mid o(X^*, m)$ can in principle be determined from any other pattern m', while under (EMAR) this is possible for $m' = 0$. On the other hand, it appears not immediately clear what is needed to identify the right conditional distribution under PMM-MAR. We formalize these insights in the next section.

2.2 Identifiability under MAR

A crucial property of all three MAR definitions (PMM-MAR, CIMAR, EMAR) presented in Section 2.2 is that

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

Thus to impute pattern m successfully, one needs to learn $p^*(o^c(x,m) \mid o(x,m))$. We are now able to summarize the three different MAR definitions in one result about the ability to identify $p^*(o^c(x,m) \mid o(x,m))$. We say $p^*(o^c(x,m) \mid o(x,m))$ is identifiable from a set $\mathcal{M}_0 \subset \mathcal{M}$ of missing patterns, if there exists $w_{m'}(o(x,m))$, with $\sum_{m' \in \mathcal{M}_0} w_{m'}(o(x,m)) = 1$, such that the mixture

$$h^*(o^c(x,m) \mid o(x,m)) = \sum_{m' \in \mathcal{M}_0} w_{m'}(o(x,m)) p^*(o^c(x,m) \mid o(x,m), M = m'), \tag{2.1}$$

satisfies $p^*(o^c(x,m) \mid o(x,m)) = h^*(o^c(x,m) \mid o(x,m))$. In particular, we say that $p^*(o^c(x,m) \mid o(x,m))$ is identifiable from a pattern $m' \in \mathcal{M}$, if $p^*(o^c(x,m) \mid o(x,m), M = m') = p^*(o^c(x,m) \mid o(x,m))$. Define in the following

$$L_m = \{m' \in \mathcal{M} : m'_j = 0 \text{ for all } j \text{ such that } m_j = 1\}.$$

Thus L_m is the set of patterns for which $o^c(x,m)$ is observed.

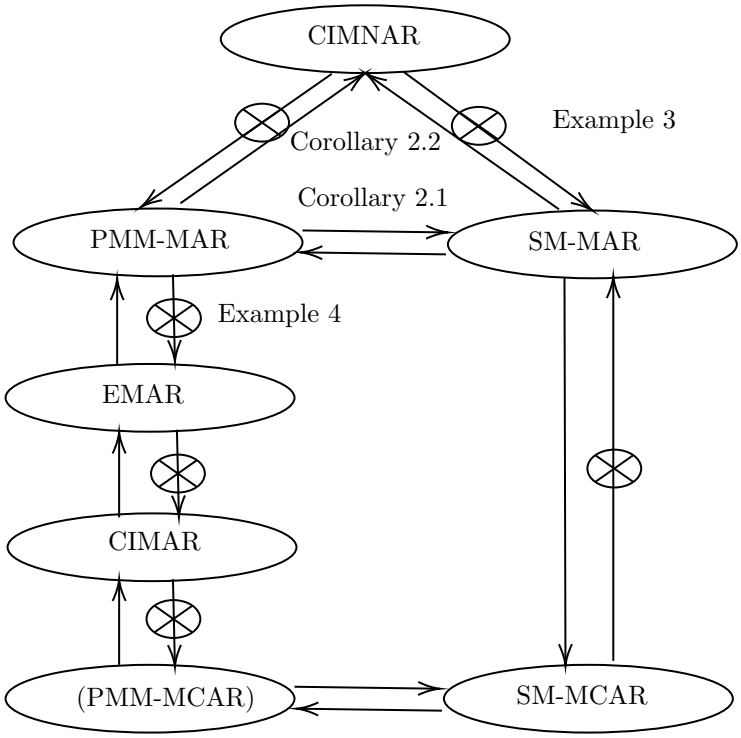

Figure 2: *Relationships between the M(N)AR conditions discussed in this paper. An arrow from condition A to condition B, encodes that A implies B. The definitions are given in Section 2.1 and 2.3.*

Proposition 2.3. *Assume* $|\mathcal{M}| > 3$. *Then for any pattern* $m \in \mathcal{M}$, $p^*(o^c(x,m) \mid o(x,m))$ *is*

- *identifiable from any other pattern* $m' \neq m$ *under CIMAR,*
- *identifiable from the pattern of fully observed data,* $m' = 0$, *under EMAR,*
- *is not identifiable from any single pattern* $m' \neq m$ *under PMM-MAR.*

In addition, if $\left|\sum_{j=1}^{d} m_j\right| > 1$, $p^*(o^c(x,m) \mid o(x,m))$ *is not identifiable from* L_m.

To appreciate Proposition 2.3, imagine for any pattern m' in which $o^c(x,m)$ is observed, X was already correctly imputed, such that $X^* \mid M = m'$ is available for all $m' \in L_m$. In this case, it would still not be possible to identify $p^*(o^c(x,m) \mid o(x,m))$ under PMM-MAR, as no mixture of the conditional distribution $p^*(o^c(X^*,m) \mid o^c(X^*,m), M = m')$ will recover the correct distribution. This is related to the fact that PMM-MAR still allows for a change in the conditional distributions over different patterns (see, e.g. Example 4 in Appendix B.4). Thus, the right conditional distributions are not trivially identified under MAR. On the contrary, under EMAR, we are able to impute $o^c(X^*,m)$ based only on the distribution of the complete input vector. However, such a practice may be difficult due to the low number of complete observations. At the price of the more stringent CIMAR assumption, we are able to build correct imputations by leveraging the information of all missing patterns.

Now, we show that the identification of the correct conditional distributions is possible, if one focusses *on one variable* X_j *at a time*. To this aim, let $L_j = \{m \in \mathcal{M} : m_j = 0\}$, be the set of patterns in which x_j is observed. Based on the missing patterns in L_j, one can build the

following mixture distribution,

$$h^*(x_j \mid x_{-j}) = \sum_{m \in L_j} \frac{p^*(x_{-j} \mid M = m)\mathbb{P}(M = m)}{\sum_{m \in L_j} p^*(x_{-j} \mid M = m)\mathbb{P}(M = m)} p^*(x_j \mid x_{-j}, M = m), \tag{2.2}$$

which is based *only on the missing patterns such that x_j is observed.* Note that the mixture (2.2) coincides with (2.1), if $o^c(x, m) = x_j$ and

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

Proposition 2.4. *Under PMM-MAR, the predictor h^* defined in* (2.2) *satisfies, for all $j \in \{1, \dots, d\}$, for all x_{-j} such that $p^*(x_{-j}) > 0$, and for all x_j,*

$$h^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j}). \tag{2.3}$$

Proposition 2.4 shows that the true distribution $p^*(x_j \mid x_{-j})$ is indeed computable 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. A similar result was already independently discussed in Ren et al. (2023), see Section 2.3. In Appendix B.1 we discuss the classical result of Rubin (1987) and highlight why this discussion of distribution shifts under MAR may not be relevant for (parametric) Maximum Likelihood Estimation (MLE).

Remark 2.1. *To illustrate these results, consider the following analogy. One could think of different missing value patterns m as different environments, such as different hospitals. Given several hospitals in which (X_j^*, X_{-j}^*) is observed, we would like to predict a variable X_j^* from (fully observed) covariates X_{-j}^* for a new hospital. Under CIMAR, covariates X_{-j}^* can arbitrarily change their distribution from one hospital to the next, but $X_j^* \mid X_{-j}^*$ remains the same in all hospitals, making it possible to learn the correct conditional distribution from any other hospital. Under PMM-MAR on the other hand, $X_j^* \mid X_{-j}^*$ may also differ from hospital to hospital. However, taking all hospitals such that (X_j^*, X_{-j}^*) is observed together, the distribution of $X_j^* \mid X_{-j}^*$ in the resulting mixture is the same as $X_j^* \mid X_{-j}^*$ needed for the new hospital.*

Link with the FCS approach Imputing one variable at a time is precisely the approach of FCS. 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 $t \geq 1$,

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

where $x_{-j}^{(t)} = \{x_l^{(t)}\}_{l \neq j}$ are the imputed and observed values of all other variables except j at the iteration t. Doing this repeatedly leads to a Gibbs sampler that converges under quite mild conditions (Little and Rubin (1986, Chapter 10.2.4.)). In practice, we do not have access to any true distribution $p^*(x_j \mid x_{-j}^{(t)})$ for $j \in \{1, \dots, d\}$, and sample instead from an estimated distribution $p_n(x_j \mid x_{-j}^{(t)})$. Proposition 2.4 shows that if we assume to have access to the true distribution $p^*(x_{-j})$, we can compute the true distribution $p^*(x_j \mid x_{-j})$ using only the observed values of X_j. As EMAR and CIMAR are both stronger than PMM-MAR, Proposition 2.4 also holds with PMM-MAR replaced with either one of those conditions. Thus, Proposition 2.4

shows that the FCS approach can identify the right conditional distributions under a weaker condition than GAN-based approaches in principle. Indeed, Deng et al. (2022) show that their GAN architecture is able to impute missingness under EMAR. 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.

Block-wise FCS The FCS approach of imputing one variable at a time has been criticized for computational reasons. Indeed, in FCS, d models have to be fitted repeatedly, which can be computationally intensive for large d. One way to remedy this would be to use multi-output methods such as DRF to impute variables as blocks, such that say 10 variables may be imputed simultaneously. The idea of using block-wise FCS was already discussed in van Buuren (2018, Chapter 4.7). However, Proposition 2.3 shows that one might not be able to recover the correct imputation distribution with this approach under PMM-MAR.

