Distribution Shift in Missing Data Imputation: A Risk-Based Perspective and Importance-Weighted Correction under MAR

Luke Shannon
School of Mathematics
University of Bristol
Bristol, United Kingdom
nq24207@bristol.ac.uk

Song Liu
School of Mathematics
University of Bristol
Bristol, United Kingdom
song.liu@bristol.ac.uk

Katarzyna Reluga
School of Business and Economics
Humboldt University of Berlin
Berlin, Germany
katarzyna.reluga@hu-berlin.de

Abstract

Missing data imputation, where a model is trained on observed data to estimate unobserved values, is a fundamental problem in machine learning. In this paper, we rigorously formulate imputation model learning as a mean-squared error risk minimisation problem. We show that when the probability of missingness depends on the data, many state-of-the-art methods fail to account for the resulting distribution shift between the observed data used for training and the full data distribution used for evaluation. Consequently, these approaches do not minimise mean-squared error on the full data distribution. Instead, we propose a novel imputation algorithm designed to learn an imputation model from the observed data while explicitly incorporating a correction for this distribution shift. Simulation studies show consistent improvements over otherwise identical uncorrected baselines, with average reductions of 3% in RMSE and 7% in Wasserstein distance.

1 Introduction

Missing data poses a fundamental obstacle to statistical inference in applied settings. Most classical estimators and modern machine learning methods are designed for complete data and break down in the presence of missingness. A well-established approach, known as imputation, is to complete the dataset by estimating the missing entries, thereby enabling the application of complete-data methods. A wide range of imputation approaches have been proposed, including generative models such as GAIN and MIWAE [Yoon et al., 2018, Mattei and Frellsen, 2019], optimal transport–based methods such as Sinkhorn [Muzellec et al., 2020], low-rank matrix completion techniques [Mazumder et al., 2010], iterative procedures such as the **e**xpectation-**m**aximisation algorithm (EM) [Dempster et al., 1977], and round robin iterative procedures such as **m**ultiple **i**mputation by **c**hained **e**quations (MICE) [van Buuren and Groothuis-Oudshoorn, 2011], MissForest [Stekhoven and Bühlmann, 2012] and HyperImpute [Jarrett et al., 2022].

We consider the problem of learning an imputation model from observed data that accurately recovers unobserved entries. We adopt a risk minimisation approach to this [Vapnik, 1998], where model performance is measured via a loss comparing true and imputed values. However, the risk, as defined with respect to the true data, has an implicit dependence on unobserved entries, which are not available

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

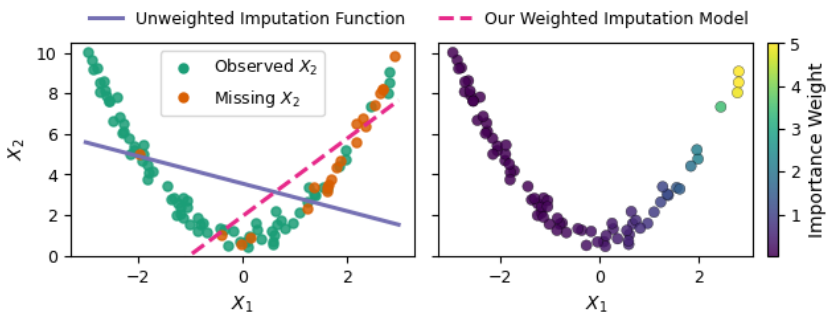


Figure 1: The left panel illustrates a two-dimensional dataset where missingness in X_2 depends on the fully observed variable X_1, such that points with higher values of X_1 are more likely to have missing values in X_2, inducing a distribution shift between the observed data (green points) and the full data distribution. As a result, an imputation function learned without correcting for the selection bias (blue line) is less accurate than a weighted approach (dashed red line) in predicting the missing values (red points). The weighted approach uses the importance weights (right panel) in its training to correct for the selection bias. The plot is best viewed in colour.

at learning time. Nevertheless, under the assumption that the missingness mechanism depends only on observed variables, we show that the risk can be written as an importance-weighted expectation over the observed data. These weights highlight a distributional shift between the observed data used for training and the full data distribution used for evaluation, which has not been explicitly accounted for in existing risk-based imputation approaches. Building on our theoretical results, we propose an importance-weighted risk minimisation algorithm that explicitly incorporates a correction for the distribution shift between observed data and full data distributions.

Related Work: Many existing imputation methods, including likelihood-based, iterative, and adversarial approaches, can be interpreted within our risk minimisation framework, enabling direct comparison with our approach. Likelihood-based methods such as EM and MIWAE minimise an observed-data negative log-likelihood risk; round-robin iterative approaches such as MICE, MissForest, and HyperImpute optimise a sequence of reconstruction-based risks defined on observed entries; and GAIN minimises a reconstruction loss on observed data under an adversarial regularisation scheme. While these methods differ in modelling assumptions and optimisation procedures, they all share a reliance on observed data and the absence of an explicit correction for the resulting mismatch with the full data distribution. In contrast, our approach explicitly incorporates a correction for this distributional shift within a risk minimisation framework via importance weighting. We place our findings in a broader literature context in Section 6.

Our contributions are as follows:

- In Section 3.1, we show the imputation risk can be expressed as an importance-weighted expectation over observed data, making explicit the effect of distributional shift on the risk.
- In Section 4, we propose an imputation algorithm that incorporates our derived importance weights in the training objective. This approach is illustrated Figure 1.
- In Section 5, we benchmark our proposed importance-weighted algorithm against a range of state-of-the-art imputation methods.

2 Problem Setup and Background

To formalise our approach, we introduce the notation for modelling missing data, and define imputation functions together with their associated risk under the full data-generating distribution.

2.1 Missing Data Notation and Mechanisms

Throughout the paper, uppercase letters denote random variables and lowercase letters their realisations, with subscripts i and $\neg i$ denoting coordinate-wise and complement indexing respectively. Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ denote a d-dimensional random vector drawn from a distribution P_X with

density $p(x)$, representing the true joint distribution of the data. Let $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ denote its support, where each $\mathcal{X}_i \subset \mathbb{R}$ is either a finite set (if X_i is discrete) or a possibly disconnected subset of $\mathbb{R}$ (if X_i is continuous), for $i = 1, \ldots, d$. We assume throughout that the data satisfies $\mathbb{E}[X_i^2] < \infty$ for all variables i, ensuring that all risks considered in this work are well-defined.

In practice, we do not observe X directly, but instead observe a partially missing version determined by a missingness pattern. Let $R = (R_1, \ldots, R_d) \in \{0, 1\}^d$ denote the missingness pattern drawn from a distribution P_R with density $p(r)$, where $R_i = 1$ indicates that X_i is observed and $R_i = 0$ that it is missing. We assume (X, R) is distributed according to $P_{X,R}$ with joint density $p(x, r)$. We model missingness by introducing a symbol $* \notin \bigcup_{i=1}^d \mathcal{X}_i$ to denote missing values, and introduce the partially observed vector $\tilde{X} \in \tilde{\mathcal{X}} := \prod_{i=1}^d \tilde{\mathcal{X}}_i$, where $\tilde{\mathcal{X}}_i := \mathcal{X}_i \cup \{*\}$. $\tilde{X}$ is defined component-wise by $\tilde{X}_i = X_i$ if $R_i = 1$ and $\tilde{X}_i = *$ if $R_i = 0$ Draws from $\tilde{X}$ correspond to the partially observed data available in practice. For any realisation $r \in \{0, 1\}^d$, the vector X induces a decomposition into observed and missing components $X_{\text{obs}} = (X_j : r_j = 1)$ and $X_{\text{miss}} = (X_j : r_j = 0)$.

Finally, the joint density factorises as $p(x, r) = p(r \mid x)p(x)$, defining the missing data mechanism. We distinguish three standard regimes: Missing Completely At Random (**MCAR**), where $p(r \mid x) = p(r)$; Missing At Random (**MAR**), where $p(r \mid x) = p(r \mid x_{\text{obs}})$; and Missing Not At Random (**MNAR**), otherwise.

We use "density" throughout to refer to both probability density and mass functions, and assume all distributions are absolutely continuous with respect to the Lebesgue measure for continuous variables and the counting measure for discrete variables.

2.2 Imputation Functions and their Risk

We cast the imputation problem within a risk minimisation framework [Vapnik, 1998], aiming to learn an imputation function that minimises discrepancy between imputed and true data. Mean squared error (MSE) is a natural choice of discrepancy measure due to its interpretability and widespread use [Yoon et al., 2018, Stekhoven and Bühlmann, 2012]. Accordingly, we focus on MSE risk minimisation. We formalise the risk via a functional $\mathcal{J} : \mathcal{G} \to \mathbb{R}$ defined over a hypothesis class $\mathcal{G}$ of measurable functions with finite variance, ensuring that $\mathcal{J}(g)$ is well-defined and finite for all $g \in \mathcal{G}$. We now define the central objects of our study.

Definition 1 (Imputation function)**.** An imputation function $g : \tilde{\mathcal{X}} \to \mathcal{X}$ maps a partially observed vector $\tilde{x}$ to a complete vector $g(\tilde{x}) = \big(g_i(\tilde{x}) : i \in \{1, \ldots, d\}\big)$, subject to the constraint that observed entries are left unchanged. That is, for each coordinate i such that $r_{\tilde{x},i} = 1$, where $r_{\tilde{x}}$ denotes the missingness pattern associated with $\tilde{x}$, we require $g_i(\tilde{x}) = \tilde{x}_i$.

Definition 2 (Imputation risk)**.** The MSE risk of an imputation function g, given in Definition 1, is

$$\mathcal{J}(g) = \mathbb{E}_{X,R}\Big[\text{MSE}\big\{g(\tilde{X}), X\big\}\Big], \tag{1}$$

where $\tilde{X}$ is induced by (X, R) as described in Section 2.1.

Let $\mathcal{P} := \{\, i \in \{1, \ldots, d\} : p(R_i = 0) > 0 \,\}$ denote the set of coordinates subject to missingness. Under MSE loss, the imputation risk decomposes additively over $\mathcal{P}$ into a collection of univariate regression problems. We restrict attention to the decomposed form of the risk given in Lemma 1 for the remainder of the paper; a derivation is provided in Appendix A.1.

Lemma 1. *For a deterministic imputation map given in Definition 1, the risk in equation* (1) *satisfies*

$$\mathcal{J}(g) = \sum_{i \in \mathcal{P}} p(R_i = 0)\, \mathcal{J}_i(g_i), \tag{2}$$

where

$$\mathcal{J}_i(g_i) = \mathbb{E}_{X,R_{\neg i}}\left[\Big\{g_i(\tilde{X}) - X_i\Big\}^2 \mid R_i = 0\right]. \tag{3}$$

Remark 1. Optimising (3) directly over consistency-constrained functions is non-standard as typical hypothesis classes (e.g. linear models, random forests) cannot naturally encode consistency. In practice, we will see in our proposed Algorithm 1, that we optimise over non-consistency-constrained functions and enforce consistency by completing missing entries within the data and leaving already observed entries unchanged — an approach that mirrors other risk-based imputation methods [van Buuren and Groothuis-Oudshoorn, 2011, Stekhoven and Bühlmann, 2012, Dempster et al., 1977].

2.3 Covariate Shift

Covariate shift is central to the theory developed in this paper. Lemma 1 shows that the imputation risk decomposes into a collection of univariate regression problems. In Section 3.1, we show that the univariate regression problems arising from Lemma 1 admit a covariate shift formulation under the observed data distribution. To formalise the connection, we recall the standard covariate shift setting. Consider a regression problem in which we aim to predict an output y from a predictor x. Let $p_{\text{train}}(x, y)$ and $p_{\text{test}}(x, y)$ denote the joint distributions of the training and testing data respectively. Covariate shift refers to the setting in which the input marginal distributions differ, $p_{\text{train}}(x) \neq p_{\text{test}}(x)$, while the conditional distribution remains invariant, $p_{\text{train}}(y \mid x) = p_{\text{test}}(y \mid x)$. In covariate shift settings, empirical risk minimisation on the training distribution is biased with respect to the test distribution. A standard correction is importance weighting, which reweights training samples by the density ratio $p_{\text{test}}(x)/p_{\text{train}}(x)$ in order to recover an unbiased estimate of the test risk [Shimodaira, 2000, Sugiyama et al., 2007].

3 Risk Minimisation Formulation for Imputation

In Section 3.1, we derive an observed data representation of the coordinate-wise imputation risks from Lemma 1, revealing a connection to covariate shift. In Sections 3.2 and 4, we build on this representation to develop the proposed imputation algorithm.

