Mask-Conditional Conformal Prediction: Valid Uncertainty For All Missing Data Mechanisms

Jiarong Fan
LaMME,
Université Paris-Saclay,
Evry

Juhyun Park
ENSIIE, LaMME,
Université Paris-Saclay,
Evry

Thi Phuong Thuy Vo
ENSIIE, LaMME,
Université Paris-Saclay,
Evry

Nicolas Brunel
ENSIIE, LaMME,
Université Paris-Saclay,
Evry;
Capgemini Invent France

Abstract

Conformal prediction (CP) offers a principled framework for uncertainty quantification, but it fails to guarantee mask-conditional coverage when faced with missing covariates. In addressing the heterogeneity induced by various missing patterns, Mask-Conditional Valid (MCV) Coverage has emerged as a more desirable property than Marginal Coverage. In this work, we adapt split CP to handle missing values by proposing a preimpute-mask-then-correct framework that can offer valid coverage. We show that our method provides guaranteed Marginal Coverage and Mask-Conditional Validity for general missing data mechanisms. A key component of our approach is a reweighted conformal prediction procedure that corrects the prediction sets after distributional imputation (multiple imputation) of the calibration dataset, making our method compatible with standard imputation pipelines. We derive two algorithms and prove that they achieve both marginal validity and MCV. We evaluate them on synthetic and real-world datasets. It reduces significantly the width of prediction intervals w.r.t standard MCV methods, while maintaining the target guarantees.

1 INTRODUCTION

Uncertainty estimation is essential for deploying machine learning models in real-world applications, such as finance, healthcare, and autonomous systems. A common approach to quantifying uncertainty is through prediction sets, such as intervals in regression or label collections in classification. The main goal of a prediction set is to contain the true label of a sample point with high probability, e.g., 90%.

To obtain valid prediction sets, Conformal Prediction (CP) has emerged as a widely adopted framework (Vovk et al., 2005). CP provides a model-agnostic and distribution-free approach to constructing prediction sets for any machine learning model (Angelopoulos et al., 2023). Given the random variable of test point $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ contains d features such that $\forall i \in \{1, ..., d\}, X_i \in \mathcal{X}_i$ with each $\mathcal{X}_i$ typically $\mathbb{R}$ or a discrete space, and a miscoverage rate α, CP aims to create a set $\hat{C}_\alpha(X)$ that contains Y with probability $1-\alpha$. Split CP (Papadopoulos et al., 2002), one of the most used CP techniques, achieves this by leveraging a calibration dataset to evaluate the model's uncertainty.

While CP seems to be well-adapted for dealing with applications under only the exchangeability assumption (weaker than i.i.d.) on the data-generating process, it fails to provide guarantees under perturbations, such as noisy labels, or missing data. (Einbinder et al., 2024; Zaffran et al., 2023, 2024; Feldman and Romano, 2024; Sesia et al., 2024; Zhou et al., 2025).

In particular, to ensure coverage guarantees with missing data, it is crucial to consider the missingness pattern or mask, denoted by the vector $M \in \{0, 1\}^d$, where $M_i = 1$ indicates the feature X_i is missing [1]. Missing data severely degrade model performance and have long been studied in ML and statistics (Emmanuel et al., 2021; Ren et al., 2023). To evaluate and correct this problem, more recent research focuses on imputing missing data and eventually choosing the best imputation algorithms to improve the reliability of predictions,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

[1] In this paper, for a vector V, we use the subscript V_i to denote the i-th element of V.

such as the impute-then-regress procedures (Le Morvan et al., 2021; Bertsimas et al., 2024; Lobo et al., 2025). It has been shown that missing values can also affect uncertainty quantification, including CP (Zaffran et al., 2023; Feldman and Romano, 2024).

Zaffran et al. (2023) introduced the concept of Mask-Conditional Validity (MCV, see Definition 2.1), emphasizing that missingness induces heterogeneity in the data, causing basic Split CP sets to exhibit under-coverage. If we denote by P the joint distribution of (X, Y, M), MCV corresponds to a group-wise control adapted to each pattern $m \in \{0,1\}^d$: $P(Y \in \hat{C}_\alpha | M = m) \geq 1 - \alpha$, which can be achieved by modifying how the calibration dataset is handled via data augmentation methods: CP-MDA-Exact and CP-MDA-Nested (Zaffran et al., 2023). However, the method is limited to MCAR and tends to be conservative in the sense of requiring a large data volume for CP-MDA-Exact, or generating a large prediction set for CP-MDA-Nested in extreme cases, e.g. Figure 12 in Zaffran et al. (2023) for data with 40% missing values.

As aleatoric uncertainty varies significantly with M (Zaffran et al., 2023), MCV is the focus of our work as it is a crucial property to ensure in practice in order to restore risk-controlled and useful prediction sets. To address the shortcomings of previous methods, we propose a novel preimpute-mask-then-correct framework built upon the standard practice of imputing first and applying the usual uncertainty quantification tools such as Split CP. A critical remark is that imputation (only) does not correct for the heterogeneity induced by different patterns of missingness; in order to control the miscoverage, we propose an alignment by missing pattern, illustrated in Figure 1, before applying Split CP. However, as we show in the experiment in Section 4.4, this approach fails to guarantee MCV due to distribution shifts introduced by imperfect imputation. We address this issue by generalizing Weighted CP (Tibshirani et al., 2019) to missing data and introducing two likelihood-based correction mechanisms, ensuring valid intervals even under MAR and MNAR conditions, which are more realistic assumptions (Kang, 2013).

Contributions.

We propose a preimpute-mask-then-correct framework with two novel correction methods, the mask-conditional Weighted CP and the Acceptance Rejection-Corrected (ARC) CP methods, to address the challenges of distribution shift and constructing valid prediction sets in datasets with missing covariates. Our ARC-CP method is a novel implementation that uses density-ratio weights to resample a subset, enabling the distribution shift reweighting problem to be solved explicitly.

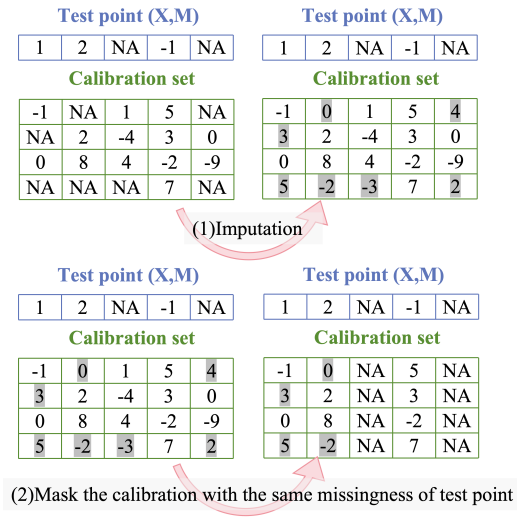

Figure 1: Illustration of the pre-imputation: the calibration data is imputed and then masked by the M of test point such that they share the same pattern. In this way, the split CP can give a prediction set conditional on M. But a distribution shift exists. Unlike CP-MDA, our method leverages all available calibration data without modifying the test point.

These approaches ensure both MV and MCV while significantly reducing the width of the prediction interval compared to previous methods, as demonstrated through rigorous theoretical analysis and extensive experiments on synthetic and real-world datasets. The proposed methods guarantee MCV without any assumptions on the missingness mechanism, which, to the best of our knowledge, has not been achieved before.

We also analyze the effect of inaccurate likelihood ratio estimators on MCV and establish lower bounds on the miscoverage, and empirically show that it remains limited and controlled. As our methodology requires only to impute once the calibration dataset, and do not require imputation of the test observation, our methodology can be integrated smoothly into a large number of data pipelines with maintained guarantees, and can be easily used in practice.

2 BACKGROUND

Missing data notation. Given the triplet random variable $Z := (X, M, Y) \in \mathcal{X} \times \{0,1\}^d \times \mathcal{Y} \sim P$, let $\mathcal{X}_{\text{NA}} = \prod_{j=1}^d (\mathcal{X}_j \cup \{\text{NA}\})$. We define the mask operator, $\text{mask} : \mathcal{X}_{\text{NA}} \times \{0,1\}^d \to \mathcal{X}$,

$$\text{mask}(x, m) := (1 - m) \odot x \; + \; m \odot \text{NA}, \qquad (1)$$

where $\odot$ is the element-wise product. Whenever there is no ambiguity, we use $\widetilde{X}$ to denote $\text{mask}(X, M)$ to

simplify the notations.

Denote by $\text{obs}(M)$ the set of indices of the observed covariates and by $\text{mis}(M)$ that of the missing covariates. For example, if $m = [0, 1, 0]$, then $\text{obs}(m) = \{1, 3\}$ and $X_{\text{obs}(m)} = [X_1, X_3]$.

Following Rubin (1976), we recall the standard classification of missingness mechanisms as MCAR (Missing Completely At Random) when $P(M = m \mid X, Y) = P(M = m)$, MAR (Missing At Random) when $P(M = m \mid X, Y) = P(M = m \mid X_{\text{obs}(M)})$ and Missing Not at Random (MNAR) when it is neither MAR nor MCAR.

To build a model on missing covariates, one may use imputation or not. Recent works have shown that for almost all deterministic imputation functions $\Phi \in \{\Phi : \mathcal{X}_{\text{NA}} \to \mathcal{X} \mid \Phi(\widetilde{X}_{\text{obs}(M)}) = X_{\text{obs}(M)}\}$, an impute-then-regress procedure with a consistent learner is Bayes consistent (Le Morvan et al., 2021). Of course, imputation is not always necessary. Tree models, such as Gradient Boosting Trees, are designed to handle covariates with missing values. For notational simplicity, we use $\hat{f}$ to denote any possible model defined in $\mathcal{X}_{\text{NA}}$.

Weighted Split CP. A popular formulation of CP is Split CP (Papadopoulos et al., 2002). Given a prefitted model $\hat{f}$ on t training data $\{(\widetilde{X}^i_{\text{tr}}, Y^i_{\text{tr}})\}_{i=1}^t$, and a holdout set of calibration data $\{\widetilde{Z}^i := (\widetilde{X}^i, Y^i)\}_{i=1}^n$, a nonconformity score $s : \mathcal{X}_{\text{NA}} \times \mathcal{Y} \to \mathbb{R}$ is calculated on all calibration points. In practice, s is induced by $\hat{f}$, which can be computed either directly on partially observed inputs or after applying an imputation strategy. Denote by $\mathcal{Q}_\beta(F)$ the β-level quantile of distribution F, i.e., for $T \sim F$, $\mathcal{Q}_\beta(T) = \inf\{z : P\{T \leq z\} \geq \beta\}$. Let $F_n := \{\Sigma_{i=1}^n \delta_{s(\widetilde{Z}^i)} + \delta_{+\infty}\}/(n+1)$, the empirical distribution of n scores, where the additional $\delta_{+\infty}$ is a standard convention in CP to guarantee finite-sample coverage. The Split CP prediction set for any test point $\widetilde{X}^{n+1} \in \mathcal{X}_{\text{NA}}$ is defined by

$$\hat{C}_\alpha(\widetilde{X}^{n+1}) := \{y \mid s(\widetilde{X}^{n+1}, y) \leq \mathcal{Q}_{1-\alpha}(F_n)\}. \quad (2)$$

As long as $\{\tilde{Z}^i\}_{i=1}^{n+1}$ are exchangeable, the prediction set contains Y^{n+1} with probability at least $1 - \alpha$ (Angelopoulos and Bates, 2021).

However, if the points are not exchangeable, as in Figure 1, the Split CP has no guarantee of coverage. Tibshirani et al. (2019) extended CP to handle non-exchangeable data under covariate shift. They considered reweighting the data using density ratio between the test and calibration set. Other scenarios of non-exchangability include repeated sampling (Sesia et al., 2023), time series (Gibbs and Candès, 2024; Zaffran et al., 2022) or label shift (Si et al., 2023; Podkopaev and Ramdas, 2021). Recently, Prinster et al. (2024) generalize conformal prediction to arbitrary data distributions beyond exchangeability under sequential, feedback-induced covariate shifts. While they extend significantly the scope of Weighted CP, it relies on the knowledge of data generating distribution (possibly factored with graphs), which is rarely available with missing data and imputation mechanisms.

Marginal Validity (MV) and Mask-Conditional Validity (MCV). In the following, we assume $t + n + 1$ i.i.d. realizations of the pair of random variables $(\widetilde{X}, Y)$, partitioned into a training set `Train` of size t, a calibration set of size n, and one test point. Let $\texttt{Cal} = \{\widetilde{Z}^i\}_{i=1}^n$ denote the calibration set. Our task is to construct a prediction set for Y^{n+1} given `Train`, `Cal`, and $\widetilde{X}^{n+1}$. Facing missing values, we are particularly interested in properties in Definition 2.1.

Definition 2.1. A method producing intervals $\hat{C}_\alpha(\widetilde{X}^{n+1})$ is MV at level α if:

$$P(Y^{n+1} \in \hat{C}_\alpha(\widetilde{X}^{n+1})) \geq 1 - \alpha,$$

and is MCV at level α if, for any $m \in \{0, 1\}^d$ such that $P(M = m) > 0$,

$$P(Y^{n+1} \in \hat{C}_\alpha(\widetilde{X}^{n+1}) \mid M^{n+1} = m) \geq 1 - \alpha.$$

The problem of uncertainty quantification under missing values has recently attracted increasing attention. Broadly two different approaches have emerged. First, Liang et al. (2024); Gui et al. (2023); Shao and Zhang (2023) addressed the problem through matrix completion, introducing novel CP strategies tailored to matrix data, with strong theoretical guarantees. However, their assumptions– such as the random observation model, where entries are missing independently according to a Bernoulli distribution (i.e., MCAR)—may not hold in typical supervised learning settings. Moreover, Shao and Zhang (2023) emphasized that matrix prediction and matrix imputation are fundamentally different problems: the former assumes structure across both rows and columns (e.g., low-rank or exchangeability), while the latter deals with fixed covariates and responses, where features are typically non-exchangeable. Second, Zaffran et al. (2023, 2024), which motivated our work, addressed it by introducing two data augmentation techniques (CP-MDA-Exact and CP-MDA-Nested) within the Conformal Prediction with Missing Data Augmentation (CP-MDA) framework, as well as a more flexible generalization, CP-MDA-Nested*. Their main algorithm is reminded in Algorithm 1.

Under MCAR, CP-MDA-Nested* has been proven to be MCV at the level 2α (Zaffran et al., 2024). The method is CP-MDA-Exact when $\widetilde{Id}_{\texttt{Cal}} = \{k \in \{1, ..., n\} : \sum_{j=1}^d M_j^k = \sum_{j=1}^d M_j^k M_j^{n+1}\}$ and CP-MDA-Nested when $\widetilde{Id}_{\texttt{Cal}} = \{1, ..., n\}$. However, for CP-MDA-Exact,

Algorithm 1 CP-MDA-Nested* Zaffran et al. (2024)

Input: $\texttt{Cal} = \{(\widetilde{X}^k, Y^k)\}_{k=1}^n$, nonconformity score function $s(\cdot,\cdot)$, significance level α, test point $\widetilde{X}^{n+1}$, any subsampled set of calibration indice $\widetilde{Id}_{\text{Cal}} \subseteq \{1, ..., n\}$. **Output:** $\widehat{C}_{n,\alpha}^{\text{MDA-Nested}^\star}(\widetilde{X}^{n+1})$.

1: $\forall k \in \widetilde{Id}_{\texttt{Cal}}$, set $\bar{M}^k = [\max(M_i^k, M_i^{n+1}), i = 1, ..., d]$

2: $\widehat{C}_{n,\alpha}^{\text{Nested}^\star}(\widetilde{X}^{n+1}(M^{n+1})) := \Big\{y : (1-\alpha)(1 + \#\widetilde{Id}_{\texttt{Cal}})$
$> \sum_{k \in \widetilde{Id}_{\texttt{Cal}}} \mathbf{1}\Big[s(\widetilde{X}^k(\bar{M}^k), Y^k) < s(\widetilde{X}^{n+1}(\bar{M}^k), y)\Big]\Big\}$

on the one hand, since only the calibration points having fewer missing values than the test point are selected, there may not be enough points for a small calibration set, which is infeasible in small-sample regimes. On the other hand, for CP-MDA-Nested, the test point is modified by augmenting the mask with additional missing values, which inevitably increases its conservativeness and generates larger prediction intervals. The CP-MDA-Nested* with a more flexible target missingness pattern is a compromise between the two methods, but how to choose the best $\widetilde{ID}_{\texttt{Cal}}$ remains a question.

Relationship to Weighted CP and Shift Correction. Beyond matrix completion and CP-MDA methods, our work is also connected to a broader research that uses importance-weighted learning to correct for distributional shifts, as in Bickel et al. (2009); Zhou et al. (2023). While previous works have largely focused on improving importance weight estimation under covariate shift, our focus is distinct: we study how such estimators can be integrated within conformal prediction to maintain mask-conditional coverage under missing data mechanisms. Thus, our main novelty lies in the preimpute-mask-then-correct framework adapted to missingness and we emphasize the role of imputation and its impact on calibration validity. In addition, we provide an evaluation of the consequences of inaccurate estimators, quantifying how poor estimation affects the miscoverage of the prediction sets.

