Predictive Uncertainty Quantification with Missing Covariates

Margaux Zaffran, Julie Josse, Yaniv Romano, Aymeric Dieuleveut

Predictive Uncertainty Quantification with Missing Covariates

Margaux Zaffran[*,1,2], Julie Josse[1], Yaniv Romano[3], and Aymeric Dieuleveut[2]

[1]PreMeDICaL project team, INRIA Sophia-Antipolis, Montpellier, France
[2]CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, Palaiseau, France
[3]Departments of Electrical Engineering and of Computer Science, Technion - Israel Institute of Technology, Haifa, Israel

Abstract

Predictive uncertainty quantification is crucial in decision-making problems. We investigate how to adequately quantify predictive uncertainty with missing covariates. A bottleneck is that missing values induce heteroskedasticity on the response's predictive distribution given the observed covariates. Thus, we focus on building predictive sets for the response that are valid *conditionally* to the missing values pattern. We show that this goal is impossible to achieve informatively in a distribution-free fashion, and we propose useful restrictions on the distribution class. Motivated by these hardness results, we characterize how missing values and predictive uncertainty intertwine. Particularly, we rigorously formalize the idea that the more missing values, the higher the predictive uncertainty. Then, we introduce a generalized framework, coined `CP-MDA-Nested`*, outputting predictive sets in both regression and classification. Under independence between the missing value pattern and both the features and the response (an assumption justified by our hardness results), these predictive sets are valid conditionally to any pattern of missing values. Moreover, it provides great flexibility in the trade-off between *statistical variability* and *efficiency*. Finally, we experimentally assess the performances of `CP-MDA-Nested`* beyond its scope of theoretical validity, demonstrating promising outcomes in more challenging configurations than independence.

Keywords: predictive uncertainty quantification, missing values, conformal prediction, distribution-free inference

1 Introduction

Predictive uncertainty quantification. Over the last decades, major research efforts on statistical and machine learning algorithms have enabled them to leverage large data sets. They are now used to support high-stakes decision-making problems such as medical, energy, or civic applications, to name just a few. To ensure the safe deployment of these models and their adoption by society, it is crucial to acknowledge that these point predictions remain uncertain, and to quantify this uncertainty, communicating the limits of predictive performance. Therefore, uncertainty quantification has received much attention in recent years, particularly in the form of building prediction sets.

Formally, the aim is to build a predictive set for the response $Y \in \mathcal{Y}$, after observing the random vector $X \in \mathcal{X} \subseteq \mathbb{R}^d$ which contains $d \in \mathbb{N}^*$ explanatory variables. Given a *miscoverage level* $\alpha \in [0, 1]$, a *marginally valid* predictive set $\mathcal{C}_\alpha(\cdot)$ is a function satisfying

[*]Corresponding author: `margaux.zaffran@inria.fr`

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(X)) \geq 1 - \alpha. \tag{1}$$

The goal is that $\mathcal{C}_\alpha(\cdot)$ is as small as possible while being marginally valid. Distribution-free uncertainty quantification tools are powerful as they require minimal assumptions on the data generation process—typically only access to a sequence of n exchangeable data points—making them usable on a wide range of applications, unlike traditional probabilistic approaches.

Importantly, it has to be noted that Equation (1) averages among all probable (X, Y), and thus might over-cover easy data points (say, e.g., young patients) at the cost of under-covering harder data points (say, e.g., older patients). Therefore, one branch of the literature studied how Equation (1) could be turned into a stronger goal. Specifically, Vovk (2012); Lei and Wasserman (2014); Barber et al. (2021a) emphasize trade-offs between theory and practice. They investigate the implications of designing a practical distribution-free method, that is one which outputs sets $\mathcal{C}_\alpha(\cdot)$ such that

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(x) | X = x) \geq 1 - \alpha, \text{ for any } x \in \mathcal{X}. \tag{2}$$

Unfortunately, they showed that Equation (2) is impossible to achieve in an informative way (i.e., typically $\mathcal{C}_\alpha(\cdot) \equiv \mathcal{Y}$ with high probability) if no assumptions on the data distributions are made. Moreover, finding natural relaxations that are compatible with informative distribution-free predictive sets seems also hard: restrictions to conditioning on $x \in \mathcal{X}$, for an arbitrary mass positive $\mathcal{X} \subseteq \mathcal{X}$, is still hard to achieve informatively (Barber et al., 2021a).