2.3 Identifiability beyond MAR

An important takeaway from the above discussion is that MAR is surprisingly weak. Indeed, it appears FCS imputation quickly breaks down when trying to weaken MAR, as we discuss here. So far, we have shown that MAR allows to identify the correct distribution for FCS imputation, assuming all other variables were already correctly imputed. However, it is not the weakest condition under which a similar argument is possible.

Definition 2.6. *The missingness mechanism is conditionally independent MNAR (CIMNAR) if, for all $j \in \{1, \dots, d\}$, for all $x \in \mathcal{X}$, we have*

$$\mathbb{P}(M_j = 1|x) = \mathbb{P}(M_j = 1|x_{-j}). \tag{CIMNAR}$$

Corollary 2.2. *Assumption SM-MAR implies Assumption CIMNAR, but Assumption CIMNAR does not imply Assumption SM-MAR.*

As shown in Corollary 2.2, CIMNAR is a strict generalization of MAR. Indeed, Example 3 in Appendix B.2 shows a setting in which CIMNAR holds, but SM-MAR does not. Closely related assumptions have already been introduced in the literature. For example, Beesley et al. (2021) suggested condition CIMNAR together with the conditional independence of the missingness indicators

$$M_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp M_d \mid X^*. \tag{2.4}$$

Taken together these two conditions do not generalize MAR, as (2.4) might not hold under SM-MAR. Condition CIMNAR also looks similar to the no self-censoring condition (NSC, Shpitser (2016)), where, for all j, $X_j^* \perp\!\!\!\perp M_j \mid X_{-j}^*, M_{-j}$. However, as demonstrated in Ren et al. (2023), neither MAR or NSC imply one another. In particular, using their adapted FCS approach for NSC data does not identify the right imputation distribution if the missingness mechanism is actually MAR. Let

$$\tilde{h}^*(x_j \mid x_{-j}) = \sum_{m \in L_j^c} \frac{p^*(x_{-j} \mid M = m)\mathbb{P}(M = m)}{\sum_{m \notin L_j} p^*(x_{-j} \mid M = m)\mathbb{P}(M = m)} p^*(x_j \mid x_{-j}, M = m), \tag{2.5}$$

be the equivalent of $h^*(x_j \mid x_{-j})$ but where we consider the patterns in L_j^c (the *unobserved patterns*) instead of L_j. In their brief discussion of FCS imputation under MAR, Ren et al. (2023) showed that MAR implies $h^*(x_j \mid x_{-j}) = \tilde{h}^*(x_j \mid x_{-j})$. This holds also true under CIMNAR:

Proposition 2.5. *Let $h^*(x_j \mid x_{-j}), \tilde{h}^*(x_j \mid x_{-j})$ be defined as in (2.2) and (2.5) respectively. Then, for all x_j, and for all x_{-j} with $p^*(x_{-j}) > 0$,*

$$h^*(x_j \mid x_{-j}) = \tilde{h}^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j}), \tag{2.6}$$

if and only if CIMNAR holds.

Under PMM-MAR, $h^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j})$ immediately implies that one can draw from the right distribution for imputation of X_j. However, PMM-MAR might not hold in general under CIMNAR, which implies that there exists $m \in L_j^c$, such that $p^*(x_j \mid x_{-j}, M = m)$ is not equal to $p^*(x_j \mid x_{-j})$. Thus, *it is no longer the right distribution to impute X_j in pattern $m \in L_j^c$*. This is true, even if we assume (2.4) in addition, hence the reason why Beesley et al. (2021) proposed adaptations to the FCS algorithm based on parametric assumptions. Nonetheless, Proposition 2.5 shows that one can correctly impute X_j from the mixture distribution $\tilde{h}^*(x_j \mid x_{-j})$, which is based on all patterns $m \in L_j^c$.

Unfortunately, Proposition 2.5 does not imply that all FCS imputation iterations are valid under CIMNAR. To see this, assume that we start with a variable X_k^* and impute it via the random variable Y_k, based on X_{-k}^*. For the next variable X_j, we assume to have access to $(Y_k, X_{-(j\cup k)}^*)$, where $X_{-(j\cup k)}^* = (X_l^*)_{l\neq j, l\neq k}$. Under MAR, the joint distribution of $(X_j, Y_k, X_{-(j\cup k)}^*)$ in all patterns $m \in L_j$ corresponds to the true distribution $p^*(x \mid M = m)$, so there is no issue for estimating $p^*(x_j \mid x_{-j})$. However, Example 3 in Appendix B.2 shows that the same is not true for CIMNAR in general. Despite this, it is sometimes possible to recover P^* under CIMNAR, as demonstrated by a modification of Example 3, as well as an empirical example in Section 4.3.

3 Implications for Imputation

In Section 2, we proved the identification of $p^*(x_j \mid x_{-j})$ for all j, based on all missing data patterns under SM-MAR. While this result is a first step to create methods that replicate the data distribution P^*, in practice it remains to estimate $p^*(x_j \mid x_{-j})$ in the FCS iterations. From now on, we will refer to any method that estimates $p^*(x_j \mid x_{-j})$ for all j, as an imputation method.

In this section, we present three essential properties that an ideal imputation method should satisfy. We then discuss which properties are met by classical FCS imputation strategies and introduce a new imputation approach denoted mice-DRF. Finally, we turn to the question of how to score imputation methods and develop our new m-I-Score.

3.1 Requirements for Imputation Methods

The goal of each iteration of an FCS algorithm is to estimate the conditional distributions $p^*(x_j \mid x_{-j})$ for all $j \in \{1, \ldots, d\}$. Thus, an imputation method is intrinsically a distributional estimator (Requirement (1) below). In order to accurately estimate the conditional distributions $p^*(x_j \mid x_{-j})$, an imputation method should be able to capture complex (potentially non-linear) interactions in the data (Requirement (2)). Finally, our identification result in Proposition 2.4 shows that $p^*(x_j \mid x_{-j})$ can be written as a mixture of all conditional distributions corresponding to patterns in which X_j is observed. Unfortunately, even under stringent assumptions such as CIMAR, distribution shifts in the observed variables may occur across different missing patterns, as highlighted in Example 1 below. Thus, an imputation method should be able to handle distributional shifts in the covariates (Requirement (3)).

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*. In this example, CIMAR holds. However, the distribution of* X_2^* *in pattern* m_1 *(*$N(0, 1)$*) is heavily shifted compared to pattern* m_2 *(*$N(5, 1)$*). Consequently, an estimation method that is able to accurately learn* $P^*_{X_1|X_2}$*, for* $X_2 \sim N(0, 1)$*, also needs to be able to extrapolate to* $P^*_{X_1|X_2}$*, for* $X_2 \sim N(5, 1)$*.*

Consequently, in the FCS framework, an imputation method should

(1) be a distributional regression method

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

(3) be able to deal with distributional shifts in the covariates.

Remark 3.1. *Murray (2018) (Section 4) also discusses best practices for general imputation methods. His points partly overlap with ours. In particular, he suggests that "imputations should reflect uncertainty about missing values" (corresponding to (1)) and "imputation models should be as flexible as possible" (corresponding to (2)). As we note in Section 5, he also emphasizes that for multiple imputation the uncertainty of the imputation model should be considered. This is not met by the imputation methods presented here and is an open problem for nonparametric imputation. While this becomes less consequential in large samples, this additional uncertainty is needed for reliable uncertainty quantification with multiple imputation.*

We now describe some of the state-of-the-art imputation methods, based on the benchmark analysis of papers by Waljee et al. (2013); Hong and Lynn (2020); Jäger et al. (2021); Wang et al. (2022); Näf et al. (2023).

We first consider two very simple methods, the *Gaussian* and *regression* imputations. Following the naming convention of the `mice` R package, we also denote the former as *mice-norm.nob* and the latter as *mice-norm.predict*. Both fit a linear regression of X_j onto X_{-j}. The Gaussian imputation then imputes X_j by drawing from a Gaussian distribution, while the regression imputation simply uses the conditional mean. Given that linear regression is known to extrapolate well, *mice-norm.predict* meets (3), while *mice-norm.nob* meets (1) and (3). A widely-used method in a variety of fields is the *missForest* of Stekhoven and Bühlmann (2011). In this method, X_j is regressed on X_{-j} with a Random Forest (RF, Breiman (2001)) and then imputed with the conditional mean in each iteration. As such, missForest, only meets (2). In contrast, *mice-cart* and *mice-RF* (Burgette and Reiter, 2010; Doove et al., 2014) use one or several trees respectively, but sample from the leaves to obtain the imputation of X_j, approximating draws from the conditional distribution. Thus these methods approximate (1), in addition to (2). As such, they may be inheriting the accuracy of missForest, while providing draws from the conditional distribution. However, they are ultimately not designed for the task of distributional regression. To this end, the distributional random forest (DRF) was recently introduced in Ćevid et al. (2022). We thus define a new imputation method, denoted

Method	(1)	(2)	(3)
missForest		✓	
mice-cart	✓	✓	
mice-RF	✓	✓	
mice-DRF	✓	✓	
mice-norm.nob	✓		✓
mice-norm.predict			✓

Table 1: *Properties (1)–(3) met by different imputation methods. Following the naming convention of the* `mice` *R package, "mice-norm.nob" refers to the Gaussian imputation, while "mice-norm.predict" refers to the regression imputation.*

mice-DRF, that regresses X_j onto X_{-j} and then imputes by sampling from the distributional regression estimator. As DRF is a forest-based distributional method, mice-DRF meets (1) and (2). However, as any local averaging estimate (kernel methods, tree-based methods, nearest neighbors), DRF still generalizes poorly outside of the training set (see, e.g., Malistov and Trushin, 2019), i.e. Requirement (3) is not met. Table 1 summarizes the properties met by the different mice methods considered in this paper.