3 METHODOLOGY

3.1 Preimputation of calibration dataset

As shown in Figure 1, the preimputation ensures that Split CP can be applied. Since the response Y is observed in the calibration set, it is natural to use it for imputation, in order to improve the reconstruction of missing covariates. Hence, we consider potentially more general imputation that use both $X_{\text{obs}(M)}$ and Y, instead of the standard ones that rely only on $X_{\text{obs}(M)}$. The latter case is covered by our results, but we find in our experiments that the general approach provides better imputations. A critical point in our work is the use of distributional (or multiple) imputation, which has been shown to provide good results and desirable properties (Näf et al., 2024), such as MICE.

Hence, we consider only distributional imputation strategies for the calibration data set. Formally, starting from the partially observed data $(\widetilde{X}, Y)$, we derive a complete version $(\hat{X}, Y)$, where law of $\hat{X}|(\widetilde{X}, Y)$ follows a certain distribution supported in $\mathcal{X}$. Let Q denote the joint distribution of $\hat{Z} := (\hat{X}, Y)$. Applying this procedure to each calibration point, we construct an imputed calibration dataset $\{\hat{Z}^i := (\hat{X}^i, Y^i)\}_{i=1}^n$ and an imputed test point $\hat{Z}^{n+1} := (\hat{X}^{n+1}, Y^{n+1})$. It should be noted that, $\hat{X}^i$ being the imputed value for $\widetilde{X}^i$, we have $\hat{X}^i_{\text{obs}(M^i)} = X^i_{\text{obs}(M^i)}$. Such imputations are commonly used in practice, such as Bayesian Linear Ridge in Scikit-learn (Pedregosa et al., 2011). The formal definition is in section A. For example, if we use a Bayesian ridge regression for imputation, then $\hat{X}_{\text{mis}(M)}|(\widetilde{X}, Y) \sim \mathcal{N}\big(\hat{\mu}(\widetilde{X}, Y), \hat{\Sigma}(\widetilde{X}, Y)\big)$, composed with estimated conditional mean $\hat{\mu}(\widetilde{X}, Y)$ and covariance matrix $\hat{\Sigma}(\widetilde{X}, Y)$.

Throughout the article, the following assumptions are made. Here, for notational simplicity, we slightly abuse notations: for $\mathcal{I}$ a set of indices (e.g. $\mathcal{I} = \text{obs}(m)$), $Z^i_{\mathcal{I}} := (X^i_{\mathcal{I}}, Y^i)$, $\hat{Z}^i_{\mathcal{I}} := (\hat{X}^i_{\mathcal{I}}, Y^i)$ and $\widetilde{Z}^i_{\mathcal{I}} := (\widetilde{X}^i_{\mathcal{I}}, Y^i)$.

Assumption 1.

(A1) The calibration and test data $\{\widetilde{Z}^i\}_{i=1}^{n+1}$ are i.i.d.

(A2) The imputer is trained on the training set and applied in the same way to all points; hence, the imputed dataset $\{\hat{Z}^i\}_{i=1}^{n+1}$ is also i.i.d.

(A3) The predictive model is fitted only on the training set.

These assumptions are standard in the CP literature and are generally satisfied in practice.

Remark 1. Based on Assumption 1, all points imputed are i.i.d. Thus, any prediction set using Split CP on the imputed calibration and test datasets is MV, as the imputation preserves exchangeability. This result generalizes Lemma 3.2 of Zaffran et al. (2023), which was limited to the deterministic imputation case. We additionally show that "impute-then-split CP" ensures marginal validity for distributional imputation.

For any mask $m \in \{0,1\}^d$, we denote $\mathcal{X}_{\text{obs}(m)} := \prod_{i \in \text{obs}(m)} \mathcal{X}_i$ and for any measurable subset $A \subseteq \mathcal{X}_{\text{obs}(m)} \times \mathcal{Y}$, we define $Q_m(A) := Q(\hat{Z}_{\text{obs}(m)} \in A)$, and the original conditional measure as $P_m(A) := P(Z_{\text{obs}(m)} \in A \mid M = m)$. After pre-imputation, conditionally on M^{n+1}, calibration points follow

$\hat{Z}^i_{\text{obs}(M^{n+1})} \sim Q_{M^{n+1}},\ i = 1, ..., n$, while the test point follows the true distribution $Z^{n+1}_{\text{obs}(M^{n+1})} \sim P_{M^{n+1}}$. Generally, $Q_{M^{n+1}} \neq P_{M^{n+1}}$, resulting in a mask-conditional distribution shift. This discrepancy limits the direct application of Split CP.

We introduce now Proposition 1 that gives a first insight into the impact of imputation on the mask-conditional coverage.

Proposition 1. Under Assumption 1, if we apply Split CP to the test point $\widetilde{X}^{n+1}$, with masked calibration dataset $\{\text{mask}(\hat{X}^i, M^{n+1}), Y^i\}_{i=1}^n$, then for almost every $m \in \{0,1\}^d$ such that $P(M = m) > 0$, the corresponding prediction set $\hat{C}_\alpha$ defined in eq. (2) satisfies:

$$P(Y^{n+1} \in \hat{C}_\alpha(\widetilde{X}^{n+1}) | M^{n+1} = m) \geq 1 - \alpha - \left(d_{TV}(P, Q) + \mathbb{E}_P \left[\left| \frac{dP(X, Y | M = m)}{dP(X, Y)} - 1 \right| \right] \right)$$

In addition, if P is MCAR:

$$P(Y^{n+1} \in \hat{C}_\alpha(X^{n+1}) | M^{n+1} = m) \geq 1 - \alpha - d_{TV}(P, Q)$$

See Section C.1 for the proof. The bound decomposes the mask-conditional miscoverage into two components: the error due to imperfect imputation and the intrinsic distribution shift caused by mask-dependent sampling. This first result indicates that under MAR or MNAR conditions, when $P(X, Y \mid M = M^{n+1}) \neq P(X, Y)$, valid coverage may not be guaranteed even when the imputation is perfect, i.e. $P = Q$. The second result highlights the importance of the quality of pre-imputation in ensuring valid coverage under MCAR. Even if our lower bounds may be loose, our results contradict apparently the results of Le Morvan et al. (2021), which focus on consistency and asymptotic Bayes-optimality, stating that "imputation doesn't matter for regression task". While these previous results are first-order, our method highlights the significant impact of pre-imputation on second-order properties, such as uncertainty and MCV. A detailed discussion on pre-imputation quality is provided in Appendix A.

3.2 Weighted conformal prediction under mask-conditional distribution shift

To solve the distribution shift caused by imputation, we introduce Assumption 2.

Assumption 2. The joint distribution of $(X, Y) \sim P$ is absolutely continuous with respect to $(\hat{X}, Y) \sim Q$, and we denote $\omega(x, y) := \frac{dP}{dQ}(x, y)$.

This assumption is satisfied in many common scenarios. For instance, if the missing covariates are continuously distributed and we use a distributional imputation method (e.g., MICE), then the imputed distribution is absolutely continuous with respect to the original one. If the missing covariates are discrete or contain mass points, the condition still holds as long as the imputation assigns nonzero probability mass to those same points. In practice, this property can be easily ensured, for example, by using distributional imputers or simply by adding a small Gaussian perturbation to the imputed values, which guarantees that the imputed distribution covers the full support of the original one.

Under Assumption 2, for any mask $m \in \{0,1\}^d$ such that $P(M = m) > 0$, the conditional distribution P_m is absolutely continuous w.r.t. Q_m with likelihood ratio $\omega_m(x_{\text{obs}(m)}, y) := \frac{dP_m}{dQ_m}(x_{\text{obs}(m)}, y)$, with convention $\omega_m = 0$ if $Q_m = 0$. Denote for all $i \in \{1, ..., n\}$, $\omega^i_m := \omega_m(\hat{X}^i_{\text{obs}(m)}, Y^i)$, and define the weights functions for all $i \in \{1, 2, ..., n\}$, $W^i_m : \mathcal{X}_{\text{obs}(m)} \times \mathcal{Y} \to \mathbb{R}$:

$$W^i_m(x_{\text{obs}(m)}, y) := \frac{\omega^i_m}{\sum_{j=1}^n \omega^j_m + \omega_m(x_{\text{obs}(m)}, y)},$$

$$\text{and } W^{n+1}_m(x_{\text{obs}(m)}, y) := \frac{\omega_m(x_{\text{obs}(m)}, y)}{\sum_{j=1}^n \omega^j_m + \omega_m(x_{\text{obs}(m)}, y)}.$$

For any nonconformity score $s : \mathcal{X}_{\text{NA}} \times \mathcal{Y} \to \mathbb{R}$, define on the same domain the score for imputed-then-masked calibration points: $S^i_m := s(\text{mask}(\hat{X}^i, m), Y^i)$.

Theorem 1. Under Assumptions 1 and 2, define the weighted split CP interval conditional on $M^{n+1} = m$:

$$\widehat{C}^W_\alpha(\widetilde{X}^{n+1}) = \{y : s(\widetilde{X}^{n+1}, y) \leq \mathcal{Q}_{1-\alpha}(\sum_{i=1}^n W^i_m(\widetilde{X}^{n+1}_{\text{obs}(m)}, y)\delta_{S^i_m} + W^{n+1}_m(\widetilde{X}^{n+1}_{\text{obs}(m)}, y)\delta_{+\infty})\} \tag{3}$$

Then for any missingness mechanism (MCAR, MAR, or MNAR): $P(Y^{n+1} \in \widehat{C}^W_\alpha(\widetilde{X}^{n+1}) | M^{n+1} = m) \geq 1 - \alpha$.

See Section C.2 for proof. The key insight of Theorem 1 is that when imputation introduces a mask-conditional distribution shift, the likelihood ratio ω_m adjusts the quantile and ensures valid coverage. Unlike Tibshirani et al. (2019) which addresses only covariate shifts with explicit prediction intervals, our method requires y-dependent weights for each calibration point in Equation (3). The formal Algorithm 3 is presented in Section C and for implementation and runtime considerations, we discuss our practical strategies in Appendix B. The result and algorithm of Theorem 1 still apply if $\omega_m(x_{\text{obs}(m)}, y)$ are known modulo a multiplicative constant. Nevertheless, it is still computationally expensive and we propose below a more efficient algorithm.

3.3 ARC conformal prediction under mask-conditional distribution shift

Generally, one would expect the pre-imputation to have at least some ability to reconstruct the original distribution, such that the gap between imputed data (proposal distribution Q_m) and complete distribution (target distribution P_m) can be reduced. Therefore, this is a well-suited scenario for drawing observations from Q_m by using acceptance-rejection (AR) to target P_m. Suppose ω_m^i is bounded by a constant $K > 0$. The idea of AR is to draw a subset from $\{\hat{Z}^i\}_{i=1}^n$ uniformly at random. Specifically, for each point, we sample $U^i \sim U(0,1)$ and retain the point if and only if: $U^i < \omega_m^i(\hat{X}^i_{\text{obs}(m)}, Y^i)/K$. The formal Algorithm 4 is presented in Appendix C. Algorithm 4 outputs accepted $\{(\hat{X}^j, Y^j)\}_{j\in\mathcal{I}_{\texttt{AR}}}$, that we mask with m to create a new calibration dataset $\widetilde{\texttt{Cal}}_m = \{(\text{mask}(\hat{X}^j, m), Y^j)\}_{j\in\mathcal{I}_{\texttt{AR}}}$. Theorem 2 shows that $\widetilde{\texttt{Cal}}_{M^{n+1}}$ has distribution $P_{M^{n+1}}$ such that we can apply standard split CP calibration to get an MCV prediction set on test point $\widetilde{X}^{n+1}$.

Theorem 2. For any $m \in \{0,1\}^d$, suppose ω_m^i is bounded by a constant $K > 0$. We generate $\widetilde{\texttt{Cal}}_m$ with index $\mathcal{I}_{\texttt{AR}}$ selected by Algorithm 4. Then: (a) The points in $\widetilde{\texttt{Cal}}_m$ are i.i.d. samples from P_m; (b) The split prediction set $\widehat{C}_\alpha^{\texttt{AR}}$ (see Algorithm 4) satisfies

$$P(Y^{n+1} \in \widehat{C}_\alpha^{\texttt{AR}}(\widetilde{X}^{n+1}) \mid M^{n+1} = m) \geq 1 - \alpha.$$

See Section C.3 for proof and the detailed ARC-CP algorithm. Theoretically, if Assumption 2 is violated, the likelihood ratio may not be well-defined or bounded, compromising our methods' validity. This situation occurs when certain observable points cannot be generated by imputation (see Appendix A for discussion), highlighting the importance of dependable imputations and the benefit of having close distributions (with the same support) to ensure a high acceptance rate.

3.4 Estimation of likelihood ratio ω_m

Similar to Tibshirani et al. (2019), we propose the following Algorithm 2 for training an estimator for the likelihood ratio ω_m. We provide here a sketch of the proof for its validity. The idea is to label the randomly generated data $\tilde{x}$ "1" if it does not contain any imputed value (i.e. $(X_{\text{obs}(m)}, Y) \sim P_m$) or "0" if it does (i.e. $(X_{\text{obs}(m)}, Y) \sim Q_m$). We train a binary classifier, such that we estimate the probability $P(C = 1 \mid \widetilde{X} = \tilde{x}, Y = y)$ with $\hat{p}(\tilde{x}, y)$. By Bayes' theorem, we get $\frac{P(C=1|\widetilde{X}=\tilde{x},Y=y)}{P(C=0|\widetilde{X}=\tilde{x},Y=y)} = \beta(dP_m/dQ_m)(x_{\text{obs}(m)}, y)$, with the constant $\beta = P(C=1)/P(C=0)$. The estimation of ω_m uses solely the calibration dataset, and does not depend on the test value (x, y).

Algorithm 2 Estimation of Likelihood Ratio

Input: Points $\{(x^i, m^i, y^i)\}_{i=1}^t$ sampled from imputed set $(\hat{X}, M, Y)$; number of random masks $q \in \mathbb{N}^\star$. **Output:** Estimator of $\omega_m(x_{\text{obs}(m)}, y)$.

1: For each (x^i, m^i, y^i), generate q random masks $\{\check{m}_j^i\}_{j=1}^q$.
2: Compute $c_j^i = \mathbb{1}\{m^i = \check{m}_j^i\}$ for all i, j, and construct dataset $\{(x^i, \check{m}_j^i, y^i, c_j^i)\}$.
3: Train a probabilistic classifier $\hat{p}$ (e.g., histogram-based GBDT) to predict c_j^i from $(\text{mask}(\tilde{x}^i, \check{m}_j^i), y^i)$.
4: $\hat{\omega}_m := \hat{p}(\tilde{x}(m), y)/(1 - \hat{p}(\tilde{x}(m), y))$ estimates ω_m.

3.5 Impact of ratio estimation on coverage

The accuracy of the estimator may impact the coverage of Weighted CP. This can be the case when it is difficult to estimate, for instance in high dimensions. Therefore, we provide below a theoretical analysis of the potential miscoverage, and complement it with an empirical study in Section D for completeness.

Let $\hat{\omega}_m(\cdot)$ be our estimator of the density ratio and we define $\tilde{\omega}_m(\cdot) := \hat{\omega}_m(\cdot)/\mathbb{E}_Q[\hat{\omega}_m(\hat{X}_{\text{obs}(m)}, Y)]$. If we denote $d\tilde{P}_m := \tilde{\omega}_m(\hat{X}_{\text{obs}(m)}, Y) \cdot dQ_m$, then, under Assumption 2, $\tilde{P}_m$ is a well defined probability measure.

Proposition 2. Under Assumption 2, for any mask m, if we use the estimated likelihood ratio $\hat{\omega}$ in (3), denoted by $\widehat{C}_\alpha^{\widehat{W}}$, then we have

$$\begin{aligned} &P(Y^{n+1} \in \widehat{C}_\alpha^{\widehat{W}}(\widetilde{X}^{n+1}) \mid M^{n+1} = m) \\ &\geq (1-\alpha) - d_{TV}(\tilde{P}_m, P_m). \end{aligned} \tag{4}$$

To gauge the loss, we propose an implementable lower bound for the coverage below.

Proposition 3. Suppose we have an extra dataset $\{(\check{X}^k, \check{Y}^k, \check{M}^k)\}_{k=1}^J$, with i.i.d. points from P. Using Algorithm 5, generate two datasets $\{Z_P^i\}_{i=1}^{l_1} \sim P_m$ and $\{Z_{\tilde{P}}^i\}_{i=1}^{l_2} \sim \tilde{P}_m$, and set $l = \min(l_1, l_2)$. Form the balanced classification sample by adding a label C, $Z_c := \{(Z_P^i, 1)\}_{i=1}^l \cup \{(Z_{\tilde{P}}^i, 0)\}_{i=1}^l$. Let $\hat{g}_m \in \arg\min_{g\in\mathcal{G}_m} \hat{R}_\gamma(g_m)$ be an empirical risk minimization over $\mathcal{G}_m$ with respect to the 0–1 loss γ on Z_c defined on Appendix C.5, and denote by $\hat{R}_\gamma(\cdot)$ the empirical risk on Z_c. Under regularity Assumptions 3 to 7, for any $\epsilon > 0$, we have with probability at least $(1 - \exp(-\frac{\epsilon^2}{2l\beta_l^2}) - \exp(-\frac{\epsilon^2 l}{2}))$ that:

$$\begin{aligned} &P(Y^{n+1} \in \hat{C}_\alpha^{\widetilde{W}}(\tilde{X}^{n+1}) \mid M^{n+1} = m) \\ &> 1 - \alpha - (1 - 2\,\hat{R}_\gamma(\hat{g}_m)) - \epsilon - K\sqrt{\frac{V_{\mathcal{G}_m}}{l}}, \end{aligned} \tag{5}$$

with $\beta_l, V_{\mathcal{G}}, K$ defined in Appendix C.5.