3.1 The Oracle Risk Minimisation Problem

The coordinate-wise risks $\mathcal{J}_i(g_i)$, defined in equation (3), are intractable since they depend on the unobserved quantity $X_i \mid R_i = 0$. In Proposition 1, we show that, under the MAR assumption and overlap condition defined in Definition 3, each $\mathcal{J}_i(g_i)$ can be equivalently written as importance-weighted expectation over observed data. The proof of Proposition 1 is given in Appendix A.2.

Definition 3 (Overlap)**.** For $i \in \mathcal{P}$, we say that the overlap condition holds for R_i if

$$0 < p(R_i = 1 \mid X_{\text{obs}} = x_{\text{obs}}, R_{\neg i} = r_{\neg i}) < 1,$$

for all $(x_{\text{obs}}, r_{\neg i})$ such that $p(x_{\text{obs}}, r_{\neg i} \mid R_i = 0) > 0$.

Proposition 1. *For $i \in \mathcal{P}$, assume that the MAR assumption holds, so that $p(r \mid x) = p(r \mid x_{\text{obs}})$, and that R_i satisfies the overlap condition in Definition 3. Then the functional $\mathcal{J}_i(g_i)$, defined in equation* (3)*, can be expressed in terms of observed data as:*

$$\mathcal{J}_i(g_i) = \mathbb{E}_{X, R_{\neg i}}\Big[w_i(X_{\text{obs}}, R_{\neg i}) \left\{g_i(\tilde{X}) - X_i\right\}^2 \mid R_i = 1\Big]. \tag{4}$$

where the weights $w_i(\cdot, \cdot)$ are defined by

$$w_i(x_{\text{obs}}, r_{\neg i}) := \frac{p(x_{\text{obs}}, r_{\neg i} \mid r_i = 0)}{p(x_{\text{obs}}, r_{\neg i} \mid r_i = 1)}. \tag{5}$$

Proposition 1 reveals an interesting phenomenon. Optimal imputation can be characterised as solving a collection of covariate shift problems, one for each partially observed coordinate in the data, making explicit a link between the fields of missing data imputation and covariate shift. Conditioning on the missingness indicator R_i induces two subpopulations: the observed entries ($R_i = 1$) and the unobserved entries ($R_i = 0$). Under MAR, the conditional distribution $X_i \mid (X_{\text{obs}}, R_{\neg i})$ is invariant across these subpopulations, while the marginal distribution of $(X_{\text{obs}}, R_{\neg i})$ are not necessarily the same. The i^{th} coordinate risk in equation (4) is therefore analogous to covariate shift, as introduced in Section 2.3, where X_i plays the role of the response variable, $(X_{\text{obs}}, R_{\neg i})$ are the covariates, and the observed and unobserved subpopulations correspond to training and test distributions respectively. The weights in equation (5), analogous to $p_{\text{test}}(x)/p_{\text{train}}(x)$, reweight the observed data so that empirical risk minimisation targets the distribution of X_i in the unobserved subpopulation.

In contrast, existing risk based imputation approaches can be viewed as minimising an approximation of equation (4) in which the weighting term is omitted. Such an approach implicitly fails to account for the distributional shift between observed and full data distributions, yielding imputation models that are optimised for the wrong distribution.

Algorithm 1 Weighted Iterative Imputation

Input: Incomplete dataset $\mathcal{D}^{\text{miss}}$, visitation order v, convergence criteria δ.
Output: Completed dataset $\hat{\mathcal{D}}$.
$\hat{\mathcal{D}} \leftarrow$ `Initial Imputation`$(\mathcal{D})$
repeat
  for column i **in** v **do**
    Estimate w_i using Algorithm 2.
    Fit g_i to minimise (6) using estimated w_i as the weights.
    Replace missing entries in column i of $\hat{\mathcal{D}}$ with predictions from $\hat{g}_i$.
  end for
until Convergence criteria δ is satisfied
return $\hat{\mathcal{D}}$

3.2 Our Surrogate Risk Objective

Although the oracle coordinate-wise risk derived in equation (4) provides a principled target for imputation, both g_i and w_i depend on X_{obs} and $R_{\neg i}$. The weights w_i depend on these quantities explicitly, while g_i depends on them implicitly through its input $\tilde{X}$. This dependence has two consequences. First, conditioning on $R_{\neg i}$ increases the effective dimensionality of both g_i and w_i, since each partially observed variable contributes additional pattern-specific features, which becomes particularly pronounced in settings with many variables subject to missingness. Second, the structure of X_{obs} itself varies across missingness patterns, leading to heterogeneity in the resulting optimisation problems across coordinates and patterns. As a consequence, a naive minimisation of equation (4) would require solving a collection of pattern-specific risk minimisation problems, which is computationally and statistically inefficient [Stempfle et al., 2023]. We therefore seek an approach that avoids the need for pattern-specific conditional models while maintaining a low-dimensional conditioning set for each model.

Low dimension modelling set: Under the MAR assumption, the missingness indicators $R_{\neg i}$ are conditionally independent of X_i given X_{obs}, implying that they do not provide additional information for the imputation task beyond X_{obs}. The conditional independence provides a principled justification for restricting attention to imputation models in which both g_i and w_i depend only on X_{obs}, yielding a lower dimensional form of our optimisation problem without loss of relevant information.

Avoiding pattern specific modelling: To overcome the need for pattern-specific modelling, we adopt a round robin iterative strategy [van Buuren and Groothuis-Oudshoorn, 2011, Stekhoven and Bühlmann, 2012]. In such approaches, missing data is typically initialised using column-wise means. The algorithm then proceeds by cycling through partially observed variables in turn. At each step, a conditional imputation model is fitted for the target variable, using the current values of the remaining variables as predictors. Missing entries are then updated using the fitted model. A key advantage of this approach is that each iteration operates on a fully imputed dataset, thereby avoiding the need to construct separate models for different missingness patterns. Round-robin imputation approaches are widely used in practice and across applied fields such as epidemiology [Resche-Rigon et al., 2013] and psychology [Xu et al., 2020] due to their flexibility, interpretability, and robust empirical performance — advantages that we retain by adopting this framework.

Surrogate risk formulation: We therefore propose minimising a data-dependent surrogate risk within a round robin scheme. At iteration t, we treat the current imputed dataset $\hat{X}^{(t)}$ as given. When updating coordinate i, the goal is to select an imputation function g_i that performs well with respect to this current state of the data. To this end, we define the surrogate objective

$$\mathcal{J}_i^{(t)}(g_i) = \mathbb{E}\Big[w_i(\hat{X}_{\neg i}^{(t)})\big(g_i(\hat{X}_{\neg i}^{(t)}) - X_i\big)^2 \mid R_i = 1\Big]. \tag{6}$$

which will form the basis of our approach.

4 Our Weighted Imputation Algorithm

We now describe our proposed weighted iterative imputation algorithm. The algorithm proceeds in a manner analogous to existing round-robin iterative approaches [van Buuren and Groothuis-

Algorithm 2 Estimate weights w_i for column i

Input: Imputed dataset $\hat{\mathcal{D}} = \{\hat{x}^{(k)}, r^{(k)}\}_{k=1}^{N}$, target column i.
Output: Estimated weights $\{w_i^{(k)}\}_{k=1}^{N}$ for column i.
Procedure:
Model $p(R_i = 1 \mid \hat{x}_{\neg i})$ using $\eta_i(\hat{x}_{\neg i})$ trained using binary classification loss.
Compute weights: $w_i^{(k)} \leftarrow \frac{1-\eta_i\left(\hat{x}_{\neg i}^{(k)}\right)}{\eta_i\left(\hat{x}_{\neg i}^{(k)}\right)}$.
return $\{w_i^{(k)}\}_{k=1}^{N}$

Oudshoorn, 2011, Stekhoven and Bühlmann, 2012, Jarrett et al., 2022]. The key difference is that the importance weights w_i in the surrogate risk (6) are unknown and must also be estimated from data at each iteration and incorporated into the estimation of the conditional imputation model g_i.

As in existing approaches, the practitioner specifies a conditional model hypothesis class $\mathcal{G}_i$ for each partially observed coordinate $i \in \mathcal{P}$, as defined in Section 2.2, from which the imputation function g_i is selected. In addition, we introduce a corresponding conditional model class $\mathcal{W}_i$, from which the importance-weighting functions are estimated at each iteration.

Algorithm 1 implements the surrogate risk minimisation procedure described in Section 3.2. The algorithm iterates through the coordinates according to a visitation order v, which is an ordered subset of $\mathcal{P}$. At iteration t when updating column i, the current imputed dataset provides an empirical realisation of $\hat{X}^{(t)}$ given in equation (6). Given this state, importance weights are estimated using a binary classification model selected from the hypothesis class $\mathcal{W}_i$, following a density ratio estimation approach [Qin, 1998, Bickel et al., 2007], as detailed in Algorithm 2. The resulting weights are then used to fit the conditional imputation function $g_i \in \mathcal{G}_i$ by minimising an empirical approximation of (6), after which missing entries in column i are updated while observed entries remain unchanged. The procedure is repeated until the stopping criterion δ is satisfied.

4.1 Practical Implementation

We now describe the implementation details of Algorithms 1 and 2. When updating column $i \in \mathcal{P}$, we proceed as follows.

Weight Estimation: Following Algorithm 2, we estimate importance weights by modelling the missingness mechanism $p(R_i = 1 \mid X_{\neg i})$ using a binary classification model from a hypothesis class $\mathcal{W}_i$ of logistic regression models. Hyperparameters are tuned over a predefined grid on a random subsample of the data, to minimise standardised mean differences (SMD) [Austin and Stuart, 2015] between observed and missing values. The estimated probabilities are transformed into importance weights and clipped at the 5th and 95th percentiles to reduce the influence of extreme values.

Clipping weights is standard practice in the importance weighting literature where, for finite samples, a small number of extreme weights can substantially inflate the variance of importance-weighted estimators, often leading to poorer empirical performance than a biased but more stable alternative under clipped weights [Cole and Hernán, 2008, Petersen et al., 2012, Ju et al., 2019]

Stabilisation Measures: Incorporating the estimated weights can reduce the effective variance of covariates under the weighted distribution, leading to unstable conditional model estimates. To mitigate this, we identify and remove, at each iteration, features with variance below 10^{-12}. This filtering step is performed prior to conditional model fitting.

Regression Estimation: Given the observed data, estimated weights, and the stabilised set of covariates, we fit a weighted regression model. To control variance, we follow [Sugiyama et al., 2007] and introduce a tempering parameter γ. The latter is selected via cross-validation on a random subset of the observed data for column i. The final model is then estimated using weighted least squares with weights w_i^{γ}. Since tuning γ at each iteration represents the primary computational cost of our procedure, we adopt a warm-starting strategy to reduce overhead and improve stability across iterations. Further details are provided in Appendix C.4.

Convergence and Early Stopping: We monitor convergence using the median change in imputed values between successive iterations, as specified by the tolerance parameter δ in Algorithm 1. Furthermore, a maximum number of iterations is used as a safeguard.

5 Experiments

For linear, Random Forest (RF), and Multi-Layer Perceptron (MLP) conditional model classes, we compare our method against an unweighted round-robin iterative baseline using the same conditional imputation model class, so that performance differences are primarily attributable to the inclusion of importance weights. The unweighted baseline corresponds to HyperImpute [Jarrett et al., 2022] restricted to linear conditional models in the linear setting, and to scikit-learn's IterativeImputer [Pedregosa et al., 2012] in the RF and MLP settings.

5.1 Experimental Setup

Datasets: We study eight datasets spanning a range of tabular, time-series, and image modalities to evaluate performance across diverse data regimes and assess robustness of our proposed method beyond any single domain. Datasets have been preprocessed to remove zero-variance columns and to cap extreme values so that skewness lies in the range $[-10, 10]$ to ensure that results are not unduly influenced by a small number of extreme observations. All dataset details and preprocessing steps are provided in Appendix C.1.