Figure 3 illustrates the behavior of different imputation strategies for Example 1. As the Gaussian imputation fits a regression in pattern m_1 and then draws from a conditional Gaussian distribution given the estimated parameters, it is the ideal method in this setting and indeed captues the distribution very well. 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 suggests that forest-based methods such as mice-cart, mice-RF, and likely also mice-DRF may be some of the most successful methods currently available, finding an imputation method that satisfies (1)–(3) is still an open problem. In general, there are many more imputation methods that could be considered, including joint modeling approaches such as GAIN (Yoon et al., 2018). Thus the ability to rank imputation methods is crucial. This will be considered in the next section.

3.2 A Proper Scoring Method Under M(N)AR

We now turn to the question of how to evaluate the performances of different imputation strategies. Requirement (1) suggests that distributional distances or scores should be used instead of classic predictive metrics such as RMSE. Recall that P refers to the distribution of X with missing values and correspondingly, $P^* \in \mathcal{P}$ refers to the distribution of X^* without missing values.

First, we assume that the true underlying distribution P^* is known. In order to evaluate the performance of an imputation strategy that produces a distribution H, we compute the (negative) energy distance between imputed and real data:

$$\tilde{d}(H, P^*) = 2\mathbb{E}_{\substack{X\sim H\\ Y\sim P^*}}[\|X - Y\|_2] - \mathbb{E}_{\substack{X\sim H\\ X'\sim H}}[\|X - X'\|_2] - \mathbb{E}_{\substack{Y\sim P^*\\ Y'\sim P^*}}[\|Y - Y'\|_2],$$

where $\|\cdot\|_2$ is the Euclidean metric on $\mathbb{R}^d$. Given samples from P^* and a sample of imputed points this can be readily estimated, see e.g., Székely (2003).

Second, we are interested in creating 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).

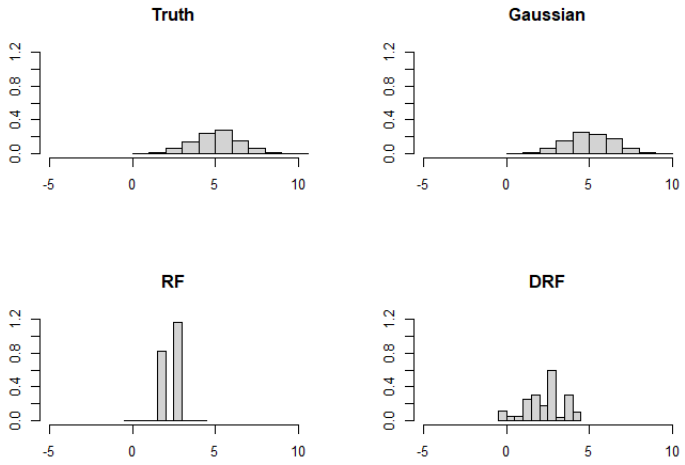

Figure 3: *The true distribution against a draw from different imputation procedures for imputing X_1 in Example 1.*

$$\mathbf{X} = \begin{pmatrix} \boxed{x_{1,1}} & \boxed{x_{1,2}} & \boxed{x_{1,3}} & \boxed{x_{1,4}} \\ NA & \boxed{x_{2,2}} & \boxed{x_{2,3}} & \boxed{x_{2,4}} \\ \boxed{x_{3,1}} & NA & x_{3,3} & x_{3,4} \\ \boxed{x_{1,4}} & NA & NA & x_{4,4} \end{pmatrix}$$

Figure 4: *Illustration of O_j, for $j = 1, 2$. For X_2, $X_{O_j} = (X_3, X_4)$ in gray, while for X_1, $X_{O_j} = X_4$ in black.*

Definition 3.1 (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 all $H \neq P^*$.

Our new scoring method for evaluating imputation performances requires that there exist $j, k \in \{1, \ldots, d\}$ such that $j \neq k$ and such that X_k is always observed when X_j is observed (see Figure 4 for an example).

Assumption 3.1. *There exists $j \in \{1, \ldots, d\}$ such that $O_j = \bigcap_{m \in L_j} \{l : m_l = 0\}$ is not empty and, for all k such that $O_k \neq \emptyset$, $X_k \perp\!\!\!\perp M_k \mid X_{O_k}$.*

Let in the following $H^*_{X_j | x_{O_j}}$ be the distribution with density

$$h^*(x_j \mid x_{O_j}) = \sum_{m \in L_j} \frac{p^*(x_{O_j} \mid M = m)\mathbb{P}(M = m)}{\sum_{m \in L_j} p^*(x_{O_j} \mid M = m)\mathbb{P}(M = m)} p^*(x_j \mid x_{O_j}, M = m), \tag{3.1}$$

and similarly, $H_{X_j|x_{O_j}}$ the distribution of the same mixture with p^* exchanged with the imputation distribution h. For all $j \in \{1, \dots, d\}$, we define the score of variable j as

$$S^j_{\mathrm{NA}}(H, P) =$$
$$\mathbb{E}_{X_{O_j} \sim P^*_{X_{O_j}|M \in L_j}} \Big[\mathbb{E}_{\substack{X \sim H_{X_j|X_{O_j}} \\ Y \sim H^*_{X_j|X_{O_j}}}} [\|X - Y\|_2] - \frac{1}{2} \mathbb{E}_{\substack{X \sim H_{X_j|X_{O_j}} \\ X' \sim H_{X_j|X_{O_j}}}} [\|X - X'\|_2] \Big], \tag{3.2}$$

where the outer expectation is taken over $X_{O_j} \sim P^*_{X_{O_j}|M \in L_j}$, the distribution of all fully observed variables wrt to variable j. This is the (expected) energy score, which is directly related to the energy distance above, see e.g., Gneiting and Raftery (2007); Gneiting et al. (2008).[2] Moreover, let

$$O = \{j : m_j = 0 \text{ for all } m \in \mathcal{M}\} \tag{3.3}$$

be the (potentially empty) set of all fully observed variables. Then, the full score is given as

$$S_{\mathrm{NA}}(H, P) = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} S^j_{\mathrm{NA}}(H, P),$$

where $\mathcal{S} = O^c \cap \{j : O_j \neq \emptyset\}$ and O^c is the set of variables with at least one missing value.

Proposition 3.1. *Under Assumption 3.1, S_{NA} is a proper I-Score.*

Näf et al. (2023) developed a first I-Score using density ratios and random projections $A \subset \{1, \dots, d\}$, denoted "DR-I-Score". The set of projections could be chosen by a practitioner as a tool to increase the number of fully observed observations. The score was shown to be proper under CIMAR on each projection. In Appendix B.3 we show that it is actually proper under EMAR (on each projection), but that it is not proper under MAR.

Compared to the DR-I-Score, this new score is not only easier to use but also proper under a modified CIMNAR condition. Here, projections are given by $A = O_j$, for each variable j we would like to test. Assumption 3.1 can thus be seen as "CIMNAR on each projection", which is considerably weaker than CIMAR on each projection. However, the projections can no longer be freely chosen but are instead given by the maximum number of observed variables available for a given variable j. We note that the construction of this score and the propriety result hinges on the discussion in Section 2. The I-Score of Näf et al. (2023) fails to be proper under MAR, even on the projected data, because it compares joint distributions of individual patterns m, m'. In contrast, $S_{\mathrm{NA}}(H, P)$ compares *conditional* distributions obtained from *mixture of patterns* that are ensured to be the same under the restricted CIMNAR condition (Assumption 3.1) for the perfect imputation.

3.3 Score Estimation

We describe now an estimation strategy for $S_{\mathrm{NA}}(H, P)$ based on a sample of n observations with missing values. Recall that $\mathcal{S} = O^c \cap \{j : O_j \neq \emptyset\}$. Fix $j \in \mathcal{S}$ and recall that L_j collects all patterns m, such that $m_j = 0$. For each i such that $m_i \in L_j$ (that is $x_{i,j}$ is observed), we build a sample of N points, $\tilde{X}^{(i)}_1, \dots, \tilde{X}^{(i)}_N$, as follows:

1. Create a new data set by concatenating the observed $(x_{i,j}, x_{i,O_j})$, $m_i \in L_j$ and the imputed $(x_{i,j}, x_{i,O_j})$, $m_i \in L^c_j$, as in Figure 5, and set the *observed* observations of X_j to missing, i.e. $x_{i,j} = \texttt{NA}$ for i with $m_i \in L_j$.

[2] 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.