See Appendix C.6 for the proof. With an imperfect estimator of the likelihood ratio, Proposition 3 yields an estimable lower bound for coverage, where the key interpretable term is $\frac{1}{2} - \hat{R}_\gamma(\hat{g}_m)$. Intuitively, if the estimator is nearly accurate, then the discriminator $\hat{g}_m$ cannot distinguish between P_m and $\tilde{P}_m$, resulting in $\hat{R}_\gamma(\hat{g}_m) \approx \frac{1}{2}$. Conversely, if the estimator is poor, $\hat{g}_m$ can easily separate and lead to $\hat{R}_\gamma(\hat{g}_m) \approx 0$, which weakens the coverage guarantee. Note that this requires estimating a g_m for each mask m, which may seem costly. However, Section 4 confirms that even with approximate estimators, our method achieves near-nominal coverage across diverse missingness settings—validating the practical robustness of our method.

4 EXPERIMENTS

The primary goals of numerical studies are: (1) to validate the MCV property of the proposed methods under missingness scenarios, (2) to compare the interval width and coverage performance with CP-MDA methods, (3) to demonstrate the necessity and utility of distribution shift corrections, (4) to study the impact of imperfect estimators of likelihood ratio and evaluate their miscoverage.

For simplicity, we consider $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. We use the nonconformity score s defined by Conformalized Quantile Regression method (Romano et al., 2019). Specifically, the model $\hat{f}$ is composed of an iterative ridge imputation and a pair of Quantile Regression Forest (Meinshausen and Ridgeway, 2006) models $(\hat{f}_{low}, \hat{f}_{high})$, which are both learned on an independent training set. s is thus defined by $s(\tilde{x}, y) = \max(\hat{f}_{low}(\tilde{x}) - y, y - \hat{f}_{high}(\tilde{x}))$. Iterative Bayesian Ridge (MICE in Scikit-learn with sample_posterior being True) is used to pre-impute the calibration. The target $1 - \alpha$ is fixed to 90%, and the likelihood ratio ω_m is estimated using a Histogram-based Gradient Boosting Classification Tree.

We present results on synthetic data and real UCI datasets under MCAR, MAR, and MNAR [2]. Missing values are generated accordingly. For each task, we generate different datasets and repeat experiments to compute a confidence interval for each mask's average coverage and interval width. All results show that our methods significantly reduce the width of prediction intervals, while also guaranteeing the MCV.

[2]The code is available on `https://github.com/jrfan/Mask-Conditional-Conformal-Prediction`

4.1 Experiments on synthetic dataset

We follow Zaffran et al. (2023, 2024) to generate synthetic datasets, as the oracle length of prediction intervals can be derived and compared. The data follows $Y = \beta^T X + \varepsilon$, $\beta \in \mathbb{R}^d$, with $d = 10$ and $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = (1, \cdots, 1)^T$ and $\Sigma = \rho(1, \cdots, 1)^T(1, \cdots, 1) + (1 - \rho)I_d$, $\rho = 0.8$. Gaussian noise $\varepsilon \sim \mathcal{N}(0, 1) \perp\!\!\!\perp (X, M)$ and $\beta = (1, 2, -1, 3, -0.5, -1, 0.3, 1.7, 0.4, -0.3)^T$. To better visualize the differences between methods, we use a relatively high missingness probability of 50% per feature. We also report results for 20% missingness in Section E to provide another perspective. The Oracle interval width, defined as the minimum width of MCV interval knowing X_M, is calculated by Proposition 4.1 in Zaffran et al. (2023). In our experiments, 500 training data and 100 calibration data are used. 100 test data are generated for 1023 types of missingness patterns ($2^{10} - 1$, the completely missing pattern is excluded). Each group is used separately to evaluate the missingness-wise coverage and interval width. We repeat each experiment 500 times.

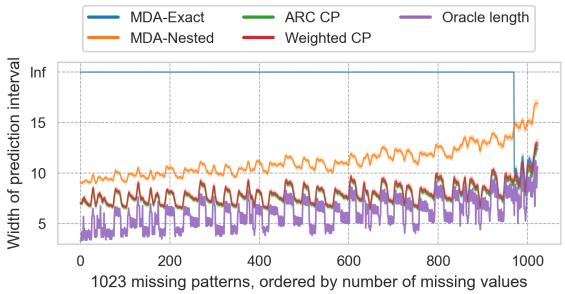

Figure 2: Synthetic dataset on MCAR: prediction interval width across 1023 masks (mean and 95% CI).

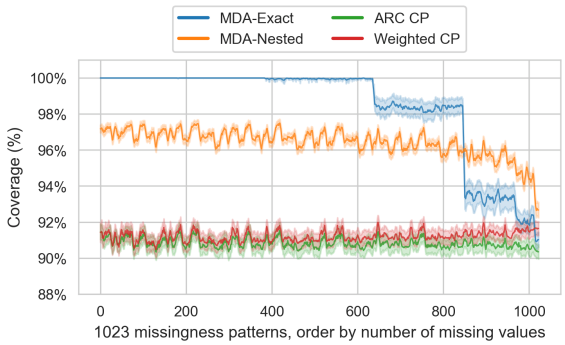

Figure 3: Synthetic dataset on MCAR: prediction coverage across 1023 masks (mean and 95% CI).

Figure 2 and Figure 3 compare our methods with CP-MDA methods missingness-wise. For clarity, we applied a window smoothing with size of 5 to the curves. The

Table 1: Worst-case coverage (95% CI) and average prediction interval width over 31 masks.

Mechanism	MDA-Exact	MDA-Nested	Weighted CP	ARC CP
MAR	87.8–89.7% / 5.84	87.9–89.8% / 5.67	89.0–91.2% / 5.79	88.3–90.2% / 5.64
MNAR	87.8–89.8% / 5.74	87.8–89.8% / 5.64	89.1–91.0% / 5.79	88.5–90.3% / 5.60

Table 2: 100-dimensional synthetic setting: average mask-conditional coverage and interval length, grouped by oracle interval length $L_{\text{oracle}}(m)$. Results of coverage and length are averaged over 1000 masks with 50 replications.

Bin of $L_{\text{oracle}}(m)$	ARC-CP cov	ARC-CP len	WCP cov	WCP len	Nested cov	Nested len
$[17.74, 20.39]$	0.911	27.92	0.884	27.14	0.926	29.58
$[20.39, 21.16]$	0.910	27.99	0.885	27.20	0.927	29.76
$[21.16, 21.57]$	0.910	28.04	0.885	27.16	0.928	29.90
$[21.57, 22.02]$	0.911	28.08	0.881	27.08	0.927	29.91
$[22.02, 22.40]$	0.911	28.20	0.882	27.17	0.928	30.04
$[22.40, 22.77]$	0.911	28.14	0.884	27.21	0.927	30.17
$[22.77, 23.14]$	0.910	28.17	0.884	27.33	0.927	30.20
$[23.14, 23.64]$	0.911	28.28	0.883	27.36	0.926	30.35
$[23.64, 24.27]$	0.910	28.25	0.883	27.27	0.928	30.53
$[24.27, 26.73]$	0.910	28.37	0.882	27.42	0.928	30.87

unsmoothed curves are provided in Section A, and both lead to the same conclusion. These results confirm our theoretical MCV guarantee as established in Theorems 1 and 2, for all missingness patterns. Our methods significantly reduce prediction interval width. In contrast, MDA-Exact fails to generate informative prediction sets for certain missingness patterns, while MDA-Nested is more conservative. Overall, our methods achieve an average **30% reduction**[3] in prediction width compared to MDA-Nested.

We also generate MAR and MNAR (self-masked) mechanisms on the same synthetic dataset by methods in Schouten et al. (2018); Zaffran et al. (2024). The first 5 covariates $\{X_1, X_2, X_3, X_4, X_5\}$ may be missing, with total missingness probability 20%. Under MAR, M is a function $h_{\text{MAR}}(X_6, ..., X_{10})$ and under MNAR, it is $h_{\text{MNAR}}(X_1, ..., X_5)$. We report the 95% confidence interval for the worst-case coverage for 31 missingness patterns in Table 1. One should be careful when generating points for each pattern as $(X, Y)|M = m$ does not have an explicit form under MAR or MNAR. We repeat 100 experiments. Our methods are MCV under both MAR and MNAR, with ARC CP providing the shortest prediction intervals.

Naive imputation baselines. We also evaluate two naive strategies on the synthetic dataset. First, one commonly adopted approach is to impute all covariates and apply standard split CP on the fully imputed data. This achieves marginal coverage of 90.6% but violates mask-conditional validity: among the $2^{10} = 1024$ masks, more than 100 have mask-conditional coverage below 90%, and the worst-case coverage drops to 61.8%. Second, we train a quantile regression model (HistGradientBoostingRegressor with pinball loss) and compute prediction intervals directly. While this yields relatively short average intervals (length ≈ 3.17), the worst-case mask-conditional coverage is only 71.2%. These baselines show that naive imputation-based approaches can achieve marginal average, but fail to adapt to heterogeneity across missingness patterns, and under-cover for some masks.

4.2 High-dimensional data ($d = 100$).

We further consider a 100-dimensional version of the same dataset in Section 4.1. The exact β is detailed in section E.7. Since there are 2^{100} possible masks, we only sample 1000 masks uniformly at random and, for each mask, generate 100 test points to estimate its mask-conditional coverage.

For each mask m, we compute the oracle 90% interval length $L_{\text{oracle}}(m)$, i.e., the shortest possible interval achieving 90% coverage under the true data-generating model. We use $L_{\text{oracle}}(m)$ as a difficulty measure and group masks into 10 bins according to this "oracle difficulty". In each bin, we report the average mask-conditional coverage and average interval length; see Table 2.

Across all oracle-difficulty bins, ARC-CP attains cov-

[3]The average of interval width over all 1023 missingness patterns is computed to compare the interval width.

erage around 91% with average interval lengths between 27.9 and 28.4, and never produces infinite intervals, whereas MDA-Exact always fails due to limited sample. Overall, ARC-CP offers a favorable trade-off between coverage, efficiency, and runtime in this 100-dimensional regime.

4.3 Experiments on real dataset

Concrete Data from the UCI repository (Yeh (1998), **licensed under CC BY 4.0**): The goal is to predict compressive strength using 8 features. 630 training data and 100 calibration data are selected. For MCAR, 100 test data are generated per mask, with 50% feature-wise missingness, resulting in 255 patterns ($2^8 - 1$). All four methods on all missingness patterns achieve 90% coverage[4]. Compared with MDA-Nested, ARC CP reduced the average width **by 12%** and Weighted CP reduced average width **by 10%**.

For MAR, the first 4 features may be missing, with rates depending on the other 4 (totaling 20%). Our methods are MCV with worst-case coverages: MDA-Exact $89.1\% \pm 1.64\%$, MDA-Nested $90.8\% \pm 1.43\%$, ARC CP $90.1\% \pm 1.63\%$, and Weighted CP $90.6\% \pm 1.42\%$. ARC CP yields the narrowest average intervals—reducing width **by 7.5%** relative to MDA-Nested (see Table 6). **Three additional real datasets** with higher dimensions are provided in the Appendix E.2. Both show the valid coverage of our methods.

4.4 Ablation study of weight correction

To demonstrate the effectiveness of correction methods, we did experiments on the same synthetic data without correction, i.e., we pre-impute the calibration and apply the mask of the test point, ignoring the distribution shift. The detailed experiments are in Section E.3, showing that the three methods have very similar interval widths Figures 8 and 9, but, without correction, the worst mask conditional coverage is around 89%, showing that without a suitable weight/acceptance for correction, the distribution shift results in an under-coverage. This confirms the necessity of correction when applying CP under missing data.

4.5 Impact of imperfect estimators

Accurately estimating the likelihood ratio is never trivial. We conduct a synthetic case study (see Appendix D) to explicitly evaluate the effect of the likelihood ratio estimator's quality on the coverage. Using a Gaussian generative model where the true density ratio is known, we show that (i) inaccurate estimators can lead to under- or over-coverage, and (ii) the empirical coverage correlates strongly with the estimator quality (e.g., Pearson correlation with the true ratio). Notably, we observe that when the correlation exceeds 0.3, the coverage approaches the target $1 - \alpha$. These results indicate that exact density ratio values might be unnecessary—what matters is preserving the relative ordering of likelihoods. This tolerance to estimation noise makes the method practically robust.

[4]See Section E, Figure 10 for details about missingness-wise coverage results

5 CONCLUSION AND FUTURE WORKS

In this work, we addressed the challenge of constructing prediction sets with missing values while maintaining both MV and MCV. We introduce two novel strategies—Mask-Conditional Weighted CP and ARC CP—that overcome the key limitations of existing methods under general missingness mechanisms, including MCAR, MAR, and MNAR. Specifically, we addressed the conservativeness of CP-MDA-Nested and the impractical sample requirements of CP-MDA-Exact by leveraging generalized distributional imputation and weighted CP.

Extensive experiments on synthetic and real-world datasets demonstrated the effectiveness of our methods. Compared to prior approaches, our methods significantly reduce the width of prediction set while preserving the desired validity properties. Notably, our ARC method achieves a competitive balance between computational efficiency and predictive informativeness. It reduces 90% of the computing time compared to traditional weighted CP in the case of distribution shift (see Section E.1), making it a promising substitution candidate.

In summary, we introduced a new framework called preimpute–mask–then–correct framework for conformal prediction with missing data, together with two correction methods that guarantee MCV under general missingness mechanisms. Future work includes deriving tighter robustness bounds under likelihood ratio estimation error or violations of Assumption 2, and developing adaptive strategies for high-dimensional or imperfect imputations.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions, which have helped improve the paper. This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie COFUND grant agreement No 101127936 (DeMythif.AI). This research was also supported by

DATAIA Convergence Institute as part of the « Programme d'Investissement d'Avenir », (ANR-17-CONV-0003) operated by Université Paris-Saclay.

References

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4): 494–591, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Dimitris Bertsimas, Arthur Delarue, and Jean Pauphilet. Simple imputation rules for prediction with missing data: Theoretical guarantees vs. empirical performance. *Transactions on Machine Learning Research*, 2024.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. *arXiv preprint arXiv:2309.13793*, 2023.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25 (328):1–66, 2024.

Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big data*, 8:1–37, 2021.

Hadi Fanaee-T. Bike Sharing. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5W894.

Shai Feldman and Yaniv Romano. Robust conformal prediction using privileged information. *arXiv preprint arXiv:2406.05405*, 2024.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pages 1651–1661. PMLR, 2020.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25 (162):1–36, 2024.

Yu Gui, Rina Barber, and Cong Ma. Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36:4820–4844, 2023.

Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning*, pages 9916–9937. PMLR, 2022.

Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5): 402–406, 2013.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

Ziyi Liang, Tianmin Xie, Xin Tong, and Matteo Sesia. Structured conformal inference for matrix completion with applications to group recommender systems. *arXiv preprint arXiv:2404.17561*, 2024.

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.

Angel David Reyero Lobo, Alexis Ayme, Claire Boyer, and Erwan Scornet. A primer on linear classification with missing data. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Vaem: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, 33:11237–11247, 2020.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5), October 2006. ISSN 0090-5364. doi: 10.1214/009053606000000786.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Jeffrey Näf, Erwan Scornet, and Julie Josse. What is a good imputation under mar missingness? *arXiv preprint arXiv:2403.19196*, 2024.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107: 107501, 2020.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pages 844–853. PMLR, 2021.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *Forty-first International Conference on Machine Learning*, 2024.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Lijuan Ren, Tao Wang, Aicha Sekhari Seklouli, Haiqing Zhang, and Abdelaziz Bouras. A review on missing values for main challenges and methods. *Information Systems*, page 102268, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

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

Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91 (434):473–489, 1996.

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

Matteo Sesia, Stefano Favaro, and Edgar Dobriban. Conformal frequency estimation using discrete sketched data with coverage for distinct queries. *Journal of Machine Learning Research*, 24(348):1–80, 2023.

Matteo Sesia, YX Rachel Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae114, 2024.

Meijia Shao and Yuan Zhang. Distribution-free matrix prediction under arbitrary missing pattern. *arXiv preprint arXiv:2305.11640*, 2023.

Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. Pac prediction sets under label shift. *arXiv preprint arXiv:2310.12964*, 2023.