MAR Simulation: To simulate MAR data, we randomly select up to four continuous features and designate them as partially observed. Let $\mathcal{M} \subseteq \{1, \dots, d\}$ denote the indices of these features. For each $i \in \mathcal{M}$, the missingness mechanism is defined to depend only on a subset of fully observed covariates. Specifically, we sample a set of indices $\mathcal{C}_i \subseteq \{1, \dots, d\} \setminus \mathcal{M}$, and define the observation probability via a logistic model $\mathbb{P}(R_i = 1 \mid X) = \sigma\Big(\alpha \sum_{j \in \mathcal{C}_i} X_j + \beta_i\Big)$ where $\sigma(z) = (1 + e^{-z})^{-1}$. Here, α controls the strength of dependence on the covariates, and β_i is chosen to achieve a desired marginal missing rate for feature X_i. We repeat experiments over 35 seeds, and within each seed we vary the MAR strength parameter $\alpha \in [-3, 3] \cap \mathbb{Z}$. For each seed, the set of variables $\mathcal{M}$ and the corresponding predictor sets $\mathcal{C}_i$ are fixed. This yields 245 simulation setups per dataset. Unless stated otherwise, β_i is chosen to simulate a 30% missingness rate.

Performance Metrics: Let D^{true} denote the complete dataset and $\hat{D}$ an imputed version of an incomplete dataset D^{miss}. We evaluate imputation accuracy using the root mean squared error (RMSE) over the missing entries, indexed by $M = \{(i, j) : D^{\text{miss}}_{i,j} \text{ is missing}\}$. To assess distributional similarity, we compute a marginal Wasserstein distance (WD), defined as the sum of one-dimensional Wasserstein distances across partially observed features: $\mathrm{W}(D^{\text{true}}, \hat{D}) = \sum_{j : X_j \in \mathcal{M}} W_1\Big(D^{\text{true}}_{(j)}, \hat{D}_{(j)}\Big)$ where W_1 denotes the 1-Wasserstein distance between empirical distributions Villani [2008]. Both metrics are calculated on the original data scale. To summarise comparative performance, for each of the 245 simulation runs we compute the performance metric under both the weighted and unweighted baselines, and report the ratio of weighted to unweighted values. The only exception is in the downstream prediction results in Figures 10 and 11, where we also compare against an additional unimputed baseline. In this setting, we report the performance metrics without further processing.

5.2 Experimental Results

Overall Performance and Runtime: Figure 2 summarises the overall results grouped by dataset and conditional imputation model class. Our method improves RMSE in 20 out of 24 simulation settings and Wasserstein distance in 23 out of 24 settings, providing strong support for the efficacy of the weighted approach. We find that RMSE gains are consistently smaller for more expressive models such as RF and MLP. This is expected behaviour, and is supported by the theory in Appendix B.0.1, where we show that the benefit of importance weighting diminishes as the approximation error of conditional imputation model classes decrease. Importantly, Figure 2 shows that while improvements in RMSE are modest in some settings, with displayed 95% confidence intervals overlapping or close to 1, Wasserstein distance exhibits consistently larger gains, with confidence intervals remaining well below 1 across most settings. These results reflect improved distributional fidelity, which Van Buuren [2018] identifies as important for high-quality imputation.

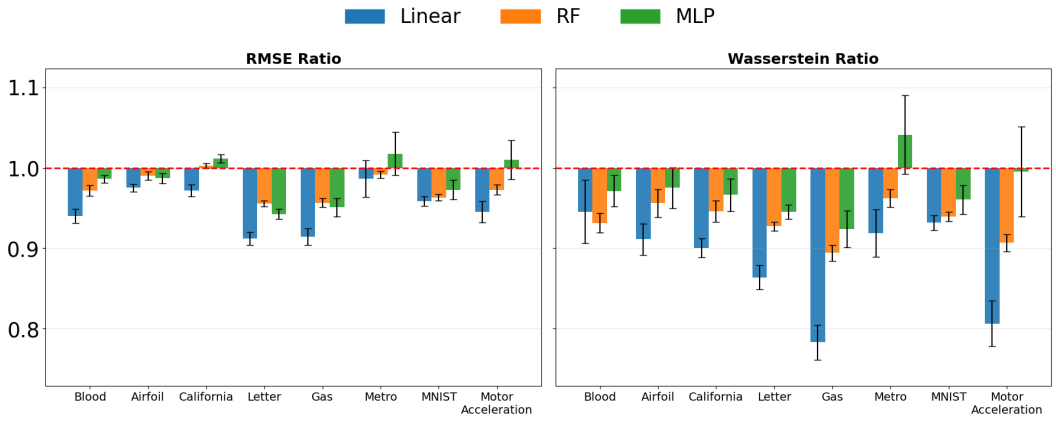


Figure 2: Values below 1 indicate improved performance. We report mean performance ratios of our weighted method relative to the unweighted baseline. Error bars represent approximate 95% confidence intervals, computed as ± 1.96 standard errors of the mean (SEM). Results are reported for RMSE and Wasserstein distance. The plot is best viewed in colour.

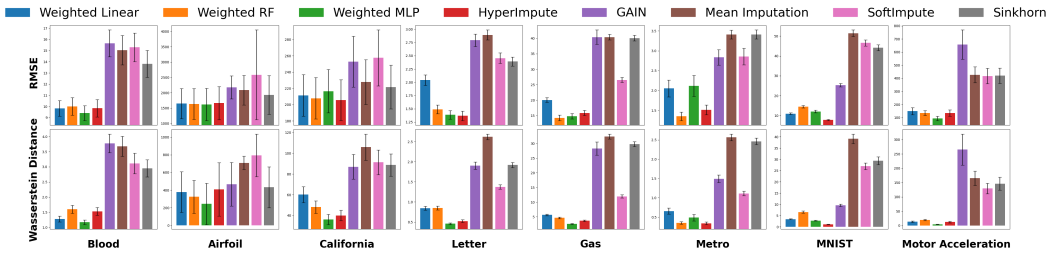


Figure 3: Lower values indicate better performance. The figure compares a broad range of methods, showing that the weighted variants are competitive across all scenarios. Bars denote mean performance, with error bars given by ± 1.96 SEM, corresponding to approximate 95% confidence intervals. Results are reported for RMSE and Wasserstein distance. The plot is best viewed in colour.

Experimental comparisons to a wider range of imputation methods are shown in Figure 3. We provide a brief description of the considered benchmarks in Appendix C.2. We observe that the round robin based imputation methods — our weighted approaches and HyperImpute — display better average RMSE and WD performance than other benchmarks across all considered settings. Unlike HyperImpute, which performs automatic model selection for each column as part of its imputation procedure, our weighted approach uses a fixed conditional model class (Linear, RF, or MLP) across all columns. Comparing our approach to HyperImpute, we observe RMSE improvements under at least one conditional model class for 5/8 datasets, and similarly for WD in 6/8 datasets. These results suggest that both the choice of conditional model and the use of importance weighting matter for accurate imputation. Integrating our weighted framework with HyperImpute's model selection, so that model selection itself accounts for the distribution shift between observed and full data, is an interesting direction for future work.

Runtime: Our proposed method achieves faster runtimes for linear conditional models and incurs only a 10% increase in runtime for RF and MLP models, indicating that extra computational cost remains manageable in practice. A full breakdown of run times are available in Appendix C.3, and details are provided in Appendix C.4 on practical steps control runtime in our method.

Sensitivity Analysis: As missingness is simulated under a known logistic regression model, the true importance weights $w_i = \frac{p(R_i=0|X_{\text{obs}})}{p(R_i=1|X_{\text{obs}})}$ can be computed for each data point. We summarise these using the 95^{th} percentile, q_{95}, where larger values indicate greater covariate shift. Figure 4 shows that for small q_{95}, corresponding to MCAR or weak MAR, weighting provides little improvement over the unweighted baseline, consistent with Appendix B.0.2. As q_{95} increases, weighting yields increasingly consistent improvements, including under substantial covariate shift. Under the MLP

model, intervals occasionally overlap 1; we discuss this further in Section 7. Although the largest weights may be atypical of real-world tabular data, they demonstrate robustness under increasingly severe violations of the positivity assumption in Definition 3.

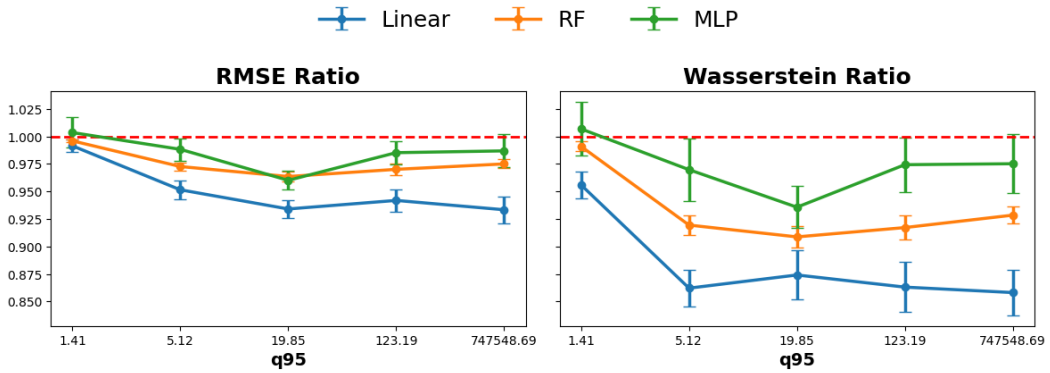


Figure 4: Values below 1 indicate improved performance. We report mean performance ratios of the weighted method relative to the unweighted baseline as a function of weight magnitude (summarised by the 95th percentile of the true weights). Lines correspond to different conditional imputation model types. Error bars denote ± 1.96 SEM, corresponding to approximate 95% confidence intervals. Results are reported for RMSE and Wasserstein distance. The plot is best viewed in colour.

We further study sensitivity to dataset characteristics on the Letter dataset, varying sample size, number of features, and missingness rate. Figure 5 shows that, in 55 of the 57 settings considered, the reported 95% confidence intervals lie below or overlap with 1, indicating that the weighted method does not degrade performance and often improves upon the unweighted baseline.

Sample Size to Dimensionality Ratio: In Appendix C.6 we subsample the datasets considered in order to study performance under varying sample size to dimensionality ratios. The results help clarify the operating regime of importance-weighted imputation: In a low sample-size-to-dimension regime, conditional model estimation is already high-variance even without weighting; our weighted approach can further exacerbate this variance, making the method more sensitive to these finite-sample effects. Conversely, as sample-size-to-dimension grows, conditional models can be stably estimated even under weighting allowing for consistent improvement over unweighted baselines.

Robustness to Model Misspecification: For violations of the MAR assumption, the plots in Appendix C.7.1 show that performance of our method degrades gradually as the probability of missingness depends increasingly on unobserved variables, with no evidence of abrupt failure. Similarly, for the weight misspecification setting, the plots in Appendix C.7.2 show that performance degrades gradually as the missingness mechanism transitions from a linear relationship to a non-linear one, with no evidence of abrupt failure.

Downstream Tasks: Although our main concern is imputation as an end goal, imputation is usually performed as a precursor to a downstream task. In Appendix C.8 we study whether improvements in imputation quality translate into improvements in downstream prediction and correlation matrix recovery. Here we provide a summary of the results.

For downstream prediction, we study regression and classification tasks on the Gas and Letter datasets, respectively, using RMSE and accuracy as performance metrics. We compare downstream XGBoost and linear models fitted after imputation under our weighted and unweighted approaches, along with an unimputed baseline—XGBoost without imputation; linear models with mean imputation and missingness indicators. Our weighted approach offers significant improvement in 2 of 6 simulation setups under a downstream XGB model. However, when an explainable linear model is preferred, our imputation method offers improvement in 5 of 6 setups. Our findings are consistent with recent empirical findings that the benefit of improved imputation is more limited when highly expressive predictive models are used [Le Morvan and Varoquaux, 2023].

For correlation matrix recovery, weighting substantially improves performance across all settings. Suggesting our weighted approach can better capture the structure of the original data. A broader evaluation across downstream tasks is left to future work.

6 Discussion

We briefly highlight connections between the techniques developed in this work and related ideas more broadly in missing data literature. Related work has also connected missing data and covariate shift. Zhou et al. [2023] study domain adaptation under missingness shift and show that the resulting problem can be interpreted as a covariate shift setting when missingness indicators are observed. While the recovered covariate shift problem is closely related to our setting, the starting assumptions and objectives differ between our two works. Zhou et al. [2023] assume a pre-specified shift and focus on downstream prediction tasks, whereas we consider imputation without an explicitly specified shift and show that a distributional mismatch naturally arises from training on observed data. Similarly, Yang et al. [2024] frame a conformal prediction problem under covariate shift as analogous to a missing data problem, enabling the use of tools from missing data theory. This perspective is likewise developed in the context of downstream prediction tasks under an assumed shift structure and therefore differs in objective from our setting. Finally, Inverse probability weighting (IPW) methods for selection bias correction in downstream inference [Robins et al., 1994, Seaman, 2018] are also related to our work, as they adjust for discrepancies between observed and full data distributions in parameter estimation. While we also employ importance weighting to address a similar issue of distributional shift, our focus is on imputation rather than downstream parameter estimation.