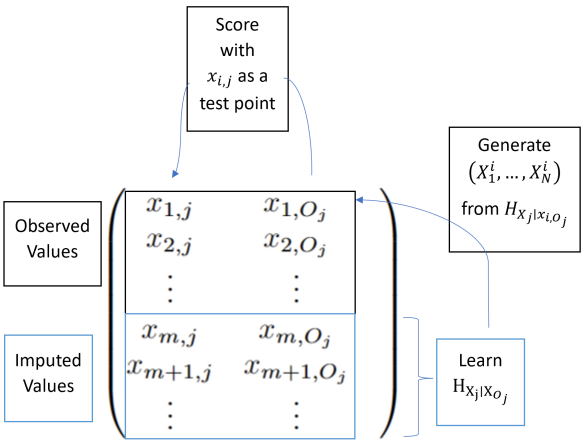

Figure 5: *Conceptual illustration of the score approximation. First, the imputed values in blue are used to learn $H_{X_j|X_{O_j}}$. Then, for each x_{i,O_j} for which $x_{i,j}$ is observed, we score the "prediction" $H_{X_j|X_{O_j}}$ using the energy score with test point $x_{i,j}$. In practice, this is done by (approximately) generating a sample $\tilde{X}_l^{(i)}, l = 1, \dots, N$ from $H_{X_j|x_{i,O_j}}$.*

2. Approximate the sampling from $H_{X_j|X_{O_j}}$ by simply imputing these artificially created NA values with H, N times.

As multiple imputation corresponds to drawing several times from the corresponding conditional distribution, this is a natural way of obtaining $\tilde{X}_l^{(i)}$, $l = 1, \dots, N$. If a method is not able to generate multiple imputations, $\tilde{X}_l^{(i)}$ is just a unique value copied N times, as $H_{X_j|x_{i,O_j}}$ is simply a point measure.

We can use the generated samples $\tilde{X}_1^{(i)}, \dots, \tilde{X}_N^{(i)}$, for all i such that $m_i \in L_j$, in order to estimate $S_{NA}^j(H, P)$:

$$\widehat{S}_{\text{NA}}^j(H, P) = \frac{1}{|\{i : m_i \in L_j\}|} \sum_{i:m_i \in L_j} \left(\frac{1}{2N^2} \sum_{l=1}^{N} \sum_{\ell=1}^{N} |\tilde{X}_l^{(i)} - \tilde{X}_\ell^{(i)}| - \frac{1}{N} \sum_{l=1}^{N} |\tilde{X}_l^{(i)} - x_{i,j}| \right), \tag{3.4}$$

as in Gneiting et al. (2008, Equation (7)). This is nothing more than the empirical counterpart of Equation (3.2). The final score is then given as

$$\widehat{S}_{\text{NA}}(H, P) = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \widehat{S}_{\text{NA}}^j(H, P). \tag{3.5}$$

Remark 3.2. *Formally, the observations in X_{O_j} are needed to ensure the test points $x_{i,j}$ are truly sampled from $h^*(x_j \mid x_{O_j})$, which in turn is equal to $p^*(x_j \mid x_{O_j})$ under Assumption 3.1, as shown in Proposition 3.1. 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 $p^*(x_j \mid x_{-j})$.*

We refer to this approach as the m-I-Score. The m-I-Score score thus uses the ability of imputation methods to generate multiple imputations naturally in its scoring. Unfortunately, this can be computationally demanding, as N should be chosen high, say at least 50 to give an accurate score. Moreover, if all variables contain missing values, this would need to be repeated

d times. This would be infeasible for realistic dimensions if the full data set had to be imputed each time. However note that in the data set created in steps 1. and 2. 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. Moreover, for large d it is possible to only consider a subset of variables X_j to calculate $\hat{S}_{NA}$. For instance, one could choose the $p < d$ X_j with the largest missingness proportion.

4 Empirical Study

The goal of this section is to illustrate the concepts discussed in this paper on both simulated and real data. We first describe the implementation of the new mice-DRF. The `mice` package provides a very convenient interface whereby new regression methods can be added to the MICE routine. We thus implement the following in MICE: for each variable j with missing values, 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 $x_{i,j}$ 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.

Imputation methods We empirically evaluate the performances of the following FCS methods

- mice-cart
- mice-DRF
- missForest
- regression imputation, named mice-norm.predict (see R-package `mice` (van Buuren and Groothuis-Oudshoorn, 2011))
- Gaussian imputation, named mice-norm.nob (see R-package `mice` (van Buuren and Groothuis-Oudshoorn, 2011))

We also compare two deep learning strategies (see Appendix A.1)

- GAIN (Yoon et al., 2018)
- MIWAE (Mattei and Frellsen, 2019).

All methods are used with their default hyperparameter values. In all examples, we standardize the scores over the 10 repetitions to lie in $(-1, 0)$.

Evaluation 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 m-I-Score, which does 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 $N = 50$. To illustrate the discussion in this paper, we also add the negative RMSE, which again uses the full data set. We also compare the m-I-Score to the DR-I-Score of Näf et al. (2023) in Appendix A.2.

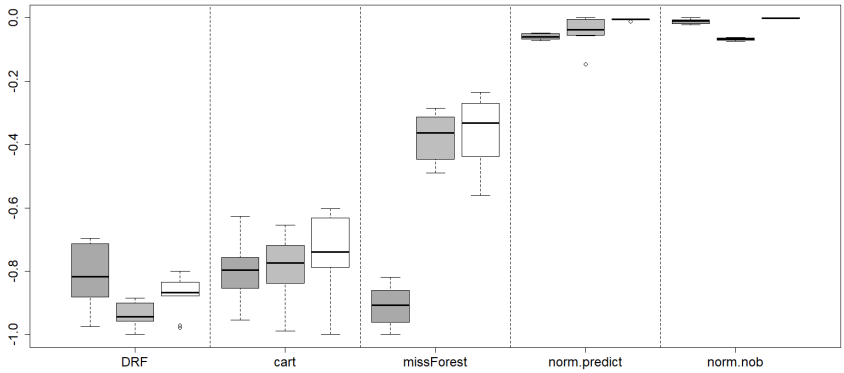

Figure 6: *Standardized scores for the Gaussian mixture model with distribution shift. Methods are ordered according to the full information score. For each method, the m-I-Score (dark gray, left), RMSE (gray, middle), and full information score (white, right) are shown.*

Results The three examples considered in this section, as well as the analysis in Appendix A.1, indicate that:

(I) The m-I-Score reliably finds the best imputation method and the ordering it produces is similar to the one of the full information score, even in the first challenging distributional shift example in Section 4.1. 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 4.2 where none of the methods perform well. Here the scores disagree quite heavily.

(II) mice-DRF and mice-cart are the most promising methods in terms of the full information score. This aligns with the findings in Wang et al. (2022); Näf et al. (2023). In particular, they tend to have higher scores than missForest, GAIN, and MIWAE, as shown in Appendix A.1.

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

4.1 Gaussian Mixture Model

We first turn to a Gaussian Mixture model to be able to put more emphasis on distribution shifts under MAR. In particular, we simulate the distribution shift of Example 1 in higher dimensions. 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, denoted X_O^*, are all drawn from three-dimensional Gaussians with means $(5, 5, 5)$, $(0, 0, 0)$ and $(-5, -5, -5)$ respectively, and a Toeplitz covariance matrix Σ with $\Sigma_{i,j} = 0.5^{|i-j|}$. Thus there are relatively strong mean shifts between the different patterns. 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 3×3 matrix of coefficients, $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are independent $N(0, 4)$ random errors and $O = \{4, 5, 6\}$ is the index of fully observed values. For $\mathbf{B}$ we copy the vector $(0.5, 1, 1.5)$ three times, such that $\mathbf{B}$ has identical rows. The data is thus 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. Moreover, Assumption 3.1 holds.

For each pattern, we generate 500 observations, resulting in $n = 1.500$ observations and around 17% of missing values. In this example, we expect that the imputation able to adapt to shift in covariates will perform well, even if they are not able to catch complex dependencies between variables. 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 may not able to deal properly 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 (standardize) full information score, m-I-Score and RMSE are given in Figure 6. The full information and m-I-Score behave as expected, with mice-norm.nob in first place, and the forest-based methods last. Interestingly, both score mice-norm.predict second. RMSE in turn, ranks mice-norm.predict as the best method. Thus, despite having access to the true underlying data, RMSE is not able to identify the method that best replicates P^*.

Appendix A.2 also shows that the DR-I-Score is not able to rank mice-norm.nob as the best method. This may be due to the difficulty of random forests to deal with covariate shifts, as the DR-I-Score implementation relies on an RF classifier to estimate the involved density ratios. This shows a clear advantage of our new score, as it does not rely on any auxiliary method to estimate $H_{X_j|X_{O_j}}$, but instead directly generates and judges samples from the imputation method.

Thus, 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 the m-I-Score scores missForest lower than the full information score. However, this makes sense as missForest gets more severely punished when it creates N imputations with very limited variation. In this sense, the m-I-Score, without having access to the true data, might actually give a more accurate picture of the correct ordering.

4.2 Mixture Model with Nonlinear Relationships

We now turn to a more complex version of the model in Section 4.1 to add nonlinear relationships to the distributional shifts. This example should indicate that the search for successful imputation methods is by no means complete.