Jake A Soloff, Rina Foygel Barber, and Rebecca Willett. Bagging provides assumption-free stability. *Journal of Machine Learning Research*, 25(131):1–35, 2024.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28 (1):112–118, 2012.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes: with applications to statistics*, pages 16–28. Springer, 1996.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5PK67.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

Seongwook Yoon and Sanghoon Sull. Gamin: Generative adversarial multiple imputation network for highly missing data. In *Proceedings of the IEEE /CVF conference on computer vision and pattern recognition*, pages 8456–8464, 2020.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.

Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. In *International Conference on Machine Learning*, pages 40578–40604. PMLR, 2023.

Margaux Zaffran, Julie Josse, Yaniv Romano, and Aymeric Dieuleveut. Predictive uncertainty quantification with missing covariates, 2024.

Helen Zhou, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under missingness shift. In *International Conference on Artificial Intelligence and Statistics*, pages 9577–9606. PMLR, 2023.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. Conformal prediction: A data perspective. *ACM Computing Surveys*, 2025.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] see Table 3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] We will upload it along with the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Discussion about quality of pre-imputation for calibration dataset

In our work, only the calibration dataset requires imputation. For the basemodel $\hat{f} : \mathcal{X}_{\text{NA}} \to \mathcal{Y}$ that is trained on the training data, it could contain implicitly an imputation (e.g. $\hat{f}$ could be a composition of a neural network $f_{\text{NN}} : \mathcal{X} \to \mathcal{Y}$ and an imputation function φ, such that $\hat{f} = f_{\text{NN}} \circ \varphi$) or it could be simple a tree model without imputation. Therefore, our method can be intergrated for any basemodel $\hat{f}$.

However, a key step for our method is the pre-imputation on calibration dataset. This section discusses the significance of pre-imputation quality for calibration datasets in uncertainty quantification. We begin by reviewing advanced imputation methods for missing values, followed by a rationale for adopting distributional imputations. Finally, we compare the performance of deterministic and distributional imputation methods through experiments.

Imputation methods can be broadly categorized into three types (Jarrett et al., 2022): (1) **Simple value imputation**: Methods such as mean or median imputation use representative values for missing data without reconstructing the original data distribution.(2) **Iterative imputation** (Liu et al., 2014; Zhu and Raghunathan, 2015):, such as MICE (Rubin, 1996), Missforest (Stekhoven and Bühlmann, 2012), etc. In these approaches, the conditional distributions of each feature are estimated based on all the others, and missing values are imputed using such univariate models in a round-robin fashion until convergence. More recently, Näf et al. (2024) proposed mice-Distributional Random Forest (DRF) by leveraging the ability to approximate the conditional distribution of DRF and argued that for MAR condition, it is important to have a distributional imputation function rather than deterministic imputation, one of the three criteria they proposed for being a good imputation candidate. All these methods require a pre-defined functional form for each variable. (3) **Deep generative models**: such as GAN-based imputations (Yoon et al., 2018; Yoon and Sull, 2020) or VAE-based imputations (Du et al., 2023; Nazabal et al., 2020; Ma et al., 2020; Mattei and Frellsen, 2019; Fortuin et al., 2020). These approaches may leverage better capacity and efficiency of learning using deep function approximators and the ability to capture correlations among covariates by amortizing the parameters (Jarrett et al., 2022). While powerful, these methods typically require large datasets and may be computationally expensive.

However, when it comes to imputation in an $X - Y$ type dataset with missing values in X, the traditional imputation is derived only as a function of $X_{\text{obs}(M)}$ predicting the missing part $X_{\text{mis}(M)}$, which limits the potential of imputation in a more general scenario. Under such a definition and for *impute-then-regression* tasks, Le Morvan et al. (2021) has shown that "imputation doesn't matter" in the sense that a universal learner can achieve the Bayes optimal for almost all imputation functions. Yet, for uncertainty quantification and coverage guarantees, this is not sufficient, so we propose a broader definition of imputation, extending to distributional and generalized distributional imputations. We will demonstrate that a good imputer can help achieve the target coverage for the prediction interval if it employs generalized distributional imputation.

In the main text, we do not give a formal definition of the imputation function; instead, we use only $(\hat{X}, Y)$ to represent imputed points. In this part, we formalize three types of imputation functions as follows and use φ to represent an imputation function such that $\hat{X} \overset{d}{=} \varphi(\widetilde{X})$ or $\hat{X} \overset{d}{=} \varphi(\widetilde{X}, Y)$, depending on whether φ takes Y as input or not.

Definition A.1 (Deterministic imputation function (Dt))**.** We say $\varphi_{Dt} : \mathcal{X}_{\text{NA}} \to \mathcal{X}$ is a deterministic imputation function, if it satisfies $\forall x \in \mathcal{X}, m \in \{0,1\}^d$

$$\varphi_{Dt}(\text{mask}(x, m))_{\text{obs}(m)} = x_{\text{obs}(m)}.$$

Definition A.2 (Distributional imputation function (Ds))**.** We say a function $\varphi_{Ds} : \mathcal{X}_{\text{NA}} \to \mathcal{Z}(\mathcal{X})$ where $\mathcal{Z}(\mathcal{X})$ denotes the space of random variables over $\mathcal{X}$ is a distributional imputation function if it satisfies $\forall x \in \mathcal{X}, m \in \{0,1\}^d$:

$$P(\varphi_{Ds}(\text{mask}(x, m))_{\text{obs}(m)} = x_{\text{obs}(m)}) = 1.$$

The missing part of the covariate, $X_{\text{mis}(M)}$, is then imputed by $\varphi_{Ds}(X)_{\text{mis}(M)}$.

The distributional imputation function can be seen as the first step of multiple imputations (MI) (Rubin, 1996), which fuses multiple estimations.

Definition A.3 (Generalized distributional imputation function (GDs))**.** We say a function $\varphi_{GDs} : \mathcal{X}_{\text{NA}} \times \mathcal{Y} \to \mathcal{Z}(\mathcal{X})$, is a generalized distributional imputation function if it satisfies $\forall x \in \mathcal{X}, m \in \{0,1\}^d, y \in \mathcal{Y}$:

$$P(\varphi_{GDs}(\text{mask}(x,m),y)_{\text{obs}(m)} = x_{\text{obs}(m)}) = 1.$$

The missing part of the covariate, $X_{\text{mis}(M)}$, is then imputed by $\varphi_{GDs}(\widetilde{X}(M), Y)_{\text{mis}(M)}$.

Here, we present an additional experiment to compare the effectiveness of deterministic and distributional imputations on synthetic datasets (section 4.1). To simulate **deterministic imputation**, we disable posterior sampling in iterative Bayesian Ridge imputation, resulting in fixed imputed values.

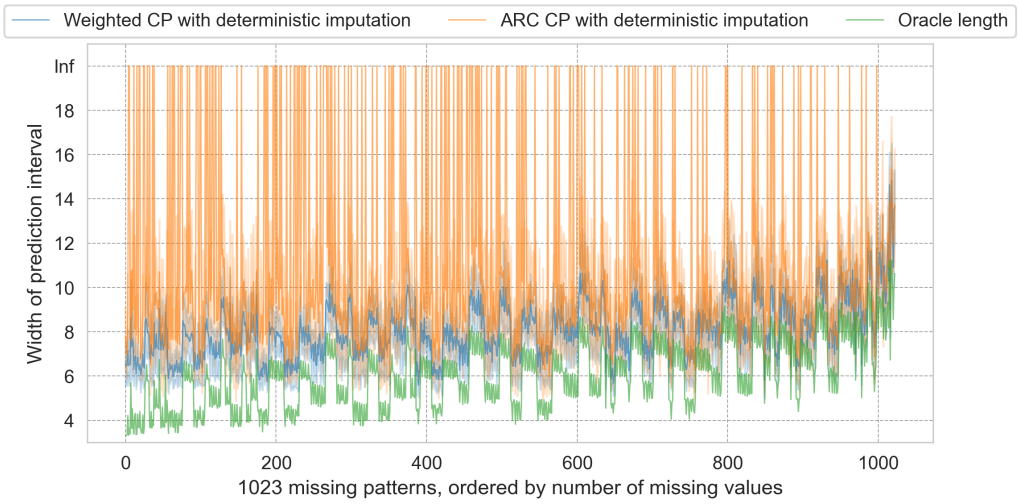

Figure 4: Synthetic dataset: mean and 99% confidence interval of prediction interval width with respect to 1023 missingness patterns, results by using deterministic imputation.

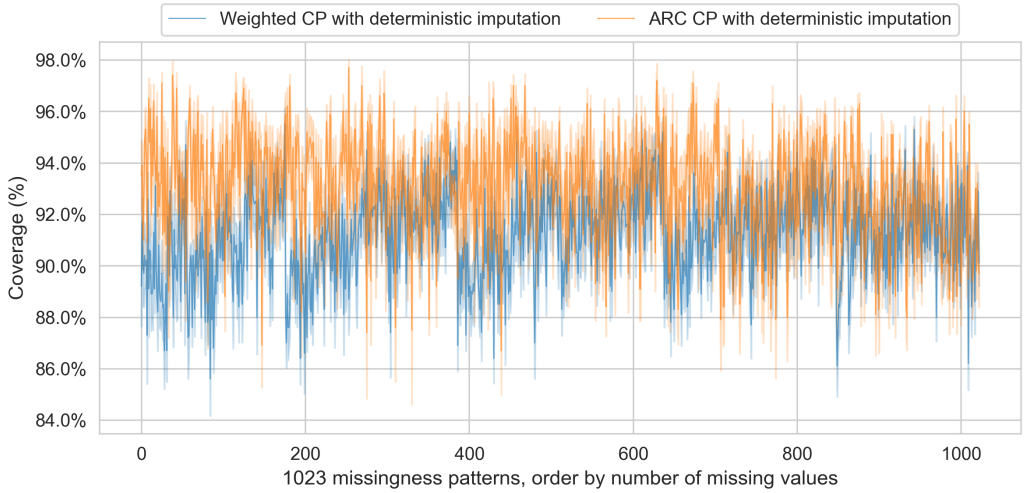

Figure 5: Synthetic dataset: mean and 99% confidence interval of prediction coverage with respect to 1023 missingness patterns, results by using deterministic imputation.

Compared with Figure 2 and Figure 3, we observe from Figures 4 and 5 that if we use a deterministic imputation instead of a distributional imputation, and keep everything else the same, the results may be suboptimal. This outcome aligns with our expectations: if we impute missing values using a deterministic function, the resulting imputed distribution would lack continuity and violate the absolute continuity assumption in Assumption 2, leading to two potential drawbacks:

1. For Weighted CP: if the likelihood ratio ω_m of a specific point is a high value, say $\omega_m(x_m^j, y^j) \approx \infty$, it is highly likely that $p_m^j = 1$ from definition of Theorem 1. Consequently, when computing the weighted quantile as in Equation (3), the weighted quantile will equal s_m^j. In other words, if even one likelihood ratio is extremely large, the weight can easily be influenced by that extreme value, distorting the weighted quantile. This distortion may result in inaccurate quantile computation, which can, in many practical scenarios, fail to guarantee MCV.

2. For ARC CP: if the absolute continuity condition is not satisfied, implying that the likelihood ratio is unbounded (i.e., $K \to \infty$), most points are likely to be rejected during the acceptance-rejection step. This would lead to an insufficient number of samples in the calibration dataset, potentially resulting in infinitely wide prediction intervals for certain missingness patterns

These results highlight that the quality of imputation plays a crucial role in our framework. If the imputation is poor or purely deterministic, the resulting method degenerates, both theoretically and practically, toward CP-MDA-Exact, because the imputed calibration data no longer provides additional distributional information beyond the observed part. In contrast, a well-calibrated distributional imputation allows our method to exploit richer information from the calibration data, effectively correcting the mask-induced distribution shift and achieving tighter prediction intervals with valid coverage. Hence, the pre-imputation step is not only a preprocessing choice, but a key component that governs the efficiency and robustness of our uncertainty quantification framework.

A.1 Sufficient conditions for absolute continuity

In this section we provide a simple sufficient condition ensuring the absolute continuity assumption in Assumption 2 holds under distributional imputation.

Proposition 4 (Sufficient conditions for absolute continuity under distributional imputation)**.** Let P_m be the true distribution of $(X_{\text{obs}(m)}, Y) \mid M = m$, and let Q_m be the distribution of the imputed pair $(\hat{X}_{\text{obs}(m)}, Y)$ as defined in Section 3.1. Assume that P_m and Q_m admit densities p_m and q_m with respect to a common dominating measure μ on $\mathcal{X}_{\text{obs}(m)} \times \mathcal{Y}$ (Lebesgue for continuous components, counting measure for discrete components), i.e. $P_m = p_m \cdot \mu$ and $Q_m = q_m \cdot \mu$. Then P_m is absolutely continuous with respect to Q_m if and only if

$$\forall (x_{\text{obs}}, y) \text{ with } p_m(x_{\text{obs}}, y) > 0, \quad q_m(x_{\text{obs}}, y) > 0. \tag{$\star$}$$

Condition ($\star$) implies that the support of P_m is contained in the support of Q_m. In particular, ($\star$) is satisfied in each of the following cases:

- **Continuous case.** P_m has a density supported on a subset of $\mathcal{X}_{\text{obs}(m)} \times \mathcal{Y}$ without point masses, and the imputation mechanism induces a density q_m that is strictly positive on the same support (e.g. Gaussian or Bayesian regression–based MICE, which has continuous support on the whole domain).

- **Discrete case.** P_m is discrete and supported on a finite or countable set of atoms. Then ($\star$) holds as soon as the imputation distribution assigns non-zero probability to every such atom.

- **Mixed case.** P_m has both a continuous part and atoms, and the imputation distribution has a positive density on the continuous support and assigns non-zero mass to all atoms.

In all three cases the likelihood ratio

$$\omega_m(z) = \frac{dP_m}{dQ_m}(z)$$

is well defined for all z in the support of P_m.

The assumption $P_m \ll Q_m$ is therefore related to the choice of distributional imputer. For MAR/MNAR cases, the missingness does not directly break absolute continuity between P_m and Q_m. In practice, MAR/MNAR make the conditional law harder to estimate, which can increase the likelihood-ratio ω_m and thus the divergence between P_m and Q_m. However, when all the covariates are continuous over $\mathcal{X}$, most of the modern distributional imputation measures lead to $P_m \ll Q_m$, typically, this is the case when Q_m is Gaussian with MICE.

Moreover, as we emphasized at the end of Appendix A (Figures 4 and 5), deterministic imputers (e.g., mean imputation) may violate this condition and cause our methods to fail, whereas simple distributional imputers, such as MICE, yield stable results.

B Implementation details of Weighted CP

As discussed in the main text, weighted CP in its general form involves the test label y during inference, which poses challenges for efficient implementation. Nevertheless, as shown in Table 3, the weighted CP method can still be implemented with success in practice. We have proposed a series of optimizations that significantly reduce the computational burden, making it a viable choice. The optimizations are as follows:

B.1 Search for Interval Bounds

It is often reasonable to assume that the prediction set defined by Equation (3) is an interval $[B_l, B_u]$, otherwise, the prediction set is included in such an interval. This assumption allows us to reduce the computational cost by searching only for the lower and upper bounds of the prediction interval, instead of performing a full grid search across $\mathcal{Y}$.

B.2 Coarse Grid Search

To further accelerate the search, we introduce a step size h and define the search grid as:

$$\mathcal{G} = \{y_{\min}, y_{\min} + h, y_{\min} + 2h, \ldots, y_{\max}\}.$$

We then define the estimated bounds as:

$$\hat{B}_l = \min\left\{y \in \mathcal{G} \mid y \notin \widehat{C}_\alpha(x),\ y + h \in \widehat{C}_\alpha(x)\right\},$$

$$\hat{B}_h = \max\left\{y \in \mathcal{G} \mid y \in \widehat{C}_\alpha(x),\ y + h \notin \widehat{C}_\alpha(x)\right\}.$$

This coarse discretization ensures that the conservativeness of the resulting interval is at most $2h$, while validity is preserved.

B.3 Predefined Search Range

We restrict the search range to a data-dependent interval to avoid unnecessary computation. Specifically, we define:

$$y_{\min} = \min\left\{y \mid s(\tilde{x}, y) \leq \max_{i=1,\ldots,n} S_m^i\right\}, \quad y_{\max} = \max\left\{y \mid s(\tilde{x}, y) \leq \max_{i=1,\ldots,n} S_m^i\right\}.$$

Then the search is performed within $[y_{\min}, y_{\max}]$. To preserve theoretical validity, we use the following rules:

$$\hat{B}_l = -\infty \quad \text{if } y_{\min} \in \widehat{C}_\alpha^W(\tilde{x}), \quad \hat{B}_h = \infty \quad \text{if } y_{\max} \in \widehat{C}_\alpha^W(\tilde{x}).$$

All the above techniques can make the Weighed CP more efficient and valid in theory, though being conservative at $2h$. While the AR-Corrected method is our primary recommendation for practical deployment due to its speed and flexibility, the weighted CP approach remains valuable. In particular, weighted CP avoids the added variance introduced by sub-sampling in the AR correction procedure, making it more stable in scenarios where variance matters.