7 Limitations

While weighting improves performance in many settings, we observe situations it hinders performance, largely reflecting the bias-variance trade-off associated with importance weighting [Shimodaira, 2000]. Any factor that increases the variance of model estimation exacerbates this trade-off. This can arise from the choice of conditional model itself: as shown in Figure 2, performance can degrade for high-variance conditional models such as MLPs. It can also arise from the sample-size-to-dimensionality ratio: as discussed in Section 5.2, our method can underperform compared to an unweighted approach when the number of samples is small relative to dimensionality, since model estimation is itself more variable in this regime. Practitioners working in high-dimensional, sample-limited settings, or with high-variance conditional models, should consider this trade-off when adopting our approach.

A second, theoretical, limitation exists: general guarantees for round-robin iterative imputation remain an open problem, and existing results are largely restricted to correctly specified conditional models, where the procedure relates closely to Gibbs sampling [Liu et al., 2014]. Our setting targets misspecified models — a regime with no known convergence theory. This limitation is shared by other round-robin imputation methods [van Buuren and Groothuis-Oudshoorn, 2011, Stekhoven and Bühlmann, 2012], whose popularity is driven by empirical performance rather than formal guarantees; we identify convergence theory under misspecification as a direction for future work.

Finally, our theoretical results rest on the MAR assumption and estimation of importance weights. Practitioners should therefore be attentive to weight estimation quality when applying our method. Appendix C.7 shows the method remains competitive under moderate departures from these conditions. Doubly robust extensions [Seaman and Vansteelandt, 2018], which augment the weighted risk in (4) with an auxiliary outcome model, could reduce sensitivity to weight estimation error, and we leave their development for round-robin imputation to future work.

8 Conclusion

We have shown that in risk-based approaches to imputation, where training is necessarily performed on observed data, one must account for a distributional shift between the observed and full data distributions when learning optimal imputation functions. We proposed a novel round-robin iterative imputation algorithm that explicitly accounts for this distributional shift and demonstrated the effectiveness of our approach across a range of experimental settings. Overall, our results suggest that weighted iterative imputation methods are a promising direction for further methodological development. Several avenues for future work remain. First, developing methods that are robust to weak distributional shift remains an open problem. Second, extending the framework beyond point estimation to probabilistic imputation objectives is a natural next step. Finally, our approach could be strengthened by incorporating more sophisticated density ratio estimation techniques, suggesting a promising connection to advances in this area.

Acknowledgments and Disclosure of Funding

Funding. This work was supported by the ProbAI Hub, funded by the UK Engineering and Physical Sciences Research Council (EPSRC). Grant EP/Y028783/1

Competing interests. The authors declare no competing interests.

References

Peter C. Austin and Elizabeth A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679, 2015. doi: https://doi.org/10.1002/sim.6607.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, pages 81–88, 2007. doi: https://doi.org/10.1145/1273496.1273507.

Stephen R. Cole and Miguel A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008. doi: 10.1093/aje/kwn164.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–22, 1977. doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Daniel Jarrett, Bogdan C. Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. HyperImpute: Generalized iterative imputation with automatic model selection. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9916–9937. PMLR, 2022.

Cheng Ju, Joshua Schwab, and Mark J. van der Laan. On adaptive propensity score truncation in causal inference. *Statistics in Medicine*, 38(14):2656–2666, 2019. doi: 10.1002/sim.8116.

Marine Le Morvan and Gaël Varoquaux. Imputation for prediction: Beware of diminishing returns! *Patterns*, 4(1):100671, 2023. doi: 10.1016/j.patter.2022.100671.

Jingchen Liu, Andrew Gelman, Jennifer Hill, Yu-Sung Su, and Jonathan Kropko. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173, 2014. doi: 10.1093/biomet/ast044.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4413–4423, 2019. URL `https://proceedings.mlr.press/v97/mattei19a.html`.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010. URL `http://jmlr.org/papers/v11/mazumder10a.html`.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7130–7140, 2020. URL `https://proceedings.mlr.press/v119/muzellec20a.html`.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2012.

Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012. doi: 10.1177/0962280210386207.

Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998. doi: https://doi.org/10.1093/biomet/85.3.619.

Matthieu Resche-Rigon, Ian R. White, Jonathan W. Bartlett, Sanne A.E. Peters, and Simon G. Thompson. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine*, 32(28):4890–4905, 2013. doi: https://doi.org/10.1002/sim.5894.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994. URL http://www.jstor.org/stable/2290910.

Shaun R. Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical Science*, 33(2):184–197, 2018. doi: 10.1214/18-STS647.

Stijn Seaman, Shaun R.; Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical Science*, 33(2), 2018. doi: 10.1214/18-STS647.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: https://doi.org/10.1016/S0378-3758(00)00115-4.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. doi: https://doi.org/10.1093/bioinformatics/btr597.

Lena Stempfle, Ashkan Panahi, and Fredrik D. Johansson. Sharing pattern submodels for prediction with missing values. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 9882–9890, 2023. doi: 10.1609/aaai.v37i8.26179.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007. URL http://jmlr.org/papers/v8/sugiyama07a.html.

Stef Van Buuren. *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, FL, 2nd edition, 2018. ISBN 9781138588318. URL https://stefvanbuuren.name/fimd/.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: https://doi.org/10.18637/jss.v045.i03.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998. URL https://openlibrary.org/books/OL689998M/Statistical_learning_theory.

Cédric Villani. *Optimal transport – Old and new*, volume 338, chapter 6, page 93. Springer, 2008. doi: https://doi.org/10.1007/978-3-540-71050-9.

Xueying Xu, Leizhen Xia, Qimeng Zhang, Shaoning Wu, Mingcheng Wu, and Hongbo Liu. The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Medical Research Methodology*, 20(1):42, 2020. doi: https://doi.org/10.1186/s12874-020-00932-0.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 86(4):943–965, 2024. doi: 10.1093/jrsssb/qkae009.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5689–5698, 2018. URL https://proceedings.mlr.press/v80/yoon18a.html.

Helen Zhou, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under missingness shift. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 9577–9606. PMLR, 2023. URL https://proceedings.mlr.press/v206/zhou23b.html.

A Theoretical Derivations

A.1 Proof of Lemma 1

Proof. For this derivation, let $\mathbf{1}_{\{A\}}$ denote the indicator function of event A. Then, under MSE loss and a deterministic imputation function Definition 1, our objective simplifies as follows:

$$\begin{aligned}
\mathcal{J}(g) &= \mathbb{E}_{X,R}\Big[\text{MSE}\big\{g(\tilde{X}), X\big\}\Big] \\
&= \mathbb{E}_{X,R}\left[\sum_{i=1}^{d}\big\{g_i(\tilde{X}) - X_i\big\}^2\right] && (7)\\
&= \mathbb{E}_{X,R}\left[\sum_{i=1}^{d}\Big\{R_iX_i + (1-R_i)g_i(\tilde{X}) - X_i\Big\}^2\right] && (8)\\
&= \mathbb{E}_{X,R}\left[\sum_{i=1}^{d}\{g_i(\tilde{X}) - X_i\}^2\mathbf{1}_{\{R_i=0\}}\right] && (9)\\
&= \sum_{i=1}^{d}\mathbb{E}_{X,R}\left[\{g_i(\tilde{X}) - X_i\}^2\mathbf{1}_{\{R_i=0\}}\right] && (10)
\end{aligned}$$

Equation (7) follows by definition of MSE. Equation (8) follows by writing $g_i(\tilde{X})$ as an R_i-weighted sum of the true value X_i and the imputed value $g_i(\tilde{X})$ which holds due to the assumed property that imputation functions preserve already observed coordinates of $\tilde{X}$. Equation (9) by noting that only coordinates with $R_i = 0$ contribute to the loss, and Equation (10) by linearity of expectation.

Focusing on the i^{th} term of the sum, we have:

$$\begin{aligned}
\mathbb{E}_{X,R}\left[\{g_i(\tilde{X}) - X_i\}^2\mathbf{1}_{\{R_i=0\}}\right] &= \mathbb{E}_{R_i}\left(\mathbb{E}_{X,R_{\neg i}|R_i}\left[\Big\{g_i(\tilde{X}) - X_i\Big\}^2\mathbf{1}_{\{R_i=0\}}\right]\right) && (11)\\
&= \mathbb{E}_{R_i}\left(\mathbf{1}_{\{R_i=0\}}\mathbb{E}_{X,R_{\neg i}|R_i}\left[\Big\{g_i(\tilde{X}) - X_i\Big\}^2\right]\right) && (12)\\
&= \sum_{r_i\in\{0,1\}}\mathbf{1}_{\{r_i=0\}}\mathbb{E}_{X,R_{\neg i}|R_i=r_i}\left[\Big\{g_i(\tilde{X}) - X_i\Big\}^2\right]p(R_i=r_i) && (13)\\
&= \mathbb{E}_{X,R_{\neg i}|R_i=0}\left[\Big\{g_i(\tilde{X}) - X_i\Big\}^2\right]\,p(R_i=0) && (14)
\end{aligned}$$

Equation (11) follows by the law of total expectation. Equation (12) follows by pulling $\mathbf{1}_{\{R_i=0\}}$ outside the inner expectation when conditioning on R_i. Equation (13) follows by the definition of expectation over R_i. Equation (14) follows because the $\mathbf{1}_{\{r_i=0\}}$ kills the term in the sum with $r_i = 1$.

Therefore,

$$\mathcal{J}(g) = \mathbb{E}_{X,R}\Big[\text{MSE}\big\{(X_{\text{obs}}, g(\tilde{X})), X\big\}\Big] = \sum_{i=1}^{d}p(R_i=0)\mathbb{E}_{X,R_{\neg i}}\left[\Big\{g_i(\tilde{X}) - X_i\Big\}^2 \mid R_i=0\right]$$

Finally, $\forall i \notin \mathcal{P}$, $p(R_i = 0) = 0$ and hence these terms drop out of the sum yielding

$$\mathcal{J}(g) = \sum_{i\in\mathcal{P}}p(R_i=0)\mathbb{E}_{X,R_{\neg i}}\left[\Big\{g_i(\tilde{X}) - X_i\Big\}^2 \mid R_i=0\right]$$

□

A.2 Proof of Proposition 1

Proof. We aim to rewrite the i^{th}- coordinate risk

$$p(R_i = 0)\mathbb{E}_{X,R_{\neg i}}\left[\left\{g_i(\tilde{X}) - X_i\right\}^2 \mid R_i = 0\right]$$

in terms of observed data only. Writing $w_i(x_{\text{obs}}, r_{\neg i}) := \frac{p(x_{\text{obs}}, r_{\neg i} | R_i=0)}{p(x_{\text{obs}}, r_{\neg i} | R_i=1)}$ and ignoring the $p(R_i = 0)$ term for now, we obtain the following,

$$\begin{aligned}
&\mathbb{E}_{X,R_{\neg i}|R_i=0}\left[\{g_i(\tilde{X}) - X_i\}^2\right] \\
&= \mathbb{E}_{X_{\text{obs}},R_{\neg i}|R_i=0}\Big(\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},R_{\neg i},R_i=0}\left[\{g_i(\tilde{X}) - X_i\}^2\right]\Big) && (15)\\
&= \mathbb{E}_{X_{\text{obs}},R_{\neg i}|R_i=0}\Big(\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},R_{\neg i},R_i=1}\left[\{g_i(\tilde{X}) - X_i\}^2\right]\Big) && (16)\\
&= \int \mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},R_{\neg i},R_i=1}\left[\{g_i(\tilde{X}) - X_i\}^2\right] p(x_{\text{obs}}, r_{\neg i} \mid R_i = 0)\, dx_{\text{obs}}\, dr_{\neg i} && (17)\\
&= \int \mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},R_{\neg i},R_i=1}\left[\{g_i(\tilde{X}) - X_i\}^2\right] p(x_{\text{obs}}, r_{\neg i} \mid R_i = 1)\, w_i(x_{\text{obs}}, r_{\neg i})\, dx_{\text{obs}}\, dr_{\neg i} && (18)\\
&= \mathbb{E}_{X_{\text{obs}},R_{\neg i}|R_i=1}\Big(\mathbb{E}_{X_{\text{miss}}|X_{\text{obs}},R_{\neg i},R_i=1}\left[w_i(x_{\text{obs}}, r_{\neg i})\{g_i(\tilde{X}) - X_i\}^2\right]\Big) && (19)\\
&= \mathbb{E}_{X_{\text{obs}},R_{\neg i}|R_i=1}\left[w_i(X_{\text{obs}}, R_{\neg i})(g_i(\tilde{X}) - X_i)^2\right] && (20)
\end{aligned}$$

Equation (15) follows by the law of total expectation. Equation (16) follows from the MAR assumption $R \perp\!\!\!\perp X_{\text{miss}} \mid X_{\text{obs}}$, which implies that the conditional distribution of X_{miss} does not depend on R_i, and we may replace $R_i = 0$ by $R_i = 1$ without altering the inner expectation. Equation (17) follows by writing out explicitly the outer expectation. Equation (18) follows by factoring $p(x_{\text{obs}}, r_{\neg i} \mid R_i = 0) = p(x_{\text{obs}}, r_{\neg i} \mid R_i = 1)\, w_i(x_{\text{obs}}, r_{\neg i})$. Equation (19) follows by bringing the weighting inside the inner expectation and recognising that the outer expectation is now over $(X_{\text{obs}}, R_{\neg i}) \mid R_i = 1$. Equation (20) follows by iterated expectation.