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

$$X^*_{O^c} = f(X^*_O) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}, \tag{4.1}$$

with

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

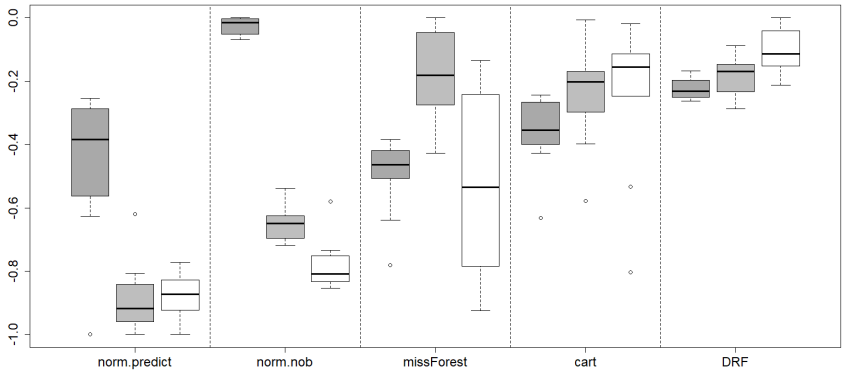

Figure 7: *Standardized scores for the nonlinear mixture model with distribution shift. Methods are ordered according to the full information score. For each method, the m-I-Score (dark gray, left), RMSE (gray, middle), and full information score (white, right) are shown.*

This introduces 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. Moverover, Assumption 3.1 is met here. For each pattern, we generate 500 observations, resulting in $n = 1'500$ observations and around 17% of missing values.

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. The ordering of the m-I-Score and the full information score shown in Figure 7 is quite different. In particular, they do not agree on the best two methods, though they both 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. Finally, the RMSE is surprisingly close to the ranking of the full information score based on the means over the 10 repetitions.

4.3 MNAR Example

We also consider an example of CIMNAR. In the previous two examples, the focus lay on the complex shifts that can occur under MAR. Here, we instead focus on a complex missingness mechanism. As such we simply take X^* to be a six-dimensional Gaussian with mean $(5, 5, 5, 5, 5, 5)$ and Toeplitz covariance matrix Σ, with $\Sigma_{ij} = 2 \cdot 0.5^{|i-j|}$. For $M \mid X^*$, we then choose:

$$\mathbb{P}(M_j = 1 \mid x) = \mathbb{P}(M_j = 1 \mid x_{-j}) = \frac{1}{1 + \exp(-1/4 \sum_{l \neq j} x_l)},$$

for $j = 1, \ldots, 3$. That is the probability of the first three variables being missing depends on all other variables, whether or not they are missing. We again take X_4, X_5, X_6 to be fully observed.

First, the mean (unstandardized) energy distance between the real data and the mice-norm.nob imputation is approximately 0.001, around the same as for the MAR example in Section 4.1. This surprisingly indicates that MICE imputation is able to recreate the distribution of this MNAR example quite closely. In fact, generating the same proportion of MCAR missing values for the same example leads to similar mean energy distance values. Figure 8

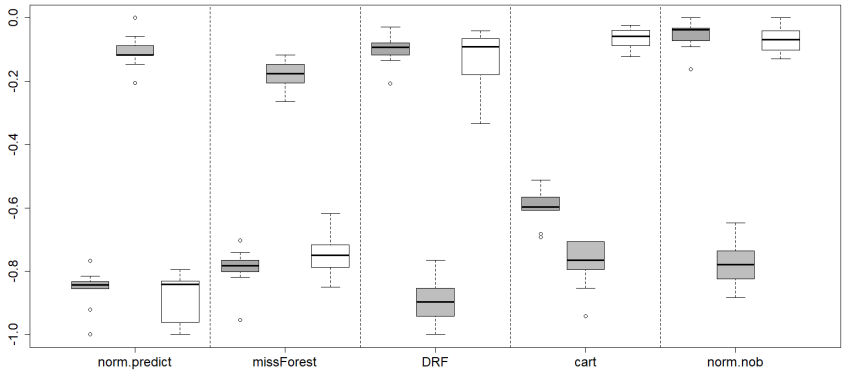

Figure 8: *Standardized scores for the MNAR example. Top: DR-I-Score over 10 iterations. Methods are ordered according to the full information score. For each method, the m-I-Score (dark gray, left), RMSE (gray, middle), and full information score (white, right) are shown.*

shows the standardized scores for this example. Despite the fact that Assumption 3.1 does not hold in this setting, the m-I-Score performs well, again correctly identifying the best imputation method to be mice-norm.nob. RMSE in turn again scores both mice-norm.predict as well as missForest higher than mice-norm.nob. In fact, the ordering of RMSE is almost the reverse of the ordering of the full information score.

4.4 Air Quality Data

We end 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 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 Schouten et al. (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 9 shows the (standardized) scores. The ordering of the m-I-Score and the full information score is again 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

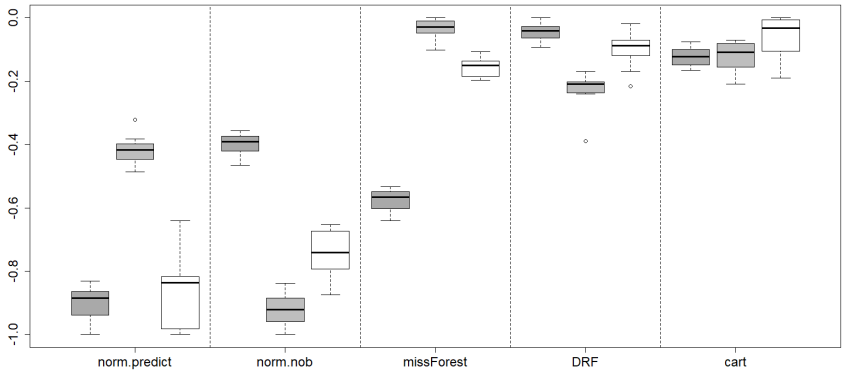

Figure 9: *Standardized scores for the air quality data example. Methods are ordered according to the full information score. For each method, the m-I-Score (dark gray, left), RMSE (gray, middle), and full information score (white, right) are shown.*

apparent nonlinearities in the data. In contrast, missForest scores higher, though interestingly the scores are not in agreement. While the full information score puts 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 again penalizes 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. RMSE on the other hand scores missForest highest, simply because it likely estimates the conditional means well. This mirrors previous analysis on real data with artificially generated missing values that found missForest to perform well (see e.g., Waljee et al. (2013); Hong and Lynn (2020); Jäger et al. (2021) among others) and indicates that this might be entirely due to the use of RMSE.

5 Discussion

This paper gives a more systematic discussion of FCS imputation. We analyse and generalize the MAR condition for imputation and, based on this analysis, propose three essential properties an ideal imputation method should meet, as well as a principled way of ranking imputation methods. We conclude with four important points.

RMSE should not be used 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 appear to be rediscovered in different fields. For instance, Hong and Lynn (2020) empirically emphasize that, while missForest achieves the smallest RMSE, parameter estimations in 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 (see, e.g., Yoon et al., 2018).

New imputation methods are needed 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, there is room for improvement, especially concerning the ability to deal with covariate shifts. In particular, Section 4.2 shows an example with distribution shifts and nonlinear relationships for which all methods fail. Appendix A.1 demonstrates that modern joint modeling approaches do not fare better in this example. In addition, when considering 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.

Identifiability does not imply consistency Our identification results, though an important first step, are far from results guaranteeing consistency of the imputation distribution and generally cannot explain the impressive performance of MICE in finite samples. In addition, a better understanding of when FCS imputation under CIMNAR is possible, might be fruitful.

Further MAR generating mechanisms may need to be considered It 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 Schouten et al. (2018), as we did in Section 4.4. Though their score is not proper under MAR, as shown in Appendix B.3, their analysis also showed mice-cart consistently in first place. Thus, tweaking the approach of Schouten et al. (2018) to produce MAR data with distribution shifts, might be an avenue for further research. In this context, the CIMAR assumption might be useful, as MAR examples with distribution shifts can easily be generated.

Acknowledgements

This work is part of the DIGPHAT project which was supported by a grant from the French government, managed by the National Research Agency (ANR), under the France 2030 program, with reference ANR-22-PESN-0017.

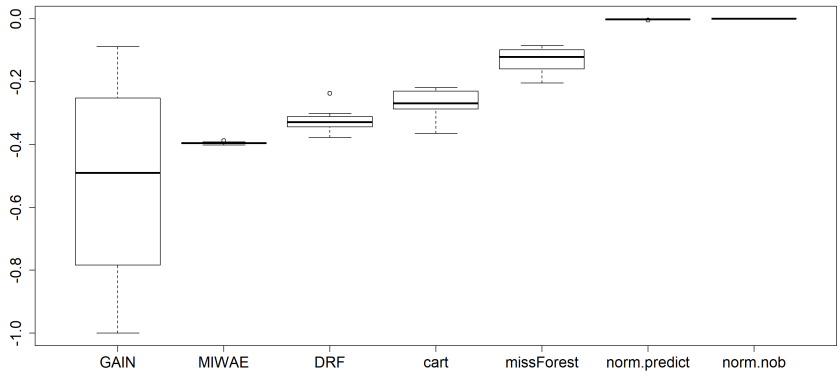

Figure 10: *Full information score for the Gaussian mixture model with distribution shift with GAIN and MIWAE.*

A Additional Empirical Considerations

A.1 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 4 to the joint modeling methods GAIN and MIWAE. The code for GAIN was taken from the original Github repository `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 10 – 13 show the results. Overall these two methods cannot compete with MICE and usually are scored last, except in the nonlinear example with distribution shift (Figure 11) 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 the air quality data application, as otherwise the implementation broke down, but did not do this for all the other examples. 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, GAIN and MIWAE tend to perform worse than missForest, even in terms of the energy distance. 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.