C Proof

Assume that P, Q, and for any $m \in \{0,1\}^d$ P_m and Q_m admit densities with respect to common dominating measure μ. We denote these densities respectively p, q, p_m, and q_m.

C.1 Proof of Proposition 1

Recall the definition of $Q_{M^{n+1}}$ and $P_{M^{n+1}}$ which are, the distribution of imputed calibration with missingness M^{n+1} and the distribution of test point, conditionally on M^{n+1}. By making a direct Split CP on the two distribution conditionally on $M^{n+1} = m$, by Theorem 2 of (Barber et al., 2023), we have for all $m \in \{0,1\}^d$ such that $P(M = m) > 0$:

$$P(Y^{n+1} \in \hat{C}_\alpha(\widetilde{X}^{n+1}) \mid M^{n+1} = m) \geq 1 - \alpha - 2d_{TV}(Q_m, P_m) \tag{6}$$

For the term of distance of total variation, we have:

$$\begin{aligned}
&2d_{TV}(Q_m, P_m) \\
&= 2 \times \frac{1}{2} \int_{\mathcal{X}^{\text{obs}(m)} \times \mathcal{Y}} \big| p_m(x_{\text{obs}(m)}, y) - q_m(x_{\text{obs}(m)}, y) \big| d(x_{\text{obs}(m)}, y) \\
&= \int_{\mathcal{X}^{\text{obs}(m)} \times \mathcal{Y}} \left| \int_{\mathcal{X}^{\text{mis}(m)}} p(x, y | M = m) dx_{\text{mis}(m)} - \int_{\mathcal{X}^{\text{mis}(m)}} q(x, y) dx_{\text{mis}(m)} \right| d(x_{\text{obs}(m)}, y) \\
&= \int_{\mathcal{X}^{\text{obs}(m)} \times \mathcal{Y}} \left| \int_{\mathcal{X}^{\text{mis}(m)}} \big\{ p(x, y | M = m) - p(x, y) + p(x, y) - q(x, y) \big\} dx_{\text{mis}(m)} \right| d(x_{\text{obs}(m)}, y) \\
&\leq \int_{\mathcal{X}^{\text{obs}(m)} \times \mathcal{Y}} \int_{\mathcal{X}^{\text{mis}(m)}} \big| p(x, y | M = m) - p(x, y) + q(x, y) - q(x, y) \big| dx_{\text{mis}(m)} d(x_{\text{obs}(m)}, y) \\
&\leq \int_{\mathcal{X}^{\text{obs}(m)} \times \mathcal{Y}} \int_{\mathcal{X}^{\text{mis}(m)}} \big\{ |p(x, y | M = m) - p(x, y)| + |p(x, y) - q(x, y)| \big\} dx_{\text{mis}(m)} d(x_{\text{obs}(m)}, y) \\
&= \int_{\mathcal{X} \times \mathcal{Y}} \big\{ |p(x, y | M = m) - p(x, y)| + |p(x, y) - q(x, y)| \big\} d(x, y) \\
&= d_{TV}(P, Q) + \int_{\mathcal{X} \times \mathcal{Y}} \left| \frac{dP(x, y \mid M = m)}{dP(x, y)} - 1 \right| dP(x, y) d(x, y) \\
&= d_{TV}(P, Q) + \mathbb{E}_P \left[\left| \frac{dP(X, Y | M = m)}{dP(X, Y)} - 1 \right| \right]
\end{aligned}$$

For the MCAR case, note that $p(X, Y|M) = p(X, Y)$; therefore, the second term in the last equation is zero.

C.2 Proof of Theorem 1

For any $m \in \{0,1\}^d$, let's consider all event conditionally on $M^{n+1} = m$: by definition of Q_m and that $M^{n+1} \perp \{\hat{X}^i_{\text{obs}(M^{n+1})}, Y^i)\}_{i=1}^n$, we have the imputed-then-masked calibration points $(\{\hat{X}^i_{\text{obs}(M^{n+1})}, Y^i)\}_{i=1}^n \mid M^{n+1} = m) \sim Q_m$ and the test point $((X^{n+1}_{\text{obs}(M^{n+1})}, Y^{n+1}) \mid M^{n+1} = m) \sim P_m$.

Recall that P_m and Q_m are probability defined on #obs(m)-dimensional space. Now consider $d-$dimensional dataset $(\{\text{mask}(\hat{X}^i, M^{n+1}), Y^i\}_{i=1}^n \mid M^{n+1} = m)$, whose distribution is denoted by $Q^\star_m$ and the test point $(\widetilde{X}^{n+1}, Y^{n+1}) \mid M^{n+1} = m)$ with its distribution denoted by P^{*}_m[5]. Suppose $Q^\star_m$ and $P^\star_m$ are dominated by the measure μ with densities $q^\star_m$ and $p^\star_m$, respectively. By definition of the mask operator, one can easily check that:

$$p^\star_m(x, y) = p_m(x_{\text{obs}(m)}, y) \mathbb{1}_{x_{\text{mis}}(m) = NA}$$

$$q^\star_m(x, y) = q_m(x_{\text{obs}(m)}, y) \mathbb{1}_{x_{\text{mis}}(m) = NA}$$

By applying Lemma 2 in (Tibshirani et al., 2019) and Assumption 2, we know that all the $n+1$ points are weighted exchangeable with weight functions $\omega_i = 1$ for $i = 1, ..., n$ and $\omega_{n+1} = \frac{p^\star_m}{q^\star_m} = \frac{p_m}{q_m} = \frac{dP_m}{dQ_m}$.

[5] Or it is equivalent to the distribution mask$(X^{n+1}, M^{n+1}) \mid M^{n+1} = m)$, since $\widetilde{X}^{n+1} = \text{mask}(X^{n+1}, M^{n+1})$

By applying Theorem 2 in (Tibshirani et al., 2019), we deduce immediately that, conditionally on $M^{n+1} = m$, the prediction set $\widehat{C}_\alpha^W(\widetilde{X}^{n+1})$ is exactly the weighted version and thus

$$P(Y^{n+1} \in \widehat{C}_\alpha^W(\tilde{X}^{n+1}) \mid M^{n+1} = m) \geq 1 - \alpha.$$

The formal algorithm is summarized in Algorithm 3:

Algorithm 3 Weighted CP with pre-imputation of Calibration

Input: Calibration set $\text{Cal} = \{(\widetilde{X}^k, Y^k)\}_{k=1}^n$, generalized distributional imputation φ, predictive model $\hat{f}$, nonconformity score function $s_{\hat{f}}(\tilde{x}, y)$, likelihood estimation function $\hat{\omega}_m(\text{mask}(x, m), y)$ for $\omega_m(x_{\text{obs}(m)}, y)$, significance level α, test point $\widetilde{X}^{n+1}$. Imputation function φ and the predictive model $\hat{f}$ are trained on an independent training set.

Output: Prediction set $\widehat{C}_\alpha^W(\widetilde{X}^{n+1})$.

1: **for** $k \in \{1, ..., n\}$ **do**
2: $(\widetilde{X}^k, Y^k) = (\varphi(\widetilde{X}^k, Y^k), Y^k)$ // Pre-impute Cal
3: $(\hat{X}^k, Y^k) = (\text{mask}(X^k, M^{n+1}), Y^k)$ // Add missingness M^{n+1} on each data point
4: $\omega^k = \hat{\omega}_m(\text{mask}(\hat{X}^k, M^{n+1}), Y^k)$ //Compute weight
5: **end for**
6: Define $W^i(x_{\text{obs}(M^{n+1})}, y)$ and $W^{n+1}(x_{\text{obs}(M^{n+1})}, y)$ as in Theorem 1.
7: Compute the weighted version of the prediction set as:

$$\begin{aligned}\widehat{C}_\alpha^W(\widetilde{X}^{n+1}) := \Big\{ & y : s_{\hat{f}}(\widetilde{X}^{n+1}, y) \\ & \leq \mathcal{Q}_{1-\alpha}\Big(\sum_{i=1}^n W^i(X^i_{\text{obs}(M^{n+1})}, y)\delta_{s_{\hat{f}}(\widetilde{X}^i, Y^i)} + W^{n+1}(X^{n+1}_{\text{obs}(M^{n+1})}, y)\delta_\infty\Big)\Big\}. \end{aligned} \tag{7}$$

C.3 Proof of Theorem 2

Before proceeding with the proof, we clarify the assumption made in Theorem 2 that each calibration sample's importance weight ω_m^i is bounded by a constant $K > 0$. In fact, this is equivalent to assuming that the likelihood ratio function $\omega_m(x_{\text{obs}(m)}, y) := \frac{P_m(x_{\text{obs}(m)}, y)}{Q_m(x_{\text{obs}(m)}, y)}$ is uniformly bounded for all $(x_{\text{obs}(m)}, y)$.

Proof. 1. For the first point, since all data have missing value NA for the index $\text{mis}(m)$ of X, we only need to prove that all points have the same marginal distribution on the index $\text{obs}(m)$ of X and Y, and they are independent.

To prove that they have the same distribution, it is sufficient to prove that the conditional distribution of $(\hat{X}^i_{\text{obs}(m)}, Y^i)$ given that $U \leq \frac{\omega_m^i}{K}$ is indeed the same distribution of P_m, That is, in terms of cumulative distribution function, we wish to show that

$$P\left(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y \mid U \leq \frac{\beta}{K}\frac{dP_m}{dQ_m}(\hat{X}_{\text{obs}(m)}, Y)\right) = P(X_{\text{obs}(m)} \leq x, Y \leq y \mid M = m), \tag{8}$$

where β is an unknown normalization constant such that $\omega_m^i = \beta\frac{dP_m}{dQ_m}$. Letting $B = \{U \leq \frac{\beta}{K}\frac{dP_m}{dQ_m}(\hat{X}_{\text{obs}(m)}, Y)\}$ be an event, $A = \{\hat{X}_{\text{obs}(m)} \leq x, Y \leq y\}$, we have firstly by the definition of uniform distribution:

$$P(B \mid \hat{X}_{\text{obs}(m)} = x, Y = y) = \frac{\beta}{K}\frac{dP_m(x, y)}{dQ_m(x, y)},$$

We have also:

$$\begin{aligned}
P(B) &= \int_{\mathcal{X}_{\text{obs}(m)}\times\mathcal{Y}} P(B \mid \hat{X}_{\text{obs}(m)} = x, Y = y) dQ_m(\hat{X}_{\text{obs}(m)} = x, Y = y) \\
&= \int_{\mathcal{X}_{\text{obs}(m)}\times\mathcal{Y}} \frac{\beta}{K} \frac{dP_m(x,y)}{dQ_m(x,y)} dQ_m(x,y) \\
&= \frac{\beta}{K},
\end{aligned}$$

$$\begin{aligned}
Q_m(B \mid A) &= \frac{\int_A \mathbb{P}_{\text{U}\sim\text{Uniform}}\Big(U \leq \frac{\beta}{K}\,\omega_m(z) \mid z\Big)\, dQ_m(z)}{Q_m(A)} && (9) \\
&= \frac{\int_A \frac{\beta}{K}\,\omega_m(z)\, dQ_m(z)}{Q_m(A)} && (10) \\
&= \frac{\beta}{K\,Q_m(A)} \int_A dP_m && (11) \\
&= \frac{\beta\, P_m(A)}{K\,Q_m(A)}. && (12)
\end{aligned}$$

By Bayes's rule:

$$\begin{aligned}
P(B|A) &= \frac{P(B,A)}{P(A)} \\
&= \frac{P(U \leq \frac{\beta}{K}\frac{dP_m}{dQ_m}(\hat{X}_{\text{obs}(m)}, Y), \hat{X}_{\text{obs}(m)} \leq x, Y \leq y)}{P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y)} \\
&= \int_{-\infty}^{(x,y)} \frac{P(U \leq \frac{\beta}{K}\frac{dP_m(w_x,w_y)}{dQ_m(w_x,w_y)} \mid w_x \leq x, w_y \leq y)}{P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y)} dQ_m(w_x, w_y)) \\
&= \frac{1}{P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y \mid M = m)} \int_{-\infty}^{(x,y)} \frac{\beta}{K}\frac{dP_m(w_x,w_y)}{dQ_m(w_x,w_y)} dQ_m(w_x,w_y) \\
&= \frac{\beta}{K \cdot P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y)} \int_{-\infty}^{(x,y)} dP_m(w_x,w_y) \\
&= \frac{\beta \cdot P(X_{\text{obs}(m)} \leq x, Y \leq y \mid M = m)}{K \cdot P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y)}.
\end{aligned}$$

Note that $P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y \mid U \leq \frac{\beta}{K}\frac{dP_m}{dQ_m}(\hat{X}_{\text{obs}(m)}, Y)) = P(A|B)$, therefore we have:

$$\begin{aligned}
P(A|B) &= P(B|A)\frac{P(A)}{P(B)} \\
&= \frac{\beta \cdot P(X_{\text{obs}(m)} \leq x, Y \leq y \mid M = m)}{K \cdot P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y)} \cdot \frac{P(\hat{X}_{\text{obs}(m)} \leq x, Y \leq y)}{\beta/K} \\
&= P(X_{\text{obs}(m)} \leq x, Y \leq y \mid M = m),
\end{aligned}$$

which is the target Equation (8). By applying the acceptance-rejection method on the imputed calibration data, the obtained data points follow the same distribution as $(X_{\text{obs}(m)}, Y)$.

Additionally, because the points in $(\hat{X}^i, Y^i) \cup (X^{n+1}, Y^{n+1})$ are independent with each other by definition and $U \perp\!\!\!\perp (X^k, Y^k)_{k=1}^{n+1}$, all data points in $\widetilde{Cal}_m$ and $(X_{\text{obs}(m)}^{n+1}, Y^{n+1})$ are mutually independent, we conclude that all the data points are i.i.d.

2. As proved before, conditionally on $M = m$, all data points in $\widetilde{Cal}_m$ and $(X_{\text{obs}(m)}^{n+1}, Y^{n+1})$ are i.i.d., satisfying the condition (exchangeability) of Conformal Prediction, yielding for any marginal valid method with target

$1-\alpha$:

$$P(Y^{n+1} \in \widehat{C}_\alpha^{AR}\left(\text{mask}(X^{n+1}, m)\right) \mid M^{n+1} = m) \geq 1-\alpha.$$

□

Algorithm 4 Acceptance-rejection corrected CP with pre-imputation of Calibration

Input: Calibration set $\text{Cal} = \{(\widetilde{X}^i, Y^i)\}_{i=1}^n$, generalized distributional imputation φ, predictive model $\hat{f}$, nonconformity score function $s_{\hat{f}}(\tilde{x}, y)$, likelihood estimation function $\hat{\omega}_m(\text{mask}(x, m), y)$ for $\omega_m(x_{\text{obs}(m)}, y)$ for $\omega_m(x_{\text{obs}(m)}, y)$ with its maximum value ω_{max}, significance level α, test point $\widetilde{X}^{n+1}$. The imputation function φ and the predictive model $\hat{f}$ are trained on an independent training set.
Output: Prediction set $\widehat{C}_\alpha^{AR}(\widetilde{X}^{n+1})$.

$\mathcal{I}_{AR} = \{\}$
for $k \in \{1, ..., n\}$ **do**
  Pre-impute Calibration point $(\varphi(\widetilde{X}^k, Y^k), Y^k)$
  Sample $U^k \sim U(0,1)$
  Compute $q^k = \frac{\hat{\omega}_m(\text{mask}(\varphi(\widetilde{X}^k, Y^k), M^{n+1}),\, Y^k)}{\omega_{\max}}$
  if $U^k < q^k$ **then**
    $\mathcal{I}_{AR} = \mathcal{I}_{AR} \cup \{k\}$
    $\hat{X}^k = \varphi(\widetilde{X}^k, Y^k)$
  end if
end for
$\widetilde{\text{Cal}} = \{(\hat{X}^i, Y^i)\}_{i \in \mathcal{I}_{AR}}$ // $|\mathcal{I}_{AR}|$ points are accepted such that $\{(\hat{X}^i, Y^i)\}_{i \in \mathcal{I}_{AR}} \subseteq \{(\varphi(\widetilde{X}^i, Y^i), Y^i)\}_{i=1}^n$
The updated dataset $\widetilde{\text{Cal}}_{M^{n+1}} = \{\text{mask}(\hat{X}^i, M^{n+1}), Y^i)\}_{i \in \mathcal{I}_{AR}}$. // Mask $\widetilde{\text{Cal}}$ with the missingness of test point M^{n+1}
$\forall i \in \mathcal{I}_{AR}$, denote $\hat{S}^i_{M^{n+1}} = s_{\hat{f}}(\text{mask}(\widetilde{X}^i, M^{n+1}), Y^i)$
Compute the AR-corrected prediction set:

$$\widehat{C}_\alpha^{AR}(\widetilde{X}^{n+1}) := \left\{ y \in \mathcal{Y} : s_{\hat{f}}(\widetilde{X}^{n+1}, y) \leq \mathcal{Q}_{1-\alpha}\left(\sum_{i \in \mathcal{I}_{AR}} \delta_{\hat{S}^i_{M^{n+1}}} + \delta_\infty \right) \right\}. \tag{13}$$

Lemma 1 (Bayes Classification Risk and Total Variation Distance)**.** Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on it. Assume there exists a σ-finite measure μ dominating both, and denote the corresponding densities by $p = \frac{d\mathbb{P}}{d\mu}$ and $q = \frac{d\mathbb{Q}}{d\mu}$.