It remains to verify that the weights are well-defined. Points with $p(X_{\text{obs}}, R_{\neg i} \mid R_i = 0) = 0$ do not contribute to the expectation and can be ignored. For points where $p(X_{\text{obs}}, R_{\neg i} \mid R_i = 0) > 0$, Bayes' rule gives

$$p(X_{\text{obs}}, R_{\neg i} \mid R_i = 1) = \frac{p(R_i = 1 \mid X_{\text{obs}}, R_{\neg i})\, p(X_{\text{obs}}, R_{\neg i})}{p(R_i = 1)} > 0$$

by the positivity assumption. Hence,

$$p(X_{\text{obs}}, R_{\neg i} \mid R_i = 1) > 0 \quad \text{whenever} \quad p(X_{\text{obs}}, R_{\neg i} \mid R_i = 0) > 0,$$

and the weights $w_i(X_{\text{obs}}, R_{\neg i})$ are well-defined.

□

Remark 2. Equation (16) highlights the source of covariate shift induced by the MAR mechanism. In practice, we may train a model to predict X_i by minimising squared error over samples for which $R_i = 1$. This corresponds to the inner expectation. However, the quantity of interest is the performance of our learned model on the population where $R_i = 0$, corresponding to the outer expectation. In general, the distributions of $(X_{\text{obs}}, R_{\neg i}) \mid R_i = 1$ and $(X_{\text{obs}}, R_{\neg i}) \mid R_i = 0$ differ, leading to a covariate shift between the training and target populations.

If this shift is not accounted for, the resulting estimator may be biased. The weighting function $w_i(X_{\text{obs}}, R_{\neg i})$ corrects for this discrepancy by re-weighting the observed-data distribution to match the target distribution.

B Further theoretical results

In this section, we derive theoretical results that help explain the empirical patterns observed in Section 5. In particular, we investigate when the optimal solutions learned under the subpopulations $R_i = 0$ and $R_i = 1$ coincide. We consider risk minimisation for coordinate i for which we are choosing an optimal imputation function from the model class $\mathcal{G}_i$.

For this section, let:

- $\mathcal{J}_i^{(r)}(g_i) = \mathbb{E}\Big[\big(g_i(\tilde{X}) - X_i\big)^2 \,\big|\, R_i = r\Big], \quad r \in \{0,1\}$, be the optimisation problems corresponding to coordinate i under $R_i = r$
- $g_i^{*,(r)} = \underset{g \in \mathcal{G}_i}{\arg\min} \Big\{\mathcal{J}_i^{(r)}(g)\Big\}$ be the loss minimising imputation function for column i chosen from a class $\mathcal{G}_i$ under $R_i = r$
- $f_i(\tilde{X}) = \mathbb{E}[X_i \mid \tilde{X}] = \mathbb{E}[X_i \mid X_{\text{obs}}]$ be the theoretically optimal imputation function. We note that, under MAR, this solution is optimal under $R_i = 1$ and $R_i = 0$ and, in actuality, depends only on X_{obs}.
- $\varepsilon^2_{i,(r)} := \min_{g_i \in \mathcal{G}_i} \mathbb{E}\Big[\big(g_i(\tilde{X}) - f_i(\tilde{X})\big)^2 \,\big|\, R_i = r\Big]$ be the approximation error of $\mathcal{G}_i$ under $R_i = r$

In the above definitions $\mathcal{J}_i^{(0)}(g_i)$ is the true coordinate-wise risk of coordinate g_i as defined in 3, whereas $\mathcal{J}_i^{(1)}(g_i)$ represents the unweighted objective minimised by existing unweighted imputation algorithms.

B.0.1 Agreement of population-optimal imputation functions under covariate shift

Proposition 2 (Distance between population minimisers under covariate shift)**.** *Suppose that:*

- $\exists \kappa > 0$ *such that* $\frac{dP(\tilde{X}|R_i=0)}{dP(\tilde{X}|R_i=1)} \leq \kappa^{-1}$ *almost surely (overlap condition).*
- *All functions* $g_i \in \mathcal{G}_i$ *belong to the intersection*

$$g_i \in L^2(\tilde{X} \mid R_i = 0) \cap L^2(\tilde{X} \mid R_i = 1),$$

i.e., they have finite second moments under both conditional distributions $P(\tilde{X} \mid R_i = 0)$ *and* $P(\tilde{X} \mid R_i = 1)$*. This ensures that all terms in the proof are well-defined and that we can apply results such as the triangle inequality in* L^2 *and have an intepretation of distance.*

Then the population minimisers satisfy

$$\mathbb{E}\Big[\big(g_i^{*,(0)}(\tilde{X}) - g_i^{*,(1)}(\tilde{X})\big)^2 \,\big|\, R_i = 0\Big] \leq \big(\varepsilon_{i,(0)} + \kappa^{-1/2}\varepsilon_{i,(1)}\big)^2.$$

Proof. For any $g_i \in \mathcal{G}_i$ and $r \in \{0,1\}$, by the bias-variance decomposition gives

$$\mathbb{E}\big[(g_i(\tilde{X}) - X_i)^2 \mid R_i = r\big] = \mathbb{E}\big[(g_i(\tilde{X}) - f(\tilde{X}))^2 \mid R_i = r\big] + \mathbb{E}\big[(X_i - f(\tilde{X}))^2 \mid R_i = r\big] \quad (21)$$

Since $\mathbb{E}\big[(X_i - f(\tilde{X}))^2 \mid R_i = r\big]$ does not depend on g_i, the minimiser, $g_i^{*,(r)}$, of the left hand side of 21 must also be the minimiser of $\mathbb{E}\big[(g_i(\tilde{X}) - f(\tilde{X}))^2 \mid R_i = r\big]$. Therefore,

$$\mathbb{E}\big[(g_i^{*,(r)}(\tilde{X}) - f(\tilde{X}))^2 \mid R_i = r\big] = \varepsilon^2_{i,(r)}$$

Now, consider the L^2 distance between the optimal functions under $R_i = 0$ and $R_i = 1$:

$$\begin{aligned}&\mathbb{E}\big[(g_i^{*,(0)}(\tilde{X}) - g_i^{*,(1)}(\tilde{X}))^2 \mid R_i = 0\big]^{1/2} \\ &\leq \mathbb{E}\big[(g_i^{*,(0)}(\tilde{X}) - f_i(\tilde{X}))^2 \mid R_i = 0\big]^{1/2} + \mathbb{E}\big[(g_i^{*,(1)}(\tilde{X}) - f_i(\tilde{X}))^2 \mid R_i = 0\big]^{1/2} \quad (22)\end{aligned}$$

by the triangle inequality in L^2.

The first term is exactly $\varepsilon_{i,(0)}$. For the second term, the triangle inequality gives us:

$$\mathbb{E}\big[(g_i^{*,(1)}(\tilde{X}) - f(\tilde{X}))^2 \mid R_i = 0\big] = \int (g^{*,(1)}(\tilde{x}) - f(\tilde{x}))^2 \, dP(\tilde{x} \mid R_i = 0)$$
$$\leq \kappa^{-1} \int (g^{*,(1)}(\tilde{x}) - f(\tilde{x}))^2 \, dP(\tilde{x} \mid R_i = 1) = \kappa^{-1} \varepsilon_{i,(1)}^2 \tag{23}$$

Combining (22) and (23), and squaring both sides gives the stated bound:

$$\mathbb{E}\big[(g^{*,(0)}(\tilde{X}) - g^{*,(1)}(\tilde{X}))^2 \mid R_i = 0\big] \leq \big(\varepsilon_{i,(0)} + \kappa^{-1/2} \varepsilon_{i,(1)}\big)^2.$$

□

Corollary 1. *[Optimisation problems are equivalent for well specified models] Suppose that $f_i \in \mathcal{G}_i$ Then $g^{*,(0)}(\tilde{X}) = g^{*,(1)}(\tilde{X})$*

Proof. If $f_i \in \mathcal{G}_i$, then $\varepsilon_0 = \varepsilon_1 = 0$. The result follows directly from Proposition 2.

□

Remark: Proposition 2 shows that as the expressive power of the model class $\mathcal{G}_i$ increases, and hence the approximation errors ε_0 and ε_1 decrease, the population-optimal solutions under $R_i = 0$ and $R_i = 1$ converge. This provides a theoretical explanation for the empirical observation that our method yields larger improvements for simpler model classes, such as linear conditional models, while the gains diminish for more expressive models such as MLPs and random forests, where approximation error is already small.

B.0.2 Agreement of population-optimal imputation functions under MCAR missingness

Proposition 3 (Equivalence of optimisation problems under MCAR)**.** *Suppose that the data satisfy the MCAR assumption. Then,*

$$g_i^{*,(0)}(\tilde{X}) = g_i^{*,(1)}(\tilde{X}).$$

Proof. As MCAR is a special case of MAR, our derived optimal weights in (5) are still valid. As argued in section 3.2, $R_{\neg i}$ can be ignored in this weighting yielding

$$w_i(X_{\text{obs}}) = \frac{p(X_{\text{obs}} \mid R_i = 0)}{p(X_{\text{obs}} \mid R_i = 1)}.$$

Under the MCAR assumption, the distribution of X_{obs} does not depend on R_i, and hence

$$p(X_{\text{obs}} \mid R_i = 0) = p(X_{\text{obs}} \mid R_i = 1),$$

so that $w_i(X_{\text{obs}}) \equiv 1$ almost surely.

Consequently, the objectives $\mathcal{J}_i^{(0)}(g)$ and $\mathcal{J}_i^{(1)}(g)$ coincide, and therefore their population minimisers are identical. □

C Supplementary Experimental Results

C.1 Dataset Metadata and Preprocessing

Table 1: Summary of datasets used in experiments with dataset URLs

Dataset	Rows	Columns	Data type	URL
Blood	748	3	Tabular	https://doi.org/10.24432/C5GS39.
Airfoil	1503	6	Tabular	https://doi.org/10.24432/C5VW2C
California	20640	8	Tabular	https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html
Letter	20000	16	Tabular	https://doi.org/10.24432/C5ZP40
Gas	416153	9	Time series	https://doi.org/10.24432/C5762W
Metro	1516948	15	Time series	https://doi.org/10.24432/C5VW3R
Motor Acceleration	511806	57	Time series	https://doi.org/10.18419/DARUS-3301
MNIST	70000	196	Image	http://yann.lecun.com/exdb/mnist/

Table 1 gives a brief summary of the datasets used in our study. Some preprocessing was applied to datasets where necessary.

- **Blood:** Remove one perfectly co-linear column.
- **California:** Capped columns so that skewness lies in the range $[-10, 10]$
- **Motor Acceleration:** Removed unary columns and capped columns so that skewness lies in the range $[-10, 10]$
- **MNIST:** To reduce computational overhead while preserving the salient digit structure, images were downsampled from 28×28 to 14×14 for the experiments.

C.2 Non-Iterative Imputation Method Comparison

Beyond the experimental results provided in the Section 5.2, here we provide a comparison of our algorithm against a wider range of state-of-the-art imputation methods. In particular we consider

- **HyperImpute**: An iterative imputation method that performs a principled model search at each iteration to choose the best performing conditional model. [Jarrett et al., 2022]
- **GAIN**: A GAN-based approach in which a generator imputes missing entries and a discriminator distinguishes them from observed entries. [Yoon et al., 2018]
- **Sinkhorn**: An approach that minimises an optimal transport divergence between distributions of observed and imputed entries. [Muzellec et al., 2020]
- **SoftImpute**: An approach that performs low-rank matrix completion using nuclear-norm regularization with iterative soft-thresholded SVD.[Mazumder et al., 2010]

The results of this wider study are shown in Figure 3.