A.2 Comparison of the m-I-Score and DR-I-Score

We start by studying the real data example of Section 4.4 with missing values generated using the ampute function. A similar setting with different data sets was studied in Näf et al. (2023). Figure 14 shows the results using the DR-I-Score with 20 random projections. Interestingly, the DR-I-Score and the m-I-Score display the same ordering, though the DR-I-Score has somewhat less power to differentiate the methods.

Next, we study the Gaussian example with distribution shift of Section 4.1. As the DR-I-Score needs EMAR on each projection to be proper in a population setting, we do not use any random projections here. As the example is CIMAR, the requirement for propriety is met.

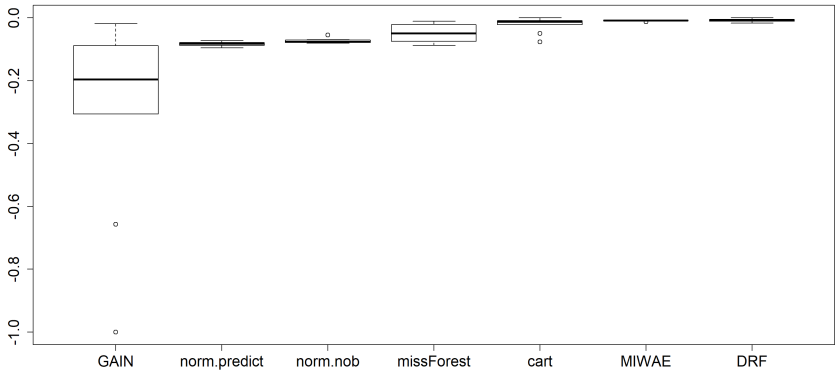

Figure 11: *Full information score for the nonlinear mixture model with distribution shift with GAIN and MIWAE.*

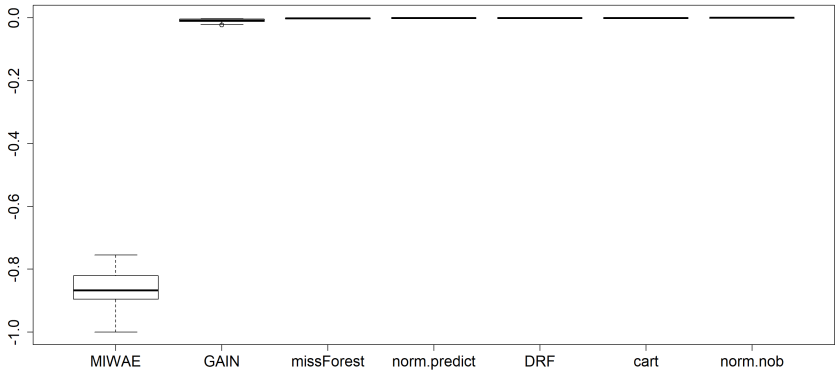

Figure 12: *Full information score for the MNAR data example with GAIN and MIWAE.*

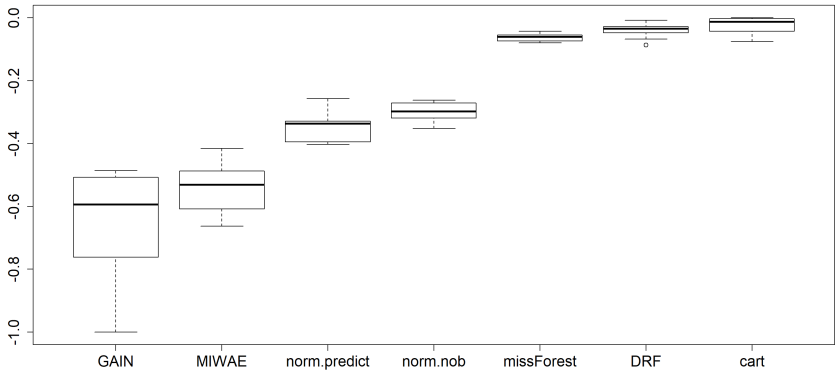

Figure 13: *Full information score for the air quality data example with GAIN and MIWAE, calculated with full data.*

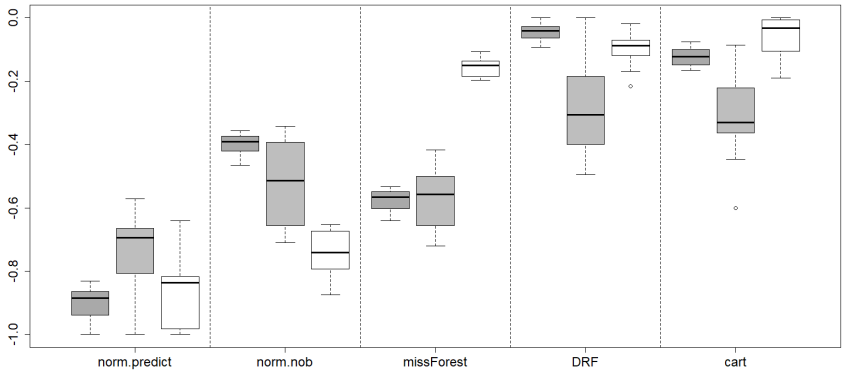

Figure 14: *Standardized scores for the air quality data example. Methods are ordered according to the full information score. For each method, the m-I-Score (dark gray, left), DR-I-Score (gray, middle), and full information score (white, right) are shown.*

Nonetheless, Figure 15 shows that the DR-I-Score erroneously scores mice-DRF and mice-cart in first place. This bias likely results because the DR-I-Score relies on a Random Forest classifier to estimate the involved density ratios. This classifier is likely not able to effectively deal with the distribution shift in this example. This showcases the advantage of the new score to let imputation methods "speak for themselves".

B 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 first highlight why the discussion of distribution shifts under MAR may not be relevant for Maximum Likelihood Estimation. In Section B.3, we then show that the score developed in Näf et al. (2023) is not proper under PMM-MAR.

B.1 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 θ. Moreover, assume the conditional distribution of $M \mid x$ is parametrized by ϕ. Then we can rewrite the MAR definition in SM-MAR slightly, as in Rubin (1976); Mealli and Rubin (2015):

$$\begin{aligned} &\mathbb{P}_\phi(M = m|x) = \mathbb{P}_\phi(M = m|\tilde{x}) \text{ for all } m \in \mathcal{M} \\ &\text{and } x, \tilde{x} \text{ such that } o(x, m) = o(\tilde{x}, m). \end{aligned} \tag{B.1}$$

As so far, ϕ and θ are not restricted to be finite-dimensional, this could in principle be assumed without loss of generality, such that (B.1) is indeed the same as condition SM-MAR. In the following, we will assume for simplicity that θ is finite-dimensional. Let Ω_θ be the space of θ, Ω_ϕ the space of possible ϕ and $\Omega_{\theta,\phi}$ the joint space of the parameters. The crucial additional condition is that:

$$\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi. \tag{B.2}$$

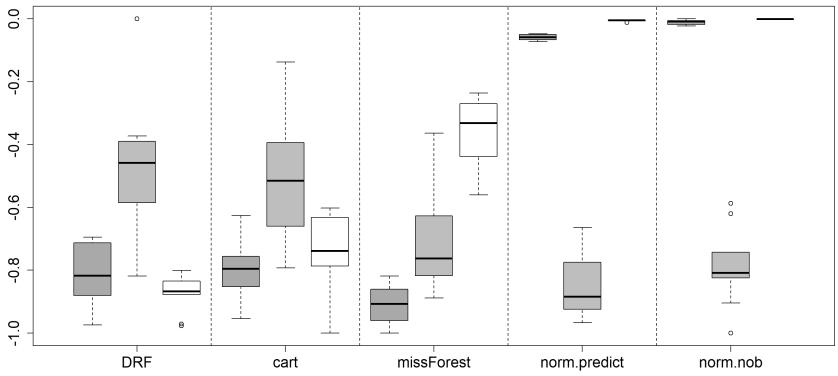

Figure 15: *Standardized scores for the Gaussian mixture example. Methods are ordered according to the full information score. For each method, the m-I-Score (dark gray, left), DR-I-Score (gray, middle), and full information score (white, right) are shown.*

This just means that ϕ is distinct from θ, so that $\mathbb{P}_\phi(M = m|x)$ does not depend on θ (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 θ. Under (B.1) 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 θ. 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 θ, ϕ (B.2) 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 θ to get the MLE. Whether this ignorability holds under MAR is a question of *parametrization*, as we illustrate in Example 4:

Example 2. *Consider the following setting:*

$$\mathcal{M} = \left\{ \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \right\},$$

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

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

We will revisit this example in Example 4 in Section B.4. 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:

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

Thus for optimization purposes, maximizing $p^*_{\theta,\phi}(x_1, x_2, x_3, M = m_1)$ over θ is equivalent to maximizing $p^*_\theta(x_1, x_2, x_3)$ over θ. In particular, being able to identify θ 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 θ, 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 θ can be recovered from only looking at the first pattern m_1. Indeed in this case:

$$p^*_{\theta,\phi}(x_1, x_2, x_3 \mid M = m_1) = \frac{\mathbb{P}(M = m_1 \mid x_1)}{\mathbb{P}(M = m_1)} p^*_\theta(x_1, x_2, x_3) = \frac{x_1}{\theta} p^*_\theta(x_1, x_2, x_3), \tag{B.4}$$

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 (B.2)*, maximizing $p^*_{\theta,\phi}(x_1, x_2, x_3, M = m_1)$ will not recover θ.*

Remark B.1. *Assuming* (B.2)*, ignorability also holds under EMAR and CIMAR.*

B.2 FCS under CIMNAR

We study an example under which FCS imputation under CIMNAR is biased.