Define the joint distribution Π on $\Omega \times \{0,1\}$ by

$$\Pi(dx, dc) = \tfrac{1}{2}\,\mathbb{P}(dx)\,\delta_1(dc) + \tfrac{1}{2}\,\mathbb{Q}(dx)\,\delta_0(dc),$$

where δ_c denotes the Dirac measure concentrated at c. Then the marginal distribution of X is

$$\mathbb{D} := \tfrac{1}{2}(\mathbb{P} + \mathbb{Q}).$$

For any measurable classifier $\mathsf{c} : \Omega \to \{0,1\}$, define the (0–1) classification risk as $\mathcal{R}(\mathsf{c}) := P_{(X,C)\sim\Pi}[\mathsf{c}(X) \neq C]$, and the Bayes-optimal risk is $\mathcal{R}^\star := \inf_{\mathsf{c}} \mathcal{R}(\mathsf{c})$. Then we have:

$$\mathcal{R}^\star = \frac{1}{2} - \frac{1}{2} d_{\text{TV}}(\mathbb{P}, \mathbb{Q}),$$

where

$$d_{\text{TV}}(\mathbb{P}, \mathbb{Q}) := \frac{1}{2} \int_\Omega |p(x) - q(x)|\, d\mu(x).$$

Proof. The posterior probability that a sample x is drawn from $\mathbb{P}$ (i.e., $C = 1$) given x is defined as

$$\eta(x) := P(C = 1 \mid X = x) = \frac{\frac{1}{2}p(x)}{\frac{1}{2}p(x) + \frac{1}{2}q(x)} = \frac{p(x)}{p(x) + q(x)}.$$

The marginal distribution of X under the joint law Π is

$$\mathbb{D}(dx) = \tfrac{1}{2}(\mathbb{P} + \mathbb{Q})(dx) = \tfrac{1}{2}(p(x) + q(x))\, d\mu(x).$$

The Bayes-optimal classifier is

$$\mathsf{c}^\star(x) = \mathbf{1}\{\eta(x) > 1/2\},$$

and the corresponding Bayes classification risk is

$$\mathcal{R}(\mathsf{c}^\star) = P_{(X,C)\sim\Pi}[\mathsf{c}^\star(X) \neq C] = \mathbb{E}_{X\sim\mathbb{D}}\big[\min(\eta(X),\, 1 - \eta(X))\big].$$

Substituting the definition of $\eta(x)$ yields

$$\min\left(\frac{p(x)}{p(x) + q(x)}, \frac{q(x)}{p(x) + q(x)}\right) = \frac{\min\{p(x), q(x)\}}{p(x) + q(x)}.$$

Therefore,

$$\mathcal{R}(\mathsf{c}^\star) = \int_\Omega \tfrac{1}{2}(p(x) + q(x))\frac{\min\{p(x), q(x)\}}{p(x) + q(x)}\, d\mu(x) = \tfrac{1}{2}\int_\Omega \min\{p(x), q(x)\}\, d\mu(x).$$

Finally, recall that the total variation distance satisfies

$$d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \tfrac{1}{2}\int_\Omega |p(x) - q(x)|\, d\mu(x) = 1 - \int_\Omega \min\{p(x), q(x)\}\, d\mu(x),$$

which gives

$$\mathcal{R}(\mathsf{c}^\star) = \tfrac{1}{2} - \tfrac{1}{2}\, d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}).$$

□

C.4 Proof of Proposition 2

We fix an $m \in \{0,1\}^d$, let $(X^{\tilde\omega}_{\mathrm{obs}(m)}, Y^{\tilde\omega})$ a random vector following $d\tilde{P}_m := \tilde\omega_m(\hat{X}_{\mathrm{obs}(m)}, Y) \cdot dQ_m$, which is a well defined new distribution via a Radon–Nikodym derivative with respect to Q_m, using a normalized weighting function $\tilde\omega_m$. We define for convenience that $X^{\tilde\omega}_{\mathrm{mis}(m)} := \mathrm{NA}$.

Since $\widehat{C}^{\widehat{W}}_\alpha$ is constructed on calibration dataset Q_m with correction weights $\tilde\omega_m$, it means that the reweighted calibration distribution is equal to the distribution of $(X^{\tilde\omega}_{\mathrm{obs}(m)}, Y^{\tilde\omega})$. By construction, this means that if we use the same Weighted CP method to compute a prediction interval for $X^{\tilde\omega}$, then we have the exact valid coverage:

$$P(Y^{\tilde\omega} \in \widehat{C}^{\widehat{W}}_\alpha(X^{\tilde\omega})) \geq 1 - \alpha. \tag{14}$$

Besides, we note that, by definition of likelihood ratio ω_m, the real $(X_{\mathrm{obs}(m)}, Y)$ follows the distribution $P_m = \omega_m(\hat{X}_{\mathrm{obs}(m)}, Y) \cdot Q_m$. Denote by A the event of Y being in the prediction interval $\widehat{C}^{W}_\alpha(\widetilde{X})$ conditionally on $M^{n+1} = m$, we have, by the basic property of total variation distance:

$$\begin{aligned}
&|P(Y^{n+1} \in \widehat{C}^{\widehat{W}}_\alpha(\widetilde{X}^{n+1}) \mid M^{n+1} = m) - P(Y^{\tilde\omega} \in \widehat{C}^{\widehat{W}}_\alpha(X^{\tilde\omega}) \mid M^{n+1} = m)| \\
=&|P_{(X_{\mathrm{obs}(m)},Y)\sim P_m}(A) - P_{(X_{\mathrm{obs}(m)},Y)\sim\tilde\omega_m(\hat{X}_{\mathrm{obs}(m)},Y)\cdot dQ_m}(A)| \\
=&|P_{(X_{\mathrm{obs}(m)},Y)\sim P_m}(A) - P_{(X_{\mathrm{obs}(m)},Y)\sim\tilde{P}_m}(A)| \\
\leq& d_{TV}(P_m, \tilde{P}_m)
\end{aligned}$$

Combine Equation (14) with the previous results, we have the result of Proposition 2.

C.5 Regularity assumptions

Notation. Fix a mask m. P_m and $\tilde{P}_m$ are probability measures on $\mathcal{X}_{\text{obs}(m)} \times \mathcal{Y}$. Introduce the balanced mixture on $(\mathcal{X}_{\text{obs}(m)} \times \mathcal{Y}) \times \{0,1\}$:

$$\Pi_m \;=\; \tfrac{1}{2}\big(P_m \times \delta_1\big) \;+\; \tfrac{1}{2}\big(\tilde{P}_m \times \delta_0\big),$$

where δ_c denotes the Dirac mass at $c \in \{0,1\}$. Let $\mathcal{G} \subset \{\, g : (\mathcal{X}_{\text{obs}(m)} \times \mathcal{Y}) \to \{0,1\} \,\}$ be a hypothesis class. Define the 0–1 loss

$$\gamma\big(g, (x_{\text{obs}(m)}, y), c\big) = \mathbf{1}\{\, g(x_{\text{obs}(m)}, y) \neq c \,\} \in [0,1],$$

its risk

$$R_\gamma(g) := \mathbb{E}_{((X_{\text{obs}(m)},Y),C)\sim\Pi_m}\big[\gamma(g, (X_{\text{obs}(m)}, Y), C)\big],$$

and the Bayes risk $R^\star_\gamma := \inf_g R_\gamma(g)$. By Lemma 1, $R^\star_\gamma = \frac{1}{2} - \frac{1}{2}\, d_{\text{TV}}(P_m, \tilde{P}_m)$.

We restrict the search for the classifier g_m to a hypothesis class $\mathcal{G}_m$, from which it is obtained via empirical minimization of the 0–1 loss defined above.

Assumption 3 (Separability of $\mathcal{G}_m$)**.** There exists a countable $\mathcal{G}' \subset \mathcal{G}_m$ such that for every $g \in \mathcal{G}_m$, there is a sequence $(g_k) \subset \mathcal{G}'$ with $\gamma\big(g_k, (x_{\text{obs}(m)}, y), c\big) \to \gamma\big(g, (x_{\text{obs}(m)}, y), c\big)$ for all $((x_{\text{obs}(m)}, y), c)$.

Assumption 4 (VC-class)**.** Let $\mathcal{A}_{\mathcal{G}_m} := \{\, \{(x_{\text{obs}(m)}, y) : g_m(x_{\text{obs}(m)}, y) = 1\} \;:\; g_m \in \mathcal{G}_m\}$ be the collection of decision sets induced by $\mathcal{G}_m$. We assume $\mathcal{A}_{\mathcal{G}_m}$ is a VC-class with finite dimension $V \geq 1$.

Assumption 5 (ERM over $\mathcal{G}_m$)**.** Given l i.i.d. samples from Π_m, let $\hat{g}_m \in \arg\min_{g\in\mathcal{G}_m} \hat{R}_\gamma(g)$ be an empirical risk minimizer of γ over $\mathcal{G}$.

Assumption 6. There exists $g^\star_m \in \mathcal{G}_m$ such that $R_\gamma(g^\star_m) = R^\star_\gamma$.

Assumption 7 (Uniform β-stability of the training algorithm)**.** Let $\check{Z} = (Z^1, \ldots, Z^l) \in \mathcal{Z}^l$ with $Z_i = (X^i_{\text{obs}(m)}, Y^i, C_i)$ i.i.d. $\sim \Pi_m$. Let $\mathcal{T}$ be a learning algorithm that outputs a classifier $\hat{g}_m = \mathcal{T}(\check{Z}) \in \mathcal{G}_m$. For $i \in \{1, \ldots, l\}$, define the dataset $\check{Z}^{(i)}$ by replacing Z^i with an independent copy $Z' \sim \Pi_m$, i.e., $\check{Z}^{(i)} = (Z^1, \ldots, Z^{i-1}, Z', Z^{i+1}, \ldots, Z^l)$. We say that the algorithm $\mathcal{T}$ is *uniformly β_l-stable* (with respect to the loss γ) if there exists $\beta_l \geq 0$ such that for all $i \in \{1, \ldots, l\}$ and all $z = (x_{\text{obs}(m)}, y, c) \in \mathcal{X}_{\text{obs}(m)} \times \mathcal{Y} \times \{0,1\}$,

$$\big|\gamma(\mathcal{T}(\check{Z}), z) - \gamma(\mathcal{T}(\check{Z}^{(i)}), z)\big| \leq \beta_l \quad \text{almost surely (w.r.t. } \Pi_m^{l+1}, \text{the joint distribution of } \check{Z}, Z').$$

These assumptions are inspired by classical statistical learning theory (Massart and Nédélec, 2006; Van Der Vaart and Wellner, 1996; Vapnik, 2013) and developments on algorithmic properties and stability (Bousquet and Elisseeff, 2002; Soloff et al., 2024). For example, kernel methods, regularized logistic regression, and neural networks with bounded weights satisfy separability and finite VC-dimension assumptions. Tree-based ensembles such as Random Forests or Gradient Boosted Trees also satisfy these assumptions when the number and depth of trees are bounded. Assumption 7 (Uniform β_l-stability) is trivially satisfied with zero-one loss. For generalizations, see Bousquet and Elisseeff (2002); Soloff et al. (2024).

C.6 Proof of Proposition 3

(1). Suppose we have another dataset of size l which is denoted by $Z_c := \{(\check{X}^i, \check{Y}^i, \check{C}^i)\}_{i=1}^l$ and an estimator of likelihood ratio $\hat{\omega}_m$. Our goal is to provide an estimation for $d_{TV}(P_m,\ \tilde{\omega}_m(\hat{X}_{\text{obs}(m)}, Y) \cdot Q_m)$.

Since $\tilde{\omega}_m$ is the normalized estimator of likelihood ratio, $\tilde{P}_m := \tilde{\omega}_m(\hat{X}_{\text{obs}(m)}, Y) \cdot Q_m$, which defines a valid probability distribution.

Under Assumptions 3 and 4, by Corollary 3 in Massart and Nédélec (2006), we have

$$\mathbb{E}_{Z_c}\big[R_\gamma(\hat{g}_m)\big] - R^\star_\gamma \;\leq\; K\sqrt{\tfrac{V_{\mathcal{G}}}{l}},$$

where the K is an universal constant independent of the sample size and the distribution and $V_{\mathcal{G}}$ is the VC-dimension of the hypothesis class $\mathcal{G}$ as in Assumption 4.

Algorithm 5 Construction of datasets from P_m and $\tilde{P}_m$ via acceptance-rejection sampling

Suppose we are given J i.i.d. samples $\{(\check{X}^i, \check{Y}^i, \check{M}^i)\}_{i=1}^J$. We first apply the same pre-imputation function used in the main text to obtain a fully observed dataset that follows distribution Q by construction.

We then apply the fixed mask m to each data point, yielding masked samples $\{(\tilde{\check{X}}^i, \check{Y}^i)\}_{i=1}^J := \{(\text{mask}(\check{X}^i, m), \check{Y}^i)\}_{i=1}^J$, which follow the distribution Q_m.

To generate i.i.d. samples from $\tilde{P}_m$, defined via the likelihood ratio $\tilde{\omega}_m$ as $d\tilde{P}_m := \tilde{\omega}_m \, dQ_m$, we apply an acceptance-rejection scheme with proposal distribution Q_m:

1. Select a candidate $Z^i = (\tilde{\check{X}}^i, \check{Y}^i)$ from the dataset, and draw $U \sim \text{Unif}(0, 1)$ independently.
2. Accept Z^i if $U \leq \tilde{\omega}_m(Z^i)/\tilde{K}$, where $\tilde{K} \geq \sup_z \tilde{\omega}_m(z)$.

Each accepted point is an exact draw from $\tilde{P}_m$, and repeating the procedure yields an i.i.d. dataset of size l_2 from $\tilde{P}_m$.

To construct the dataset from P_m, we select all samples from the original collection $\{(\check{X}^i, \check{Y}^i, \check{M}^i)\}_{i=1}^J$ that satisfy $\check{M}^i = m$. These points follow P_m by definition.

This procedure yields two independent datasets, each consisting of i.i.d. samples from P_m and $\tilde{P}_m$, respectively.

(2). Consider the construction of prediction interval $\hat{C}_\alpha^W$, which is independent of the dataset Z_c. Therefore, we can apply the Proposition 2 and have

$$P\left(Y^{n+1} \in \hat{C}_\alpha^W(\tilde{X}^{n+1}) \mid M^{n+1} = m\right) - (1-\alpha) \geq -d_{TV}(P_m, \tilde{P}_m) \tag{15}$$

$$= -(1 - 2R_\gamma^\star)\,, \tag{16}$$

where the last equality follows from Lemma 1. Since $\hat{C}_\alpha^W$ is constructed independently of Z_c, the right-hand side of eq. (15) can be rewritten as

$$\begin{aligned} & P(Y^{n+1} \in \hat{C}_\alpha^W(\tilde{X}^{n+1}) \mid M^{n+1} = m) - (1-\alpha) \\ \geq & -(1 - 2\,R_\gamma^\star) = -1 + 2\,R_\gamma^\star \\ \geq & -1 + 2\mathbb{E}_{Z_c}[R_\gamma(\hat{g}_m)] - K\sqrt{\frac{V_{\mathcal{G}_m}}{l}}\,, \end{aligned} \tag{17}$$

where $K > 0$ is a universal constant that may absorb absolute numerical constants such as 2. Under Assumption 7, we can apply McDiarmid's inequality (McDiarmid et al., 1989) to obtain the following bound: for any $\epsilon > 0$, $P_{Z_c}\left(R_\gamma(\hat{g}_m) - \mathbb{E}_{Z_c}[R_\gamma(\hat{g}_m)]) \geq \epsilon\right) \leq \exp(-\frac{2\epsilon^2}{l\beta_l^2})$. Since all the risk are bounded by 0 and 1, we apply Hoeffding's inequality and get $P_{Z_c}(\hat{R}_\gamma(\hat{g}_m) - R_\gamma(\hat{g}_m) \geq \epsilon) \leq \exp(-\frac{2\epsilon^2 l^2}{l})$.

Replacing ϵ by $\epsilon/2$, we have

$$P_{Z_c}(|\hat{R}_\gamma(\hat{g}_m) - \mathbb{E}_{Z_c}[R_\gamma(\hat{g}_m)]| \leq \epsilon) > 1 - \exp(-\frac{\epsilon^2}{2l\beta_l^2}) - \exp(-\frac{\epsilon^2 l}{2}). \tag{18}$$

Substituting into Equation (17), we obtain the desired result.

D Impact of Estimator Quality on Coverage

In this section, we study the practical implications of density ratio estimation quality on the mask-conditional coverage of weighted conformal prediction. The analysis builds on synthetic data where the true density ratios are analytically tractable.