C.3 Runtimes

Table 2 summarises the runtimes for our experiments.

Across all experiments, we consistently observe improved runtime efficiency for linear models when using our weighted approach. When aggregated across all datasets, the weighted linear method runs in approximately 47% of the time of the corresponding unweighted approach, demonstrating a substantial computational advantage. We note that HyperImpute runtimes are not directly comparable to the our approach under linear conditional models as HyperImpute performs hyperparameter search at each iteration, incurring overhead beyond base model fitting that accounts for its longer runtimes.

For both RF and MLP based models, we observe relatively larger proportional increases in runtime on smaller datasets. However, these differences are small in absolute terms, typically amounting to only a few seconds, and therefore have limited practical impact.

Importantly, as dataset size increases (from the Gas dataset onwards), runtimes between weighted and unweighted approaches become more comparable, showing our proposed method scales well without introducing prohibitive computational overhead.

At the overall level, the weighted variants are approximately **12.5% slower for RF models** and **11.3% slower for MLP models**. Despite this modest overhead, these results demonstrate that a weighted iterative imputation approach can be deployed in practice without requiring expensive additional estimation or tuning of weighting parameters, and without introducing significant computational burden.

Table 2: Mean runtime (seconds) across datasets with overall mean highlighted in bold.

		Blood	Airfoil	California	Letter	Gas	Metro	MNIST	Motor Acc.	**Overall**
Linear	Weighted	1.26	0.25	0.42	0.69	2.28	11.91	42.76	17.27	**9.11**
	Unweighted	4.70	9.67	1.53	1.73	5.16	27.23	57.25	51.08	**19.30**
RF	Weighted	4.53	14.85	25.23	21.22	112.26	385.59	224.17	891.74	**210.45**
	Unweighted	0.29	2.13	7.69	5.03	93.26	415.20	51.79	926.87	**187.00**
MLP	Weighted	3.19	15.66	74.99	90.31	826.31	1603.54	344.30	852.64	**476.37**
	Unweighted	0.28	4.22	60.05	69.27	815.18	1581.80	203.27	792.08	**428.02**

C.4 Gamma Tuning

A naive implementation of our algorithm requires re-estimating weights and tuning the tempering parameter γ at each iteration, via grid search with cross-validation. This can lead to a substantial increase in computational cost on top of the already iterative nature of the imputation procedure. Careful handling of γ is therefore essential to ensure tractable runtime.

We propose a simple guiding principle: **as the expressivity of the conditional model increases, γ can be tuned over a coarser grid and updated less frequently**. This is supported by the theoretical results in Appendix B.0.1, which show that the impact of importance weighting diminishes as model expressivity increases. Consequently, less computational effort needs to be spent on precisely tuning γ in these settings.

In practice, we adopt a two-stage strategy: (i) an initial grid search to select γ, followed by (ii) local refinement via warm-starting in subsequent iterations.

Initial grid search: We select the initial γ over a model-dependent grid reflecting the expressivity of the conditional model. Less expressive models require finer tuning, while more expressive models permit coarser grids.

Warm starting: After the initial selection, γ is updated, in successive iterations, within a small neighbourhood of the previous value. Specifically, at iteration t, we search over

$$\{\gamma^{(t-1)} - 0.05,\ \gamma^{(t-1)},\ \gamma^{(t-1)} + 0.05\},$$

where $\gamma^{(t-1)}$ is the gamma value previously selected for our column of interest. This approach significantly reduces computational cost while ensuring stable convergence.

Weight update frequency: We are not required to update weights at every iteration. For linear models, weights and γ are updated at every iteration. For more expressive models (RF and MLP), we reduce update frequency: an initial unweighted pass is performed, followed by two iterations where weights estimates are updated and γ is tuned, after which both are fixed for the remaining iterations.

To support this approach, we analyse the average change in weights over iterations. The results in Table 4 show that the average change in weights stabilise rapidly after the first iteration, suggesting that further weight updates provide only marginal benefit relative to the cost.

The specific configurations used in our experiments are summarised in Table 3.

Table 3: Summary of γ tuning strategy across model classes.

Model	Initial γ Grid	Warm Start	Update Strategy
Linear	$\{0, 0.1, \ldots, 1\}$	± 0.05	Every iteration
RF	$\{0, 0.5, 1\}$	± 0.05	Iterations 2–3, then fixed
MLP	$\{0, 0.25, 0.5, 0.75, 1\}$	± 0.05	Iterations 2–3, then fixed

Table 4: Average change in estimated weights between successive iterations (mean $\pm$ SEM deviation) for the letter dataset under linear conditioanl imputation models.

Method	$0 \to 1$	$1 \to 2$	$2 \to 3$	$3 \to 4$	$4 \to 5$
Linear	0.42 ± 0.05	0.03 ± 0.01	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
RF	0.38 ± 0.04	0.02 ± 0.01	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
MLP	0.46 ± 0.06	0.04 ± 0.01	0.01 ± 0.01	0.00 ± 0.00	0.00 ± 0.00

C.5 Sensitivity Analysis

To assess the sensitivity of our method to different data characteristics, we evaluate performance under varying data sizes, feature dimensionality, and missingness rates. Data size is controlled by subsampling rows from the original dataset, while feature count is varied by subsampling columns. Missingness is adjusted by tuning β_i to achieve the desired missingness level as described in Section 5. Figure 5 shows that our method remains robust across the range of scenarios considered, consistently outperforming the unweighted baseline.

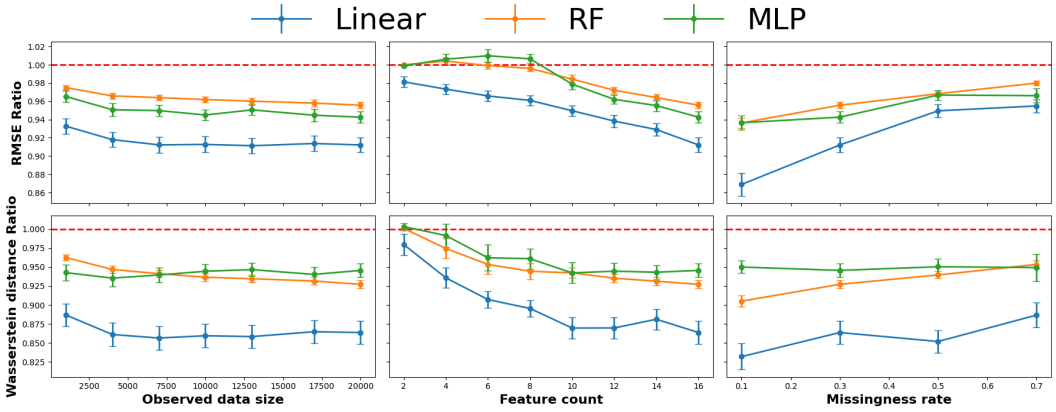


Figure 5: Values below 1 indicate improved performance. We report mean performance ratios of the weighted method relative to its unweighted counterpart across varying dataset conditions and conditional model types. Lines denote mean ratios across dataset settings, with error bars given by ± 1.96 SEM, corresponding to approximate 95% confidence intervals. Results are reported for RMSE and Wasserstein distance across data variation regimes (e.g., sample size, number of features, and missingness rate). The plot is best viewed in colour.

C.6 Sample Size to Dimensionality Ratio

To investigate behaviour under varying sample-size-to-dimension ratio, we subsample each dataset to $a * p$ observations where p is the number of variables in the dataset and a controls the number of observations per variable.

Tables 5 and 6 show relative improvement in RMSE and WD for our weighted approach over an unweighted baseline for linear conditional models. We observe that for small values of a, corresponding to low sample-size-to-dimension settings, our proposed method can underperform the unweighted baseline. Conversely, as a increases, we begine to observe consistent improvement over the unweighted baseline for all but RMSE on the Motor Acceleration dataset. Motor Acceleration requires larger values of a before we observe such improvements. Figure 2 illustrates that, when

Table 5: RMSE ratio of WI-Ridge relative to IterBayesRidge, reported as mean ± SEM.

Dataset	$a = 4$	$a = 8$	$a = 16$	$a = 32$	$a = 64$	$a = 128$	$a = 256$
Blood	0.99 ± 0.06	0.93 ± 0.03	0.87 ± 0.02	0.86 ± 0.02	0.85 ± 0.02	0.84 ± 0.02	0.82 ± 0.01
Airfoil	0.89 ± 0.03	0.88 ± 0.02	0.89 ± 0.02	0.87 ± 0.01	0.86 ± 0.01	0.93 ± 0.01	0.93 ± 0.01
California	1.09 ± 0.03	1.04 ± 0.02	1.01 ± 0.01	1.01 ± 0.01	1.02 ± 0.01	1.01 ± 0.01	0.98 ± 0.01
Letter	1.04 ± 0.01	1.02 ± 0.01	0.98 ± 0.01	0.95 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.91 ± 0.01
Gas	1.08 ± 0.03	0.96 ± 0.02	0.95 ± 0.01	0.96 ± 0.01	0.92 ± 0.01	0.90 ± 0.01	0.89 ± 0.01
Metro	0.94 ± 0.04	0.93 ± 0.03	0.93 ± 0.03	0.88 ± 0.02	0.89 ± 0.02	0.91 ± 0.01	0.91 ± 0.01
MNIST	1.24 ± 0.02	1.12 ± 0.01	1.05 ± 0.01	1.01 ± 0.01	0.98 ± 0.01	0.96 ± 0.01	0.94 ± 0.02
Motor Acceleration	4.87 ± 1.70	1.83 ± 0.34	1.58 ± 0.30	1.50 ± 0.25	1.30 ± 0.20	1.30 ± 0.19	1.34 ± 0.19

Table 6: Wasserstein distance (WD) ratio of WI-Ridge relative to IterBayesRidge, reported as mean ± SEM.

Dataset	$a = 4$	$a = 8$	$a = 16$	$a = 32$	$a = 64$	$a = 128$	$a = 256$
Blood	0.96 ± 0.07	0.92 ± 0.04	0.85 ± 0.03	0.87 ± 0.02	0.89 ± 0.02	0.92 ± 0.02	0.93 ± 0.02
Airfoil	0.76 ± 0.04	0.72 ± 0.03	0.72 ± 0.02	0.71 ± 0.02	0.71 ± 0.02	0.78 ± 0.02	0.78 ± 0.02
California	0.93 ± 0.04	0.89 ± 0.05	0.83 ± 0.02	0.87 ± 0.02	0.88 ± 0.01	0.87 ± 0.01	0.84 ± 0.01
Letter	0.88 ± 0.02	0.89 ± 0.01	0.91 ± 0.01	0.88 ± 0.01	0.87 ± 0.01	0.88 ± 0.01	0.86 ± 0.01
Gas	1.07 ± 0.03	0.96 ± 0.02	0.93 ± 0.02	0.91 ± 0.02	0.80 ± 0.02	0.76 ± 0.02	0.71 ± 0.02
Metro	0.93 ± 0.06	0.84 ± 0.03	0.83 ± 0.04	0.79 ± 0.02	0.77 ± 0.02	0.79 ± 0.02	0.79 ± 0.02
MNIST	0.86 ± 0.02	0.89 ± 0.01	0.93 ± 0.01	0.95 ± 0.01	0.95 ± 0.00	0.94 ± 0.01	0.93 ± 0.02
Motor Acceleration	4.52 ± 1.40	1.81 ± 0.33	1.43 ± 0.27	1.20 ± 0.14	1.01 ± 0.10	0.94 ± 0.09	0.89 ± 0.09

using the full dataset (corresponding to the largest effective value of a), we also observe improvement for the Motor Acceleration dataset.

C.7 Robustness to Model and Mechanism Misspecification

In this section, we investigate how performance degrades under progressively stronger violations of the MAR assumption and under misspecification of the weighting model.

We consider data generated according to the following process:

$$\begin{aligned}
&X_1 \sim \mathcal{N}(0,1),\\
&X_2 \sim \mathcal{N}(0,1),\\
&X_3 = X_1 X_2 + \epsilon_3, \quad \epsilon_3 \sim \mathcal{N}(0,1),\\
&X_4 = 0.5X_1^2 + X_2 + \epsilon_4, \quad \epsilon_4 \sim \mathcal{N}(0,1),\\
&X_5 = X_1 X_3 + X_2^2 + \epsilon_5, \quad \epsilon_5 \sim \mathcal{N}(0,1).
\end{aligned}$$

C.7.1 MNAR

To simulate MNAR mechanisms, we introduce missingness in X_4 and X_5 according to

$$P(R_j = 1 \mid X) = \sigma(c + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3), \quad j \in \{4, 5\},$$

where $\sigma(\cdot)$ is the logistic function and c is calibrated to control the overall missingness rate.