Example 3. *We consider*

$$\mathcal{M} = \left\{ \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \end{pmatrix} \right\},$$

(X^*_1, X^*_2) to be standard Gaussian random variables with correlation ρ and $X^*_3 \sim N(0, 1)$ independently of X^*_1, X^*_2. We assume that M_1, M_2, M_3 are independent given X^* (i.e., (2.4)*), and*

$$\begin{aligned} \mathbb{P}(M_1 = 1 \mid x) &= \mathbb{P}(M_1 = 1 \mid x_2) = \mathbf{1}\{x_2 > 0\} * 0.8 \\ \mathbb{P}(M_2 = 1 \mid x) &= \mathbb{P}(M_2 = 1) = p. \end{aligned}$$

That is, X_2 is missing with probability p independently of x, while the probability of X_1 being missing depends on X_2. If $p = 0$, we recover a MAR mechanism. In the following, we set $p = 0.5$. In this case, it holds that:

$$\begin{aligned} \mathbb{P}(M = m_1 \mid x) &= \mathbb{P}(M = m_3 \mid x) = (1 - \mathbf{1}\{x_2 > 0\} \cdot 0.8) \cdot 0.5 \\ \mathbb{P}(M = m_2 \mid x) &= \mathbb{P}(M = m_4 \mid x) = \mathbf{1}\{x_2 > 0\} \cdot 0.8 \cdot 0.5. \end{aligned}$$

As X_2 is missing in patterns m_3 and m_4, SM-MAR II does not hold. However CIMNAR holds here and thus,

$$h^*(x_2 \mid x_1, x_3) = \tilde{h}^*(x_2 \mid x_1, x_3) = N(\rho x_1, 1 - \rho^2)(x_2) = p^*(x_2 \mid x_1, x_3).$$

However, if we would like to impute $X_2 \mid M = m_4$, it holds that

$$p^*(x_2 \mid x_{-2}, M = m_4) = p^*(x_2 \mid x_{-2}) \frac{\mathbb{P}(M = m_4 \mid x)}{\mathbb{P}(M = m_4 \mid x_{-2})} = 2N(\rho x_1, 1 - \rho^2)(x_2) \mathbf{1}\{x_2 > 0\},$$

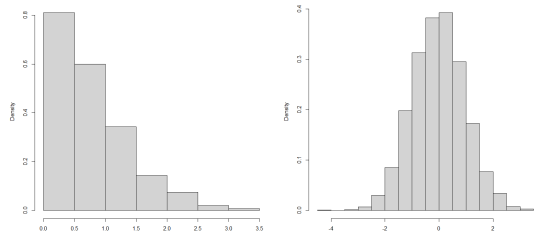

Figure 16: *Left: Actual distribution of $X_2^* \mid M = m_4$. Right: $N(0,1)$ Imputation.*

a half-normal distribution. Following the above modeling approach, we first impute X_2 in patterns m_3 and m_4 with $\tilde{h}^*(x_2 \mid x_{-2}) = N(\rho x_1, 1-\rho^2)(x_2)$. Again, this means we will replicate the distribution of X_2^* correctly, that is $Y_2 \overset{D}{=} X_2^*$, where $\overset{D}{=}$ means equal in distribution. But now assume we need to impute X_1 using this imputed X_2. To learn the desired $p^*(x_1 \mid x_{-1})$, we need to consider patterns in L_1, i.e., patterns m_1 and m_3, and then impute in patterns m_2 and m_4. Unfortunately, in these two combinations of patterns, Y_2 no longer has the correct distribution. This leads to two problems: First, it implies we cannot learn the correct conditional distribution from (X_1, Y_2) in patterns m_1 and m_3. While the correct conditional distribution would be $p^*(x_1 \mid x_2, M \in L_1) = p^*(x_1 \mid x_2) = N(\rho x_2, 1-\rho^2)(x_1)$, the actual conditional distribution of $X_1 \mid Y_2$ is proportional to $N(\rho y_2, 1-\rho^2)(x_1)(1 - \mathbf{1}\{y_2 > 0\} \cdot 0.8)$. Thus the bias in the imputation of X_2^* propagates into the conditional distribution of $X_1 \mid Y_2$. Note that this would not be an issue if we were able to observe (X_1, Y_2) in patterns m_1 and m_2. A similar problem arises on the imputation side. The joint distribution for (X_1^*, X_2^*) in patterns m_2 and m_4 is given as:

$$p^*(x_1, x_2 \mid M \notin L_1) = N(\rho x_2, 1-\rho^2)(x_1) \cdot \frac{\mathbf{1}\{x_2 > 0\} \cdot 0.8 \cdot 0.5}{0.5 \cdot 0.8 \cdot 0.5} N(0,1)(x_2),$$

such that the mixture of $p^*(x_1, x_2 \mid M \in L_1)$ and $p^*(x_1, x_2 \mid M \notin L_1)$ corresponds to $N(\rho x_2, 1-\rho^2)(x_1)N(0,1)(x_2)$. Even if one assumes that we were able to obtain the correct conditional distribution $p^*(x_1 \mid x_2) = N(\rho x_2, 1-\rho^2)(x_1)$ and use it to impute X_1 in patterns m_2 and m_4, obtaining Y_1 and imputed version of X_1^*, the joint distribution of (Y_1, Y_2) would be $N(\rho y_2, 1-\rho^2)(y_1)N(0,1)(y_2)$. This again does not correspond to $p^*(x_1, x_2 \mid M \notin L_1)$, rendering the distribution of (Y_1, Y_2) different than the one of (X_1^*, X_2^*).

We notice that in the above example there would be no issue for $\rho = 0$, such that X_1^* and X_2^* are independent. Figure 16 illustrates how $X_2 \mid M = m_4$ is still wrongly imputed, plotting the true density of $X^* \mid M = m_4$ against the imputation distribution ($N(0,1)$). Nonetheless, Figure 17 shows that the Gaussian imputation manages to accurately impute the distribution overall.

B.3 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

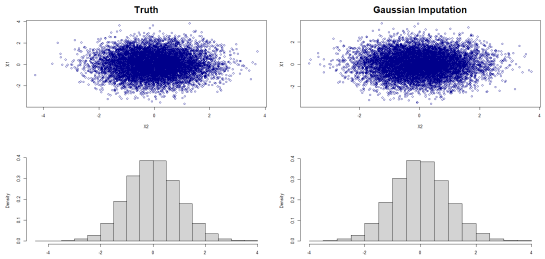

Figure 17: *Left: True joint distribution of (X_1^*, X_2^*) (top) and distribution of X_2^* (bottom). Right: Gaussian imputation of the joint distribution of (X_1, X_2) (top) and X_2 (bottom).*

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. $A = \{1, \dots, d\}$, this is a stronger condition than PMM-MAR, as was shown above. In particular, in Example 4, 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

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

Thus we have just shown that

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

However, inspecting the proof of (Näf et al., 2023, Proposition) reveals that their score is actually proper under EMAR:

Corollary B.1. *The I-Score defined in Näf et al. (2023) is proper under EMAR.*

B.4 Proofs

Corollary 2.1. *Condition SM-MAR is equivalent to SM-MAR II.*

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

$$\begin{aligned}
&\mathbb{P}(M = m|x) = \mathbb{P}(M = m|\tilde{x}) \Leftrightarrow \\
&\frac{p^*(x|M = m)\mathbb{P}(M = m)}{p^*(x)} = \frac{p^*(\tilde{x}|M = m)\mathbb{P}(M = m)}{p^*(\tilde{x})} \Leftrightarrow \\
&\frac{p^*(o(x, m), o^c(x, m) \mid M = m)}{p^*(o(\tilde{x}, m), o^c(\tilde{x}, m) \mid M = m)} = \frac{p^*(o(x, m), o^c(x, m))}{p^*(o(\tilde{x}, m), o^c(\tilde{x}, m))} \Leftrightarrow \\
&\frac{p^*(o^c(x, m) \mid o(x, m), M = m)}{p^*(o^c(x, m) \mid o(x, m))} = \frac{p^*(o^c(\tilde{x}, m) \mid o(x, m), M = m)}{p^*(o^c(\tilde{x}, m) \mid o(x, m))} \Leftrightarrow \\
&p^*(o^c(x, m) \mid o(x, m), M = m) = \frac{p^*(o^c(\tilde{x}, m) \mid o(x, m), M = m)}{p^*(o^c(\tilde{x}, m) \mid o(x, m))} p^*(o^c(x, m) \mid o(x, m)) \quad \text{(B.5)}
\end{aligned}$$

Integrating (B.5) 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. □

Example 4. *Consider*

$$\mathcal{M} = \left\{ \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \right\}, \tag{B.6}$$

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

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

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 both CIMAR and EMAR do not hold and that $p^*(x_1 \mid x_2, x_3)$ *is not identifiable from pattern* m_1*. The same argument shows that* $p^*(x_1 \mid x_2, x_3)$ *is also not identifiable from pattern* m_2*. Figure 18 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*. Indeed, PMM-MAR allows for a change in the conditional distributions over different patterns, and requires only that the distribution* $X_1^* \mid X_2^*, X_3^*$ *in pattern* m_3 *corresponds to the unconditional one.*

Proposition 2.3. *Assume* $|\mathcal{M}| > 3$*. Then for any pattern* $m \in \mathcal{M}$*,* $p^*(o^c(x,m) \mid o(x,m))$ *is*