D.1 Synthetic Data Generation

We consider the same synthetic setup as in section 4. Let $X \in \mathbb{R}^d$ be a d-dimensional Gaussian random variable and Y be the linear response with additive Gaussian noise:

$$X \sim \mathcal{N}(\mu_X, \Sigma_X),$$
$$Y \mid X = \beta^\top X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

The joint distribution (X, Y) is Gaussian with mean and covariance:

$$\mu_{XY} = \begin{bmatrix} \mu_X \\ \beta^\top \mu_X \end{bmatrix},$$
$$\Sigma_{XY} = \begin{bmatrix} \Sigma_X & \Sigma_X \beta \\ \beta^\top \Sigma_X & \beta^\top \Sigma_X \beta + \sigma^2 \end{bmatrix}.$$

We simulate Missing Completely at Random (MCAR) by randomly masking each feature X_j independently with probability 0.5. Let $M \in \{0, 1\}^d$ denote the missingness indicator.

By Gaussian conditioning, for each missing pattern $M = m$, the conditional distribution of the missing covariates given observed covariates and Y is:

$$X_{\text{mis}} \mid X_{\text{obs}}, Y \sim \mathcal{N}(\mu_{\text{mis}|\text{obs},Y}, \Sigma_{\text{mis}|\text{obs},Y}),$$

where:

$$\mu_{\text{mis}|\text{obs},Y} = \mu_{\text{mis}} + \Sigma_{\text{mis},[X_{\text{obs}},Y]} \Sigma^{-1}_{[X_{\text{obs}},Y]} \left(\begin{bmatrix} X_{\text{obs}} & Y \end{bmatrix} - \mu_{[X_{\text{obs}},Y]} \right),$$
$$\Sigma_{\text{mis}|\text{obs},Y} = \Sigma_{\text{mis},\text{mis}} - \Sigma_{\text{mis},[X_{\text{obs}},Y]} \Sigma^{-1}_{[X_{\text{obs}},Y]} \Sigma_{[X_{\text{obs}},Y],\text{mis}}.$$

D.2 Perturbed Imputation Strategy

To emulate estimation noise in imputation, we perturb the conditional distribution:

$$\varphi(\tilde{x}, y)_{\text{mis}} \sim \mathcal{N}(\tilde{\mu}_{\text{mis}}, \tilde{\Sigma}_{\text{mis}}),$$

with:

$$\tilde{\mu}_{\text{mis}} = \alpha \cdot \mu_{\text{mis}|\text{obs},Y} + \Delta,$$
$$\tilde{\Sigma}_{\text{mis}} = \Sigma_{\text{mis}|\text{obs},Y} + \tau I,$$

where $\alpha > 0$, $\Delta \in \mathbb{R}^{|\text{mis}|}$ is a bias term, and $\tau > 0$ is variance inflation.

As discussed in the main text, the key ingredient for weighted conformal prediction is the likelihood ratio:

$$\omega_m(x_{\text{obs}}, y) = \frac{dP_m(x_{\text{obs}}, y)}{dQ_m(x_{\text{obs}}, y)},$$

where P_m is the true mask-conditional distribution of $(X, Y)|M = m$, and Q_m is the induced distribution from imputed calibration data. In this setup, both are computable: P_m is the original Gaussian, and Q_m is a mixture Gaussian from the perturbed imputation. Under our perturbed imputation strategy, all the terms are tractable.

D.3 Empirical Evaluation

We perform controlled simulations under fixed mask $m = [1, 1, 1, 0, 0]$ with $d = 5$, $\mu_X = \mathbf{1}_d$, and covariance $\Sigma_X = \rho \mathbf{1}_d \mathbf{1}_d^\top + (1-\rho) I$, $\rho = 0.8$. We generate 500 training, 200 calibration, and 100 test samples. For perturbation, we set $\Delta = 0.5, \alpha = 1.5, \tau = 0.2$. The target coverage is 90%. We then evaluate the following:

1. **No correction:** miscoverage due to distribution shift.

2. **Oracle correction:** use exact ω_m.
3. **Estimated correction:** learn an estimator $\hat{\omega}_m$ of ω_m using algorithm 2.
4. **Perturbed estimation:** apply controlled noise to $\hat{\omega}_m$ test robustness of different qualities of density estimators (see Section D.4 for details).

Each setting is repeated 100 times to compute coverage statistics and its sensitivity to estimator quality. We find that:

- **Uncorrected CP** underestimates coverage (86%) due to test/calibration shift.
- **Oracle-corrected CP** achieves near-perfect coverage (89–90%).
- **Estimated weights** by using Algorithm 2 achieve our the desired coverage, though showing conservative behavior (93% coverage).
- **Estimator degradation** leads to systematic under-coverage as correlation between estimated and true weights drops (see Section D.4 for details).

D.4 Impact of Controlled Perturbation in Density Ratio Estimation

To examine empirically the robustness of our method under imperfect density ratio estimation, we introduce controlled perturbations. This complements the theoretical and empirical findings discussed in the main text and addresses concerns regarding the sensitivity of our approach to density ratio quality.

Noise Injection into Estimated Density Ratio. In our implementation, density ratios are estimated using Algorithm 2 (a histogram-based gradient boosting classifier). To simulate various levels of estimation quality, we add Gaussian noise to the logarithm of the estimated density ratios:

$$\log \tilde{\omega}_m = \log \hat{\omega}_m + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

then exponentiate back to obtain noisy $\tilde{\omega}_m$. This ensures the perturbation is multiplicative and realistically models deviations in the ratio space.

Empirical Evaluation. For each of 100 repetitions, we pertube the estimator of density ratio and we evaluate the quality of density ratio estimator on calibration data using the Pearson correlation coefficient between true and estimated density ratio values. We then measure the mask-conditional coverage obtained from our weighted CP method. Building on this, for each repetition, we have a different quality of density ratio estimator (in terms of Pearson correlation) and an associated coverage. We then draw the scatter plot of the correlation against empirical coverage in Figure 6. We observe a strong positive relationship: when correlation exceeds 0.3, coverage stabilizes around the nominal level of 90%, demonstrating the method's robustness to moderate estimation error.

D.5 Evaluation of theoretical Miscoverage Bound.

Our goal is to empirically verify the validity of Proposition 3, which asserts that the miscoverage due to imperfect density ratio estimation can be upper bounded by quantities estimated from data. Specifically, we construct the distributions P_m and $\tilde{P}_m$ using an auxiliary dataset of 500 i.i.d. samples, train a classifier $\hat{g}_m$ to distinguish them, and evaluate whether the resulting classification error yields a valid upper bound on the observed miscoverage.

By Section C.6, we have:

$$P(Y \in \widehat{C}_\alpha^{\widetilde{W}}(\tilde{X}) \mid M = m) - 90\% \geq -(1 - 2\widehat{\mathcal{R}}(\hat{g}_m)).$$
$$\leftrightarrow 90\% - P(Y \in \widehat{C}_\alpha^{\widetilde{W}}(\tilde{X}) \mid M = m) \leq (1 - 2\widehat{\mathcal{R}}(\hat{g}_m)).$$

For each repetition of experiment, we can compute a miscoverage by

$$\text{Miscoverage} = (90\% - \text{True coverage}).$$

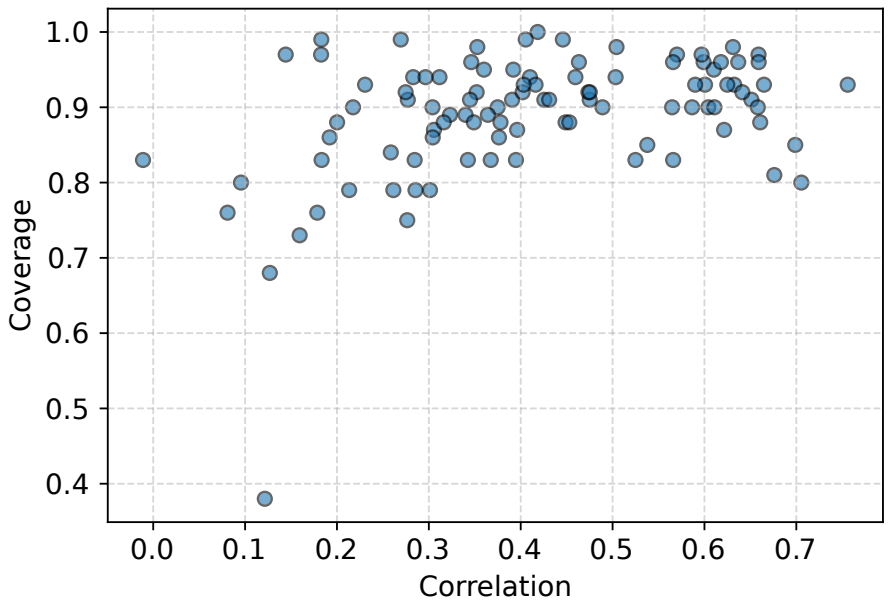

Figure 6: Coverage vs. Correlation of estimated and true density ratios. Each point represents one experiment. High correlation yields near-nominal coverage.

We can also compute the empirical error $\widehat{\mathcal{R}}(\hat{g}_m)$ of the classifier between P_m and $\tilde{P}_m$ by using an extra dataset , and the empirical miscoverage $(1 - 2\widehat{\mathcal{R}}(\hat{g}_m))$. Ideally, we should observe that all the miscoverage is controlled by $(1 - 2\widehat{\mathcal{R}}(\hat{g}_m))$. In Figure 7, we compare empirical miscoverage with the theoretical bound $(1 - 2\widehat{\mathcal{R}}(\hat{g}_m))$. Most observed miscoverage values lie below the red line $y = x$, verifying that the bound is valid.

Together, these findings demonstrate that our method is empirically robust to imperfect density ratio estimation, and theoretical bounds on miscoverage can be derived and validated in practice.

E Detailed experiments

E.1 Running time comparison

We compare the running time of all 4 methods on the same dataset, i.e., 100 points for each missingness pattern. All experiments are running on the same server with CPU Intel XeonGold 6252 @ 2.10GHz. The running time of 4 methods on two datasets is shown in Table 3. Our proposed AR-Corrected method has a running time comparable to MDA-Exact, which is much faster than MDA-Nested and weighted CP. This aligns with our expectation since the MDA-Nested requires modifying test points and calibration points differently, and the weighted method requires using a "grid search" like method to evaluate the prediction interval, which cannot be solved explicitly. Therefore, our proposed acceptance rejection-based method can be a good candidate for substituting weighted conformal prediction, especially when the weighted CP interval has no simple and explicit form.

Table 3: Comparison of running time of 4 methods.

Method	Synthetic dataset	Concrete dataset
MDA-Exact	17034s	2031s
MDA-Nested	510932s	43225s
Weighted CP	228638s	37983s
ARC CP	21320s	4766s

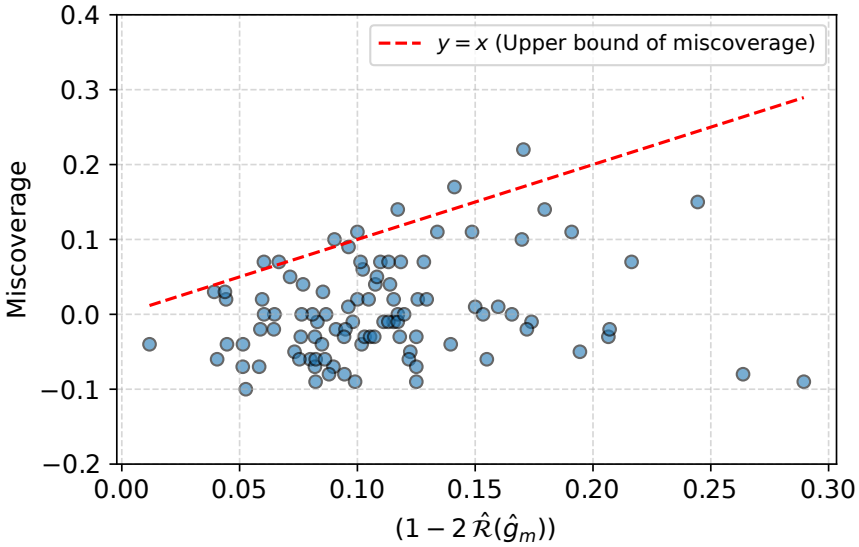

Figure 7: Miscoverage vs estimated bound. Points below $y = x$ confirm the bound's conservativeness.

E.2 Another real dataset

Bike Sharing Dataset from the UCI repository(Fanaee-T, 2013) The target value is the total count of bikes rented during each hour. There are four continuous covariates and 12 discrete covariates. 5000 training data and 100 calibration data are selected. Another 100 test data for each pattern of missingness are generated under MCAR mechanism with 50% missing values for each feature. Since we have four continuous covariates, 15 $(2^4 - 1)$ types of missingness patterns exist.

Table 4: Comparison on Bike sharing dataset for different methods. The average coverage and interval width of 15 different missingness patterns are listed.

Method	Avg. Cov	Avg.Width
MDA-Exact	92.9 ± 0.7	∞
MDA-Nested	91.8 ± 0.5	336.3 ± 2.3
Weighted CP	91.9 ± 0.6	$\mathbf{306.4 \pm 2.2}$
ARC CP	93.0 ± 0.6	$\mathbf{309.9 \pm 3.1}$

The results are shown in Table 4, we can observe similar results: MDA-Exact may not work for certain masks, thus leading the average width of the interval to be infinite, and MDA-Nested over-estimates the uncertainty, whereas our methods provide a prediction set with shorter width while being mask-conditional valid.

MEPS19 dataset The MEPS19 dataset refers to the 2019 Medical Expenditure Panel Survey conducted by the Agency for Healthcare Research and Quality (AHRQ)[6]. It is a large-scale, nationally representative survey of families and individuals in the United States, collecting detailed information on healthcare utilization, expenditures, insurance coverage, and health status. The dataset contains both continuous and categorical covariates: in our setup, we use 5 continuous and 134 discrete covariates.

The target variable in our analysis, *UTILIZATION_reg*, is an aggregated measure of healthcare utilization, derived from MEPS variables describing the frequency and intensity of medical service use. Due to the wide range of the target values, applying the Weighted Conformal Prediction (Weighted CP) method is computationally

[6] `https://meps.ahrq.gov/data_stats/download_data_files.jsp`

challenging, as it requires evaluating the conformity score across a dense grid of possible outcomes. Therefore, we focus on comparing MDA-Exact, MDA-Nested, and our ARC-CP method.

We randomly select 1000 training samples and 500 calibration samples. The 5 continuous covariates are subject to missingness following a Missing Completely at Random (MCAR) mechanism. We repeat the experiment 50 times to estimate the average coverage and corresponding confidence intervals. The results, shown in Table 5, indicate that all methods achieve the desired coverage, while our ARC-CP method yields the shortest average interval length.

Table 5: Comparison on Bike sharing dataset for different methods. The worst-case coverage and interval width of 32 different missingness patterns are listed.

Method	Coverage	Interval width
MDA-Exact	94.03 – 95.01	33.0
MDA-Nested	94.29 – 95.15	34.1
ARC CP	94.04 – 95.04	**32.08**

Communities and Crime dataset To further explore high-dimensional behavior, we added experiments on the Communities and Crime dataset (Redmond and Baveja, 2002), a regression task with 122 real-valued features. We allow all features to be missing, such that there are 2^{122} possible masks, and randomly sample 1000 masks to evaluate mask-conditional coverage for each. For ARC-CP, the worst-case mask-conditional coverage among these 1000 masks is 87.8%; only 49 masks have coverage below 90% (between 87.8% and 90%), indicating that ARC-CP remains robust and reasonably close to the target even in this challenging high-dimensional real-data setting. MDA-Exact always gives infinite intervals, while Nested and Weighted CP cannot finish in a reasonable time (one week).

E.3 The necessity for correction

To show that our correction method can guarantee MCV effectively, we do experiments on the dataset in Section 4.4 with and without applying the correction methods. The comparison for whether using the correction method in terms of missingness coverage and interval width is shown in Figures 8 and 9. Without correction, the worst mask conditional coverage is 89.1%, showing that without a suitable weight/acceptance for adjusting original Calibration, the distribution shift results in an under-coverage of the true label.

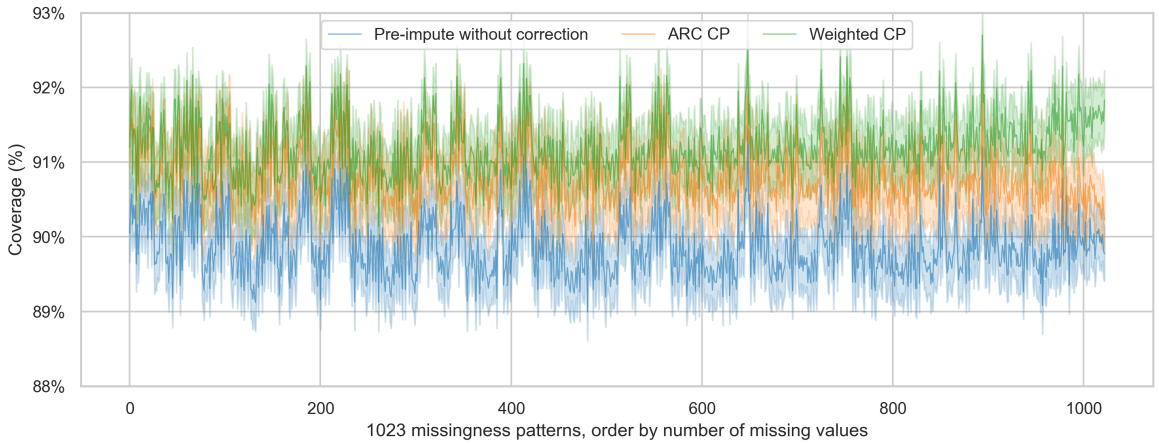

Figure 8: Synthetic dataset: Comparison of prediction coverage (mean and 99% confidence interval) across 1023 missingness patterns, with and without the proposed correction method.