Simulation setup 1 (missingness driven by X_1): We fix $\beta_2 = \beta_3 = 2$ and vary β_1 over the range $[0, 10]$. After generating missingness, the driver variable X_1 is removed from the partially observed dataset. This yields an MNAR setting in which missingness depends on an unobserved variable, with the strength of this dependence controlled by β_1.

Figure 6 reports the ratio of weighted to unweighted performance for linear conditional imputation models. When $\beta_1 = 0$, the mechanism reduces to MAR, as missingness depends only on observed variables. As β_1 increases, the degree of MNAR violation increases.

The vertical line indicates the value $\beta_1^\star$ at which the majority of variance in the missingness mechanism is explained by the unobserved variable X_1, rather than the observed covariates. As β_1 increases, we observe a smooth degradation in performance for RMSE, while improvements in Wasserstein distance remain relatively stable across a wide range of MNAR settings.

We note that large values such as $\beta_1 = 10$ are included for illustrative purposes to probe extreme regimes. Overall, these results suggest that our proposed approach continues to offer improvement provided that the primary sources of missingness dependence are captured within the observed data.

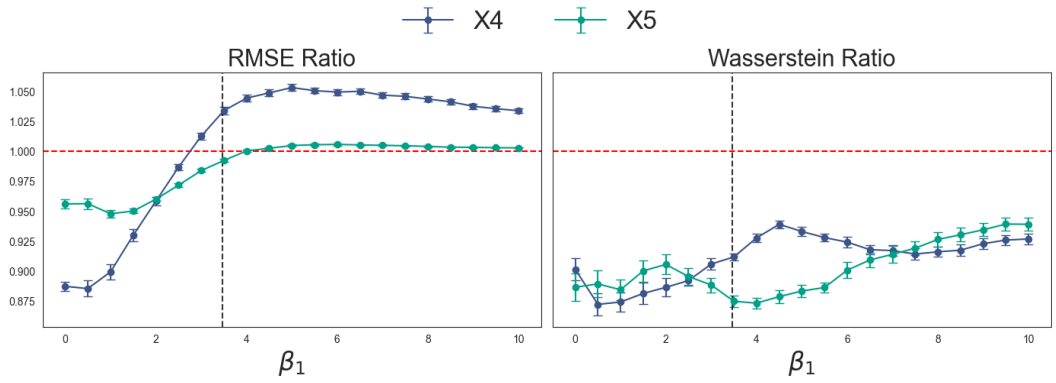


Figure 6: Values below 1 indicate improved performance. Mean performance ratios of the weighted method relative to its unweighted counterpart are shown for linear conditional models. Lines denote mean ratios, with error bars given by ± 1.96 SEM, corresponding to approximate 95% confidence intervals. The vertical line indicates the value $\beta_1^\star$, at which the majority of variance in the missingness mechanism is explained by the unobserved variable X_1 rather than the observed covariates. The plot is best viewed in colour.

Simulation setups 2 & 3 (missingness driven by X_2 and X_3): We repeat the above procedure for two additional settings. In the first, missingness is driven by X_2, after which X_2 is removed from the dataset prior to imputation. In the second, missingness is driven by X_3, and X_3 is similarly removed. In both cases, we vary the corresponding coefficient to control the strength of MNAR dependence. The results are shown in Figures 7 and 8. These experiments exhibit qualitatively similar behaviour. In particular, performance degrades gradually as the strength of MNAR dependence increases, rather than failing abruptly. Moreover, we typically continue to observe improvements over the unweighted baseline in regimes where the dominant drivers of missingness are observed. This suggests that our proposed method shows improvement provided that the primary sources of missingness dependence are captured within the observed data.

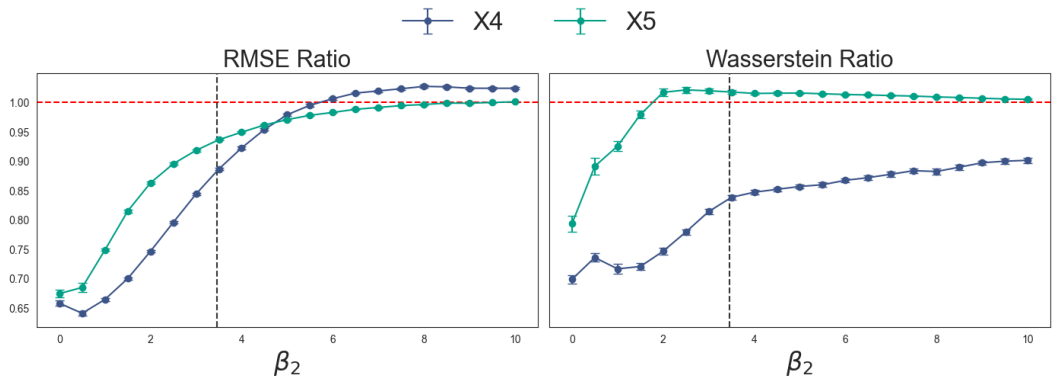


Figure 7: Values below 1 indicate improved performance. Mean performance ratios of the weighted method relative to its unweighted counterpart are shown for linear conditional models. Lines denote mean ratios, with error bars given by ± 1.96 SEM, corresponding to approximate 95% confidence intervals. The vertical line indicates the value $\beta_2^\star$, at which the majority of variance in the missingness mechanism is explained by the unobserved variable X_2 rather than the observed covariates. The plot is best viewed in colour.

C.7.2 Misspecified Weights Model

We next assess robustness to misspecification of the weight model. In contrast to the MNAR experiments above, we retain a MAR mechanism but introduce controlled misspecification in the functional form used to estimate the importance weights.

Simulation setup: We generate data according to the same base process as above. Missingness in X_4 and X_5 is then introduced via logistic models depending on nonlinear transformations of the

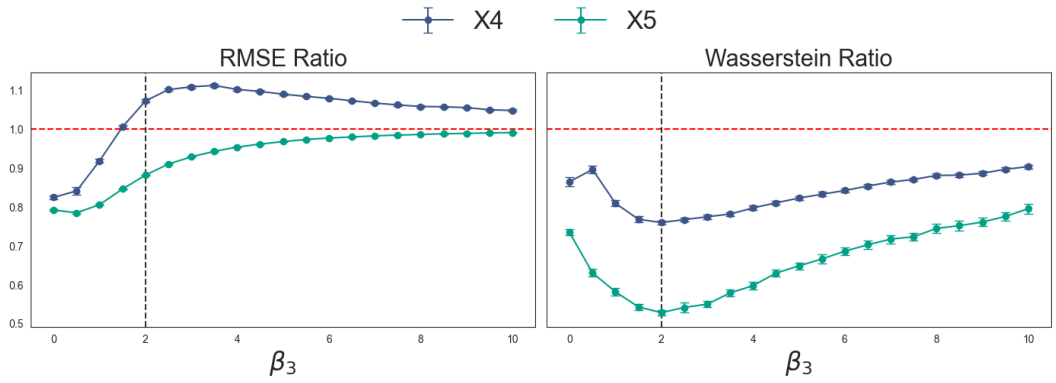


Figure 8: Values below 1 indicate improved performance. Mean performance ratios of the weighted method relative to its unweighted counterpart are shown for linear conditional models. Lines denote mean ratios, with error bars given by ± 1.96 SEM, corresponding to approximate 95% confidence intervals. The vertical line indicates the value $\beta_3^\star$, at which the majority of variance in the missingness mechanism is explained by the unobserved variable X_3 rather than the observed covariates. The plot is best viewed in colour.

covariates:

$$P(R_j = 1 \mid X) = \sigma\Big(c + \beta_1 \tilde{X}_1 + \beta_2 \tilde{X}_2 + \beta_3 \tilde{X}_3\Big), \quad j \in \{4, 5\},$$

where $\tilde{X}_k$ denotes a nonlinear transformation of X_k defined as

$$\tilde{X}_k = (1 - \eta) X_k + \eta\, h(X_k), \quad k \in \{1, 2, 3\},$$

with

$$h(x) = \tfrac{1}{3}x^3 + 0.5x^2 - x - 1.$$

The parameter $\eta \in [0, 2]$ controls the degree of nonlinearity in the missingness mechanism. When $\eta = 0$, the mechanism is linear in (X_1, X_2, X_3) and therefore correctly specified by a standard logistic model. As η increases, the mechanism becomes increasingly nonlinear, inducing progressive misspecification when weights are estimated using a linear model.

Evaluation: We estimate importance weights using a logistic model linear in (X_1, X_2, X_3), thereby introducing misspecification whenever $\eta > 0$. Performance is again evaluated using RMSE and Wasserstein distance, reported as ratios relative to the unweighted baseline.

Figure 9 shows that performance degrades smoothly as η increases. In particular, the proposed method continues to provide improvements over the unweighted baseline under mild to moderate misspecification. This suggests that our method is robust to moderate departures from correct weight specification.

Abalation Study: Beyond the study in C.7.2, we conducted a further ablation study. In this study we introduce missingness in the letter dataset under the same procedure as described in Section 5.1 but with the linear logistic regression mechanism replaced with a bi-modal, non linear mechanism. Specifically, using the notation established in Section 5.1, for each $i \in \mathcal{M}$, missingness is introduced according to

$$\mathbb{P}(R_i = 1 \mid X) = \sigma\left(\alpha \sum_{j \in \mathcal{C}_i} \tilde{X}_j + \beta_i\right)$$

where

$$\tilde{X}_k = (1 - \eta) X_k + \eta\, q(X_k), \quad k \in \mathcal{C}_i,$$

with

$$q(x) = -3(x^2 - 1)^2$$

The parameter $\eta \in [0, 1]$ controls the degree of nonlinearity in the missingness mechanism. The resulting dataset is then imputed under three different mechanisms

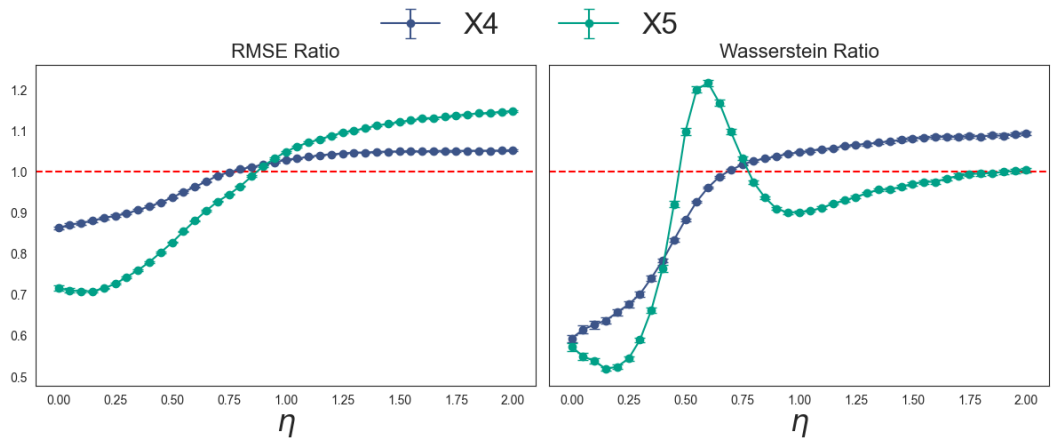


Figure 9: Values below 1 indicate improved performance. Mean performance ratios of the weighted method relative to its unweighted counterpart are shown for linear conditional models. Lines denote mean ratios across dataset settings, with error bars given by ± 1.96 SEM, corresponding to approximate 95% confidence intervals. The plot is best viewed in colour.

1. **WI-Linear**: Our weighted imputation approach introduced in Section 4 under linear conditional imputation models.
2. **WI-Linear-Oracle**: An adaptation of our weighted imputation approach introduced in Section 4 under linear conditional imputation models using the true oracle weights at each stage of the algorithm without capping and tempering.
3. **WI-Linear-Oracle-Transform**: An adaptation of our weighted imputation approach introduced in Section 4 under linear conditional imputation models using the true oracle weights at each stage of the algorithm including capping and tempering at each stage.