- *identifiable from any other pattern* $m' \neq m$ *under CIMAR,*
- *identifiable from the pattern of fully observed data,* $m' = 0$*, under EMAR,*
- *is not identifiable from any single pattern* $m' \neq m$ *under PMM-MAR.*

In addition, if $\left| \sum_{j=1}^d m_j \right| > 1$*,* $p^*(o^c(x,m) \mid o(x,m))$ *is not identifiable from* L_m*.*

Proof. By definition of CIMAR and EMAR it directly follows that

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

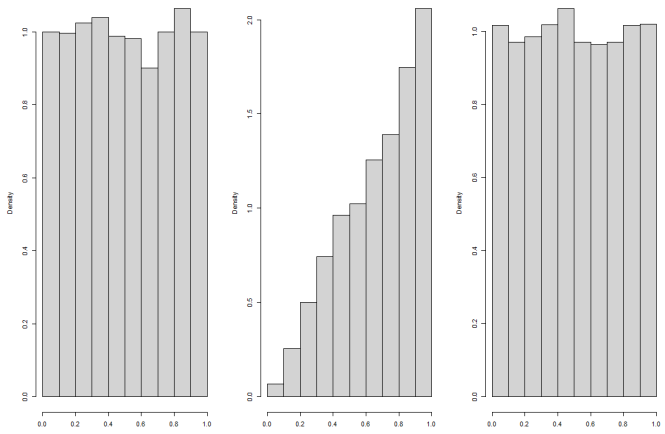

Figure 18: *Illustration of Example 4. Left: Distribution we would like to impute $X_1^* \mid M = m_3$. Middle: Distribution of X_1 in the fully observed pattern $(X_1 \mid M = m_1)$. Right: Distribution of all patterns for which X_1 is observed (Mixture of the distribution of X_1 in patterns m_1 and m_2).*

for all $m' \in \mathcal{M}$ and

$$p^*(o^c(x,m) \mid o(x,m), M = 0) = p^*(o^c(x,m) \mid o(x,m)),$$

respectively. Example 4 shows that under PMM-MAR,

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

might not hold for any $m' \neq m$. Finally Example 5 shows that there exists a MAR situation, where for some $m \in \mathcal{M}$,

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

is not equal to $p^*(o^c(x,m) \mid o(x,m))$ for any set of weights $w_{m'}(o(x,m))$. □

Example 5. *Consider*

$$\mathcal{M} = \left\{ \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} \right\}, \tag{B.7}$$

and $(X_1^*, X_2^*, X_3^*, X_4^*)$ are independently uniformly distributed on $[0,1]$. We further specify that

$$\begin{aligned}
\mathbb{P}(M = m_1 \mid x) &= \mathbb{P}(M = m_1 \mid x_1, x_2) = (x_1 + x_2)/8 \\
\mathbb{P}(M = m_2 \mid x) &= \mathbb{P}(M = m_2 \mid x_2) = 1/4 - x_2/8 \\
\mathbb{P}(M = m_3 \mid x) &= \mathbb{P}(M = m_3 \mid x_1) = 1/4 - x_1/8 \\
\mathbb{P}(M = m_4 \mid x) &= \mathbb{P}(M = m_4) = 2/4.
\end{aligned}$$

Again, $\mathbb{P}(M = m \mid x) = \mathbb{P}(M = m \mid o(x,m))$ and thus the MAR condition SM-MAR holds. Consider now $m = m_4$, such that $o^c(x,m) = (x_1, x_2)$. Then it holds that

$$\begin{aligned}
p^*(o^c(x,m) \mid o(x,m), M = m_1) &= \frac{1}{2}(x_1 + x_2)\, p^*(x_1, x_2 \mid x_3, x_4) \\
p^*(o^c(x,m) \mid o(x,m), M = m_2) &= (2 - x_2) p^*(x_1, x_2 \mid x_3, x_4)
\end{aligned}$$

Consider the mixture as in (2.1)*,*

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

with $\sum_{m' \in L_m} w_{m'}(o(x,m)) = 1$*. Then for* $h^*(o^c(x,m) \mid o(x,m)) = p^*(o^c(x,m) \mid o(x,m))$ *to hold, it must hold that for all* x_1, x_2*,*

$$w_{m_1}(x_3, x_4)\frac{1}{2}(x_1 + x_2) + w_{m_2}(x_3, x_4)(2 - x_2) = 1.$$

It is impossible to find a set of weights that meet this condition for all x_1, x_2 *simultaneously.*

Proposition 2.4. *Under PMM-MAR, the predictor* h^* *defined in* (2.2) *satisfies, for all* $j \in \{1, \dots, d\}$*, for all* x_{-j} *such that* $p^*(x_{-j}) > 0$*, and for all* x_j*,*

$$h^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j}). \tag{2.3}$$

Proof. Recall that

$$L_j = \{m \in \mathcal{M} : m_j = 0\}. \tag{B.8}$$

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

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

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

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

Thus in fact:

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

It only remains to show that

$$\frac{\sum_{m \in L_j} \mathbb{P}(M = m \mid x)}{\sum_{m \in L_j} \mathbb{P}(M = m \mid x_{-j})} = 1. \tag{B.9}$$

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 (B.9) indeed holds. □

Corollary 2.2. *Assumption SM-MAR implies Assumption CIMNAR, but Assumption CIMNAR does not imply Assumption SM-MAR.*

Proof. Proposition 2.4 showed that under SM-MAR for all x,

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

showing that $\mathbb{P}(M_j = 1 \mid x) = \mathbb{P}(M_j = 1 \mid x_{-j})$. □

Proposition 2.5. *Let $h^*(x_j \mid x_{-j}), \tilde{h}^*(x_j \mid x_{-j})$ be defined as in (2.2) and (2.5) respectively. Then, for all x_j, and for all x_{-j} with $p^*(x_{-j}) > 0$,*

$$h^*(x_j \mid x_{-j}) = \tilde{h}^*(x_j \mid x_{-j}) = p^*(x_j \mid x_{-j}), \tag{2.6}$$

if and only if CIMNAR holds.

Proof. Recall that $L_j = \{m \in \mathcal{M} : m_j = 0\}$ is the set of patterns in which x_j is observed and note that $\sum_{m \in L_j} \mathbb{P}(M = m \mid x) = \mathbb{P}(M_j = 1 \mid x)$. Thus under CIMNAR, it holds that

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

and similarly for $h^*(x_j \mid x_{-j})$. Now assume there exist x such that $h^*(x_j \mid x_{-j}) \neq p^*(x_j \mid x_{-j})$. This implies that

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

and thus that CIMNAR does not hold. □

Proposition 3.1. *Under Assumption 3.1, S_{NA} is a proper I-Score.*

Proof. We show that for each j,

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

holds. To ease notation, we define

$$es(H, y) = \frac{1}{2} \mathbb{E}_{\substack{X \sim H \\ X' \sim H}} [|X - X'|] - \mathbb{E}_{X \sim H}[|X - y|]$$

By propriety of the energy score (see e.g., Gneiting and Raftery (2007)) $\mathbb{E}_{Y\sim H^*_{X_j|x_{O_j}}}[es(H_{X_j|x_{O_j}}, Y)] \leq \mathbb{E}_{Y\sim H^*_{X_j|x_{O_j}}}[es(H^*_{X_j|x_{O_j}}, Y)]$. Taking expectations over $X_{O_j} \sim P^*_{X_{O_j}|M\in L_j}$ on both sides shows that

$$S^j_{NA}(H, P) = \mathbb{E}[\mathbb{E}[es(H_{X_j|X_{O_j}}, Y)]] \leq \mathbb{E}[\mathbb{E}[es(H^*_{X_j|X_{O_j}}, Y)]], \tag{B.10}$$

where we omitted the subscripts for a lighter notation. Moreover, Assumption 3.1 implies that

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

Thus, it follows that $H^*_{X_j|X_{O_j}} = P^*_{X_j|X_{O_j}}$, and thus:

$$\mathbb{E}[\mathbb{E}[es(H^*_{X_j|X_O}, Y)]] = \mathbb{E}[\mathbb{E}[es(P^*_{X_j|X_O}, Y)]] = S^j_{NA}(P^*, P). \tag{B.11}$$

Combining (B.10) and (B.11) 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.

Beesley, L. J., Bondarenko, I., Elliot, M. R., Kurian, A. W., Katz, S. J., and Taylor, J. M. (2021). Multiple imputation with missing data indicators. *Statistical Methods in Medical Research*, 30(12):2685–2700.

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.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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.

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., and Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing lc-ms metabolomics data: a comparative study. *BMC Bioinformatics*, 20(1):492.

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.

Little, R., Rubin, D., and Safari, a. O. M. C. (2019). *Statistical Analysis with Missing Data., 3rd Edition*. Wiley.

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.

Malistov, A. and Trushin, A. (2019). Gradient boosted trees with extrapolation. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 783–789. IEEE.

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 auto-encoders. *GigaScience*, 9(8):giaa082.

Ren, B., Lipsitz, S. R., Weiss, R. D., and Fitzmaurice, G. M. (2023). Multiple imputation for non-monotone missing not at random data using the no self-censoring model. *Stat Methods Med Res*, 32(10):1973–1993.

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.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

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

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

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

Shpitser, I. (2016). Consistent estimation of functions of data missing non-monotonically and not at random. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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.