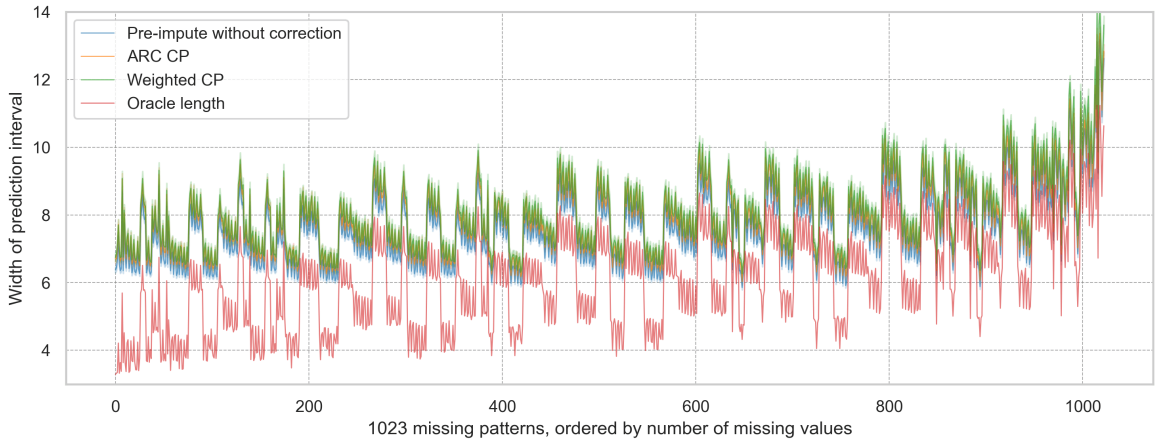

Figure 9: Synthetic dataset: Comparison of the width of prediction interval (mean and 99% confidence interval) across 1023 missingness patterns, with and without the proposed correction method.

E.4 Experiments of concrete dataset

We provide the missingness-wise coverage for concrete dataset, as shown in Figure 10.

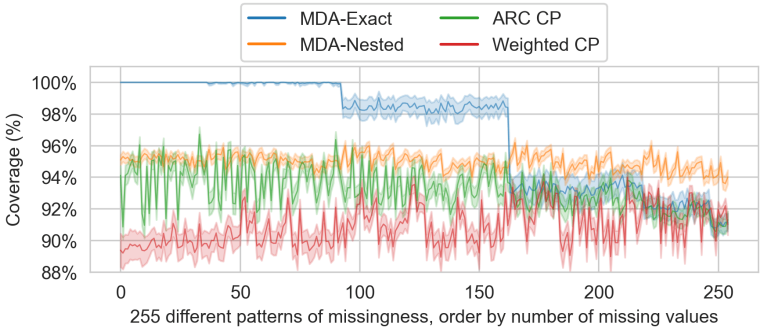

Figure 10: Concrete dataset: mean and 99% confidence interval of prediction coverage with respect to 255 missingness patterns

For MAR settings, the worst-case missingness-wise coverage and the marginal prediction interval width for the four methods are in Table 6:

Table 6: Worst-case coverage (95% CI) and average prediction interval width over 15 masks, concrete dataset with 4 possible MAR missing covariates

Mech.	MDA-Exact	MDA-Nested	Weighted CP	ARC CP
MAR	89.1% ± 1.64% / 29.1	90.8% ± 1.43% / 31.4	89.1% ± 1.91% / 29.8	90.6% ± 1.42% / 28.9

For self-masked MNAR settings, it is generally difficult to generate enough test points per pattern, we allow 2 covariates to be missing. The worst-case missingness-wise coverage and the marginal prediction interval width for the four methods are in Table 7:

Table 7: Worst-case coverage (95% CI) and average prediction interval width over 15 masks, concrete dataset with 4 possible missing covariates

Mech.	MDA-Exact	MDA-Nested	Weighted CP	ARC CP
MNAR	89.0% ± 1.78% / 30.1	88.7% ± 1.76% / 30.4	90.0% ± 1.67% / 30.8	89.0% ± 1.83% / 30.1

E.5 Complementary of synthetic experiments for other settings

E.5.1 Less missing values

We consider the same setting of the synthetic dataset, except we have 20% missing values instead of 50%. As we said, when there are many missing values in the dataset, and we want to achieve MCV for a test point with almost no missing values, the MDA-Exact could give an uninformative prediction interval due to the lack of sample points in the calibration dataset. Still, this method should work well when we have fewer missing values. The comparison of MDA-Exact, MDA-Nested, as well as our methods, is in Figures 11 and 12. We observe that MDA-Exact is less likely to give an uninformative prediction set, but both the MDA-Exact and Nested show a missingness conditional over-coverage when the test point has fewer missing values. Our method shows more stability in coverage for all patterns of missingness. The AR-corrected method reduces 10.7% the average width of prediction interval than MDA-Nested for all patterns of missingness.

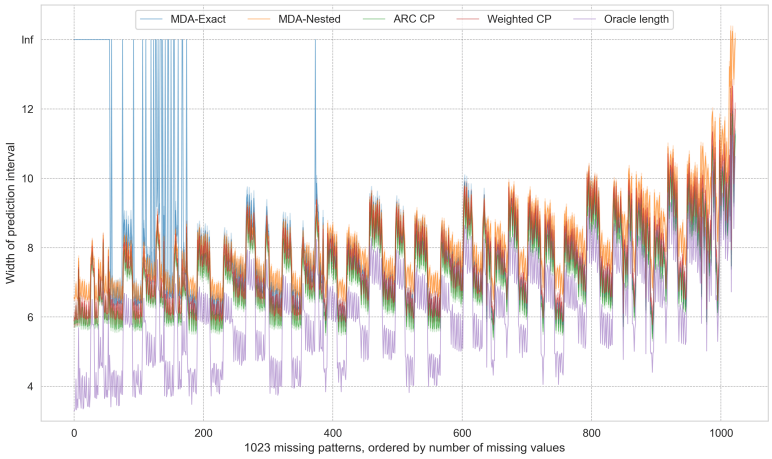

Figure 11: Synthetic dataset with 20% missing values: mean and 99% confidence interval of prediction interval width with respect to 1023 missingness patterns, comparison for whether using correction method.

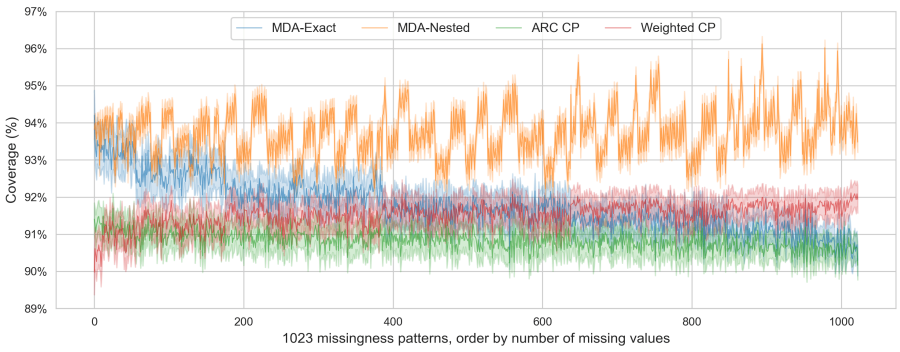

Figure 12: Synthetic dataset with 20% missing values: mean and 99% confidence interval of prediction coverage with respect to 1023 missingness patterns, comparison for whether using correction method.

E.5.2 Synthetic dataset with independent features $\rho = 0$

Similarly to Zaffran et al. (2024), we run another experiment on a dataset in which the features are independent ($\rho = 0$), leading to an imputation that can not be better than the marginal expectation of the features. We also keep 20% missing values. The same results are shown in Figures 13 and 14. We observe similar results: the AR-corrected method reduces 8.9% the average width of prediction interval than MDA-Nested for all patterns of missingness.

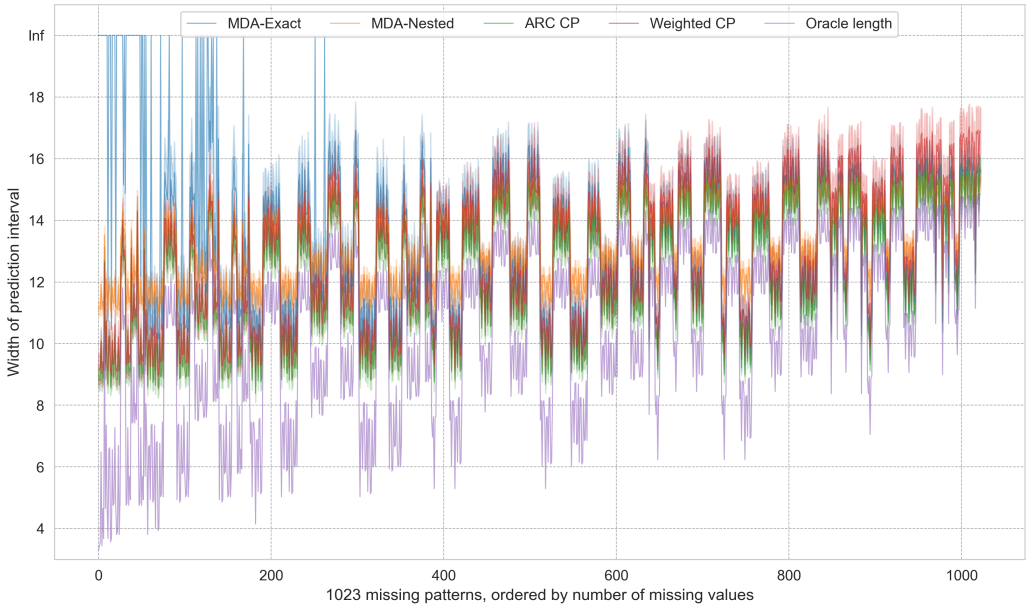

Figure 13: Synthetic dataset with 20% missing values and independent features: mean and 99% confidence interval of prediction interval width with respect to 1023 missingness patterns, comparison for whether using correction method.

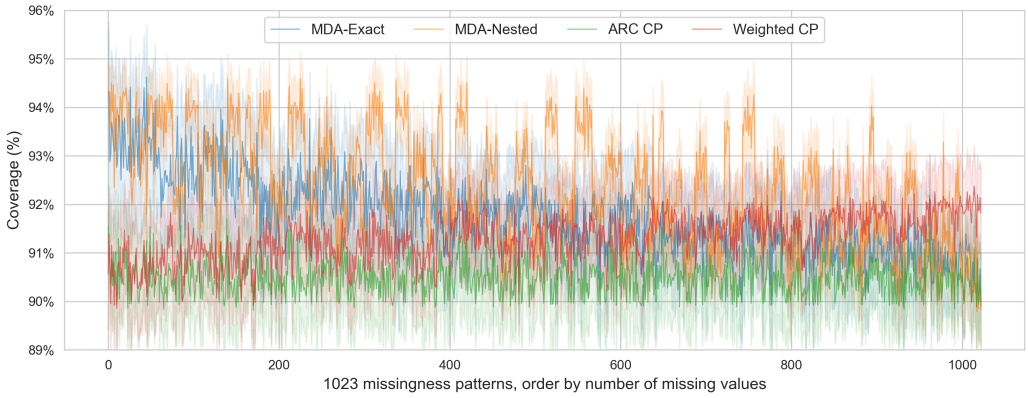

Figure 14: Synthetic dataset with 20% missing values and independent features: mean and 99% confidence interval of prediction coverage with respect to 1023 missingness patterns, comparison for whether using correction method.

E.6 Unsmoothed results for Figures 2 and 3

The unsmoothed results in the Experiments Section 4 are shown in Figures 15 and 16.

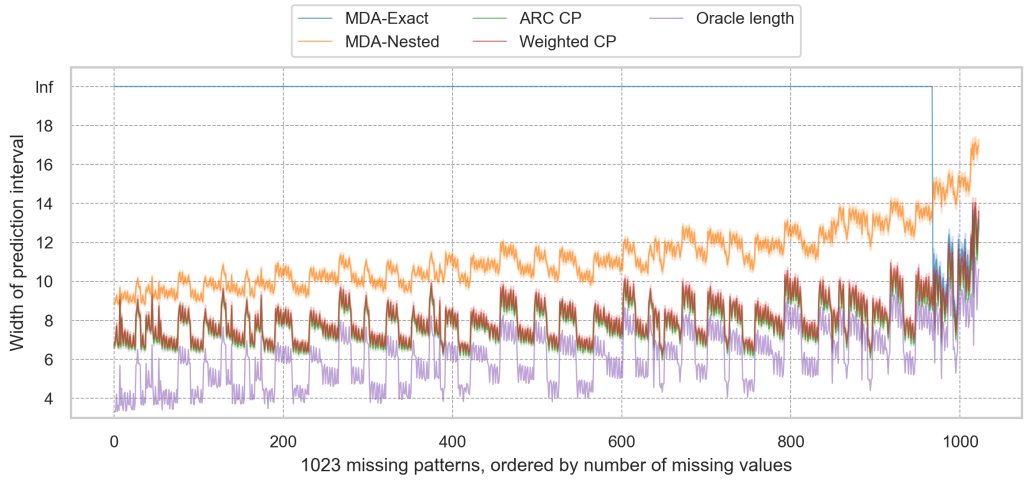

Figure 15: Synthetic dataset: unsmoothed results for mean and 99% confidence interval of prediction width with respect to 1023 missingness patterns.

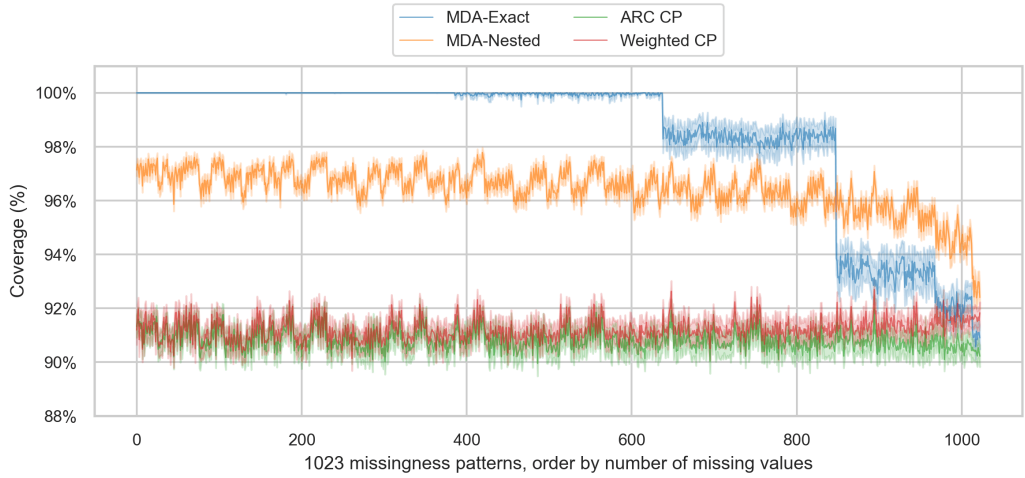

Figure 16: Synthetic dataset: unsmoothed results for mean and 99% confidence interval of prediction coverage with respect to 1023 missingness patterns.

E.7 Coefficient vector β for the 100-dimensional synthetic experiment

In the 100-dimensional synthetic experiment of Section 4.2, we fix the linear model coefficients as $\beta =($ 1, 2, -1, 3, -0.5, -1, 0.3, 1.7, 0.4, -0.3, -2.84, 0.91, 2.77, -0.68, 1.53, -1.94, 0.44, -2.21, 2.05, -0.17, -1.46, 0.23, 1.88, -2.63, 2.41, -0.82, 0.71, -1.12, 2.58, -0.49, -2.05, 1.37, -0.94, 2.63, -1.59, 0.52, -2.77, 1.96, -0.08, 2.24, -1.23, 0.35, 2.49, -2.36, 1.14, -0.71, 2.81, -1.76, 0.62, -2.48, 1.79, -0.27, 2.92, -1.65, 0.84, -2.58, 2.18, -0.41, 1.25, -2.93, 2.67, -0.56, 1.48, -2.12, 0.19, 2.32, -1.88, 0.77, -2.69, 2.55, -0.63, 1.61, -2.44, 0.58, 2.86, -1.31, 0.05, -2.29, 1.92, -0.14, 2.37, -1.57, 0.49, 2.99, -2.72, 1.08, -0.35, 2.14, -1.99, 0.66, -2.11, 0.88, -1.45, 2.63, -0.72, 1.57, -2.33, 0.12, 2.48, -1.08), as shown in the released code.