Table 7 shows the results of this study relative to an unweighted baseline so that so that values < 1 show an improvement over an otherwise identical unweighted approach. We see that our proposed method continues to outperform the unweighted baseline across all levels of misspecification considered suggesting our proposed method is robust to misspecification in the weights model. We observe that clipped and tempered oracle weights consistently outperform the corresponding raw oracle weights showing that more accurate importance weights do not necessarily yield better finite-sample performance and provides further support to our choice to clip and temper weights in our implementation of our algorithm in Section 4

Table 7: Imputation RMSE on the letter dataset under linear conditional imputation models for varying levels of nonlinearity η. Values are reported as mean $\pm$ 1.96 SEM.

Method	$\eta = 0$	$\eta = 0.25$	$\eta = 0.5$	$\eta = 0.75$	$\eta = 1$
WI-Linear	0.902 ± 0.022	0.978 ± 0.007	0.983 ± 0.005	0.983 ± 0.004	0.982 ± 0.004
WI-Linear-Oracle	0.919 ± 0.023	0.973 ± 0.011	0.978 ± 0.012	0.984 ± 0.013	0.995 ± 0.013
WI-Linear-Oracle-Transform	0.911 ± 0.017	0.962 ± 0.008	0.964 ± 0.007	0.962 ± 0.008	0.961 ± 0.008

C.8 Downstream Application

We evaluate downstream performance across three tasks: downstream regression, downstream classification, and correlation matrix recovery. Below, we describe the experimental setup and results for each task.

C.8.1 Downstream Regression

We evaluate downstream regression on the **Gas** dataset, where **COValue** is held out as the prediction target. Missingness is introduced into the remaining features following the MAR simulation procedure outlined in Section 5.1. Each dataset is then imputed using the competing methods.

For each imputed dataset, we train tuned Ridge and XGBoost regression models to predict the target variable. We additionally include an "unimputed" baseline for comparison. For XGBoost, this corresponds to using the raw data directly, as the model can natively handle missing values. For linear models, we use mean imputation and augment inputs with missingness indicator variables.

We evaluate performance using RMSE. The results are shown in Figure 10.

Downstream XGBoost regressor: For linear imputation methods, unweighted imputation can harm downstream performance, whereas our weighted approach improves it. For RF-based imputation, we observe improvements over the unimputed baseline, but insignificant difference from weighting. For MLP-based imputation, our weighted approach yields substantial improvements over both the unweighted and unimputed baselines.

Downstream Ridge regressor: For linear imputation methods, our weighted approach significantly improves performance over the unweighted variant, although both are comparable to the unimputed baseline. For RF-based imputation, we again observe improvements over the unimputed baseline, with minimal difference between weighted and unweighted methods. For MLP-based imputation, unweighted imputation can degrade performance relative to the unimputed baseline, whereas our weighted method yields substantial improvements.

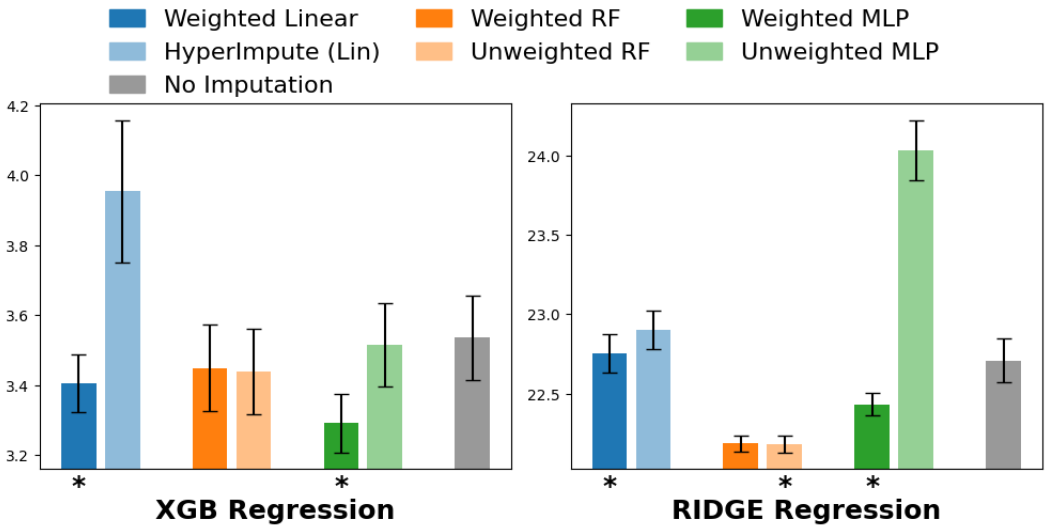


Figure 10: Lower values indicate better performance. We report mean RMSE performance of weighted and unweighted imputation methods, alongside an unimputed baseline, on downstream regression tasks evaluated on a held-out test set. Asterisks (*) indicate statistically significant improvements of weighted methods over their unweighted counterparts, based on a paired Wilcoxon signed-rank test ($p < 0.05$). Error bars denote ± 1.96 SEM, corresponding to approximate 95% confidence intervals. The plot is best viewed in colour.

C.8.2 Downstream Classification

We evaluate downstream classification on the **Letter** dataset using the letter label as the prediction target. This amounts to a multiclass problem. We follow the same procedure as in the downstream regression setting, ultimately training a tuned XGBoost classifier and logistic multiclass regression model, alongside corresponding unimputed baselines.

We evaluate performance using accuracy. The results are shown in Figure 11.

Downstream XGBoost classifier: We observe that the best-performing method (unweighted MLP imputation) improves accuracy by only 0.3% over the worst-performing configuration (unweighted linear imputation), indicating that effect sizes are small in this setting.

Downstream logistic regression classifier: We observe that all imputation methods improve downstream performance over the unimputed baseline. Furthermore, our weighted approach consistently outperforms its unweighted counterparts across all imputation families.

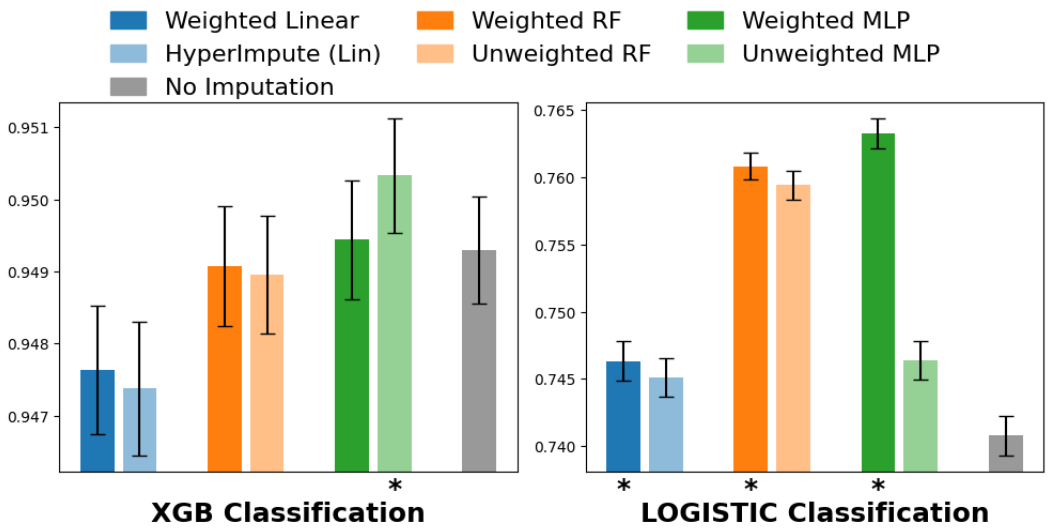


Figure 11: Higher values indicate better performance. We report mean accuracy of weighted and unweighted imputation methods, alongside an unimputed baseline, on downstream classification tasks evaluated on a held-out test set. Asterisks (*) indicate statistically significant improvements of weighted over unweighted methods based on a paired Wilcoxon signed-rank test ($p < 0.05$). Error bars denote ± 1.96 SEM, corresponding to approximate 95% confidence intervals. The plots are best viewed in colour.

C.8.3 Correlation Matrix Recovery

We also evaluate our imputation methods on correlation recovery as a measure how well we can recover the joint distribution in the data. We evaluate performance by measuring the frobenius error of the correlation matrix on the imputed data against the true data. Figure 12 shows that across all settings, the weighted model significantly improves over the unweighted variant.

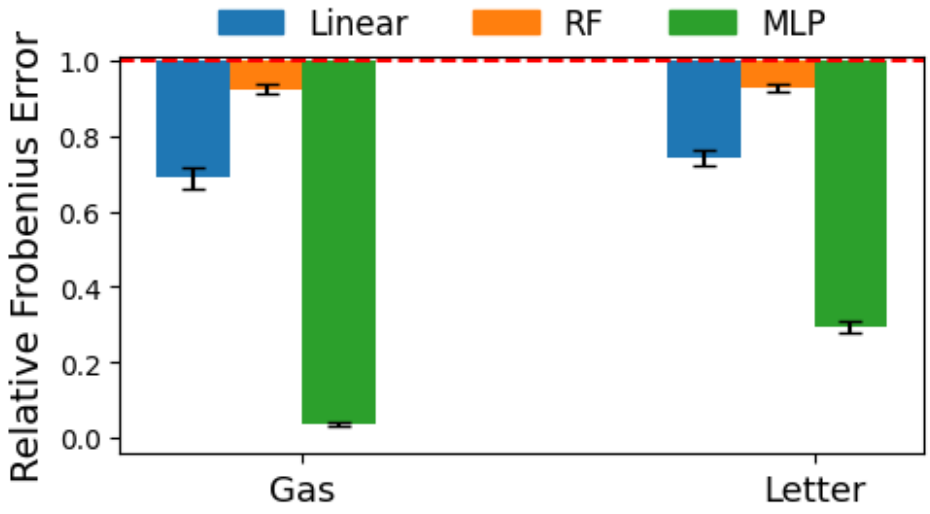


Figure 12: Values below 1 indicate improved performance. We report mean performance ratios of the weighted method relative to the unweighted baseline across the Gas and Letter datasets and conditional model types. Bars denote mean ratios, with error bars given by ± 1.96 SEM, corresponding to approximate 95% confidence intervals. The plots are best viewed in colour.

D Computational Resources

All experiments reported in Section 5 were conducted on the Isambard 3 high-performance computing facility hosted by the University of Bristol. Computations were performed on CPU-only nodes; no GPU acceleration was used in any experiments.

For the main experimental results, all models were run with a maximum memory allocation of 8GB per job. This was sufficient for all reported experiments. In practice, smaller datasets required substantially less memory, and could be executed with significantly lower resource allocations.

We did not observe memory bottlenecks across any of the reported settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the three contributions listed in Section 1: the theoretical result (Section 3.1), the proposed algorithm (Section 4), and the empirical evaluation (Section 5).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section (Section 7) discusses performance degradation for high-variance conditional imputation models, limited benefit under weak covariate shift, and non-extension to MNAR settings.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results (Lemma 1, Propositions 1, 2 and 3, Corollary 1) are stated with their full assumptions and complete proofs are provided in Appendices A and B. The MAR assumption and overlap condition required for Proposition 1 are explicitly stated in the theorem.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setup is fully described in Section 5.1 and Appendix C, including datasets (with URLs in Table 1), the MAR simulation procedure, evaluation metrics, model classes, and hyperparameter tuning strategies. Implementation details of our algorithm are provided in Section 4.1

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Anonymised code implementing the proposed imputation algorithm and simulation studies are provided as supplemental material. All datasets are publicly available from established repositories, with URLs provided in Table 1

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Full experimental details are given in Section 5.1 and Appendix C, including the MAR simulation parameters, number of seeds, hyperparameter tuning procedures, and dataset preprocessing steps.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All performance ratio figures include ± 1.96 SEM error bars corresponding to approximate 95% confidence intervals, as stated in each figure caption. Downstream task results include paired Wilcoxon signed-rank test significance indicators (Figures 10 and 11).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details of computational resources are given in Appendix D. Mean runtimes per dataset and model class are reported in Table 2.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work involves methodological research on publicly available benchmark datasets. The paper conforms in all respects with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper presents foundational methodological research on missing data imputation with no direct path to specific harmful applications. There are no foreseeable negative societal impacts specific to this work.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models or scraped datasets, and poses no high-risk misuse concerns. All datasets used are publicly available benchmark datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets are cited with their original sources and URLs in Table 1. All baseline methods are cited with their original papers throughout the paper.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pre-trained models are introduced.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects and therefore does not require IRB approval or equivalent.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs do not form any part of the core methodology of this research. The proposed algorithm relies on logistic regression for weight estimation and standard regression models for imputation.