Missing values. Somewhat paradoxically, as the quantity of data rises, the number of missing data also increases. This phenomenon is modeled through the introduction of a third random variable called the *mask* or *missing pattern*, denoted by $M \in \mathcal{M} \subseteq \{0, 1\}^d$, encoding which variables have not been observed. That is, the mask M is the indicator vector such that for any $j \in [\![1, d]\!]$, $M_j = 1$ whenever X_j is missing (not observed), and $M_j = 0$ otherwise. As a consequence, we are working on $\mathcal{P} :=$ {distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$}. For a given pattern $m \in \mathcal{M}$, $X_{\text{obs}(m)}$ is the random vector of observed features, and $X_{\text{mis}(m)}$ is the random vector of unobserved ones. For example, if we observe $(\texttt{NA}, 6, 2)$ then $m = (1, 0, 0)$ and $x_{\text{obs}(m)} = (6, 2)$. Notice that the number of different missing patterns, i.e., the size or cardinality of $\mathcal{M} := \#\mathcal{M}$, typically grows exponentially in the dimension (for $\mathcal{M} = \{0, 1\}^d$ there are 2^d different patterns).

The way we deal with those missing values will typically depend on the downstream task at hand. While there is a vast range of studies in the inferential setting (Little, 2019; Josse and Reiter, 2018) with numerous implementations (Mayer et al., 2022), the research effort is scarcer on the prediction framework (Josse et al., 2024; Le Morvan et al., 2020b,a, 2021; Ayme et al., 2022; Van Ness et al., 2022; Ayme et al., 2023; Zaffran et al., 2023; Ayme et al., 2024). It is mostly limited to *point prediction*, except for Zaffran et al. (2023). The literature on both inference and prediction highlights the necessity of taking into account the missingness distribution. Following Rubin (1976), we consider three well-known missingness mechanisms.

Definition 1.1 (Missing Completely At Random (MCAR))**.** The missing pattern distribution is said to be Missing Completely At Random (MCAR) if $M \perp\!\!\!\perp X$. We denote $\mathcal{P}_{\text{MCAR}}$ the corresponding set of distributions, i.e. $\mathcal{P}_{\text{MCAR}} := \{P \in \mathcal{P}$, such that for any $m \in \mathcal{M}, \mathbb{P}_P(M = m|X) = \mathbb{P}_P(M = m)$, that is $M \perp\!\!\!\perp X\}$.

Definition 1.2 (Missing At Random (MAR))**.** The missing pattern distribution is said to be Missing At Random (MAR) if M only depends on the observed components of X. We denote $\mathcal{P}_{\text{MAR}}$ the corresponding set of distributions, i.e. $\mathcal{P}_{\text{MAR}} := \{P \in \mathcal{P}$, such that for any $m \in \mathcal{M}$, $\mathbb{P}_P(M = m|X) = \mathbb{P}_P\left(M = m|X_{\text{obs}(m)}\right)\}$.

Definition 1.3 (Missing Non At Random (MNAR))**.** The missing pattern distribution is said to be Missing Non At Random (MNAR) if M can depend on the observed values of X but also on its missing components. We denote $\mathcal{P}_{\text{MNAR}}$ the corresponding set of distributions, i.e. $\mathcal{P}_{\text{MNAR}} := \mathcal{P}$.

Remark 1.4. We thus have $\mathcal{P}_{\text{MCAR}} \subset \mathcal{P}_{\text{MAR}} \subset \mathcal{P}_{\text{MNAR}} = \mathcal{P}$.

Predictive framework with missing covariates. In a predictive framework, the dependence between Y and M plays a key role, maybe even bigger than the relationship between (X, M). Indeed, in some situations, Y can be a direct function of M: the missingness conveys in itself information about the label. Therefore, these cases are particularly challenging in a predictive framework, as some patterns on the one hand can induce an important label distributional shift, and on the other hand be rarely observed due to the high cardinality of $\mathcal{M}$. Thus, we focus on settings where there is *not* such a direct dependency, that is Assumption A1. Yet, as we will show in the paper, it remains that the lack of observation of some features influences the uncertainty of the prediction of Y from $X_{\text{obs}(M)}$.

Assumption A1 (M does not explain Y)**.** We say that Y is independent of M given X if $Y \perp\!\!\!\perp M \,|X$. The associated distribution belongs to $\mathcal{P}_{\text{Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$.

Definitions 1.1 to 1.3 and Assumption A1 will be our main assumptions on the joint distribution of (X, M, Y) throughout the manuscript. Our interest is in building predictive sets from n observations $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ on a new test point $\left(X^{(n+1)}, M^{(n+1)}, Y^{(n+1)}\right)$. We thus also make assumptions on the *links between those samples*: the usual backbone assumption is that we have access to $n+1$ independent and identically distributed (i.i.d.) draws from a distribution Q in a set $\mathcal{Q}$, with $\mathcal{Q}$ being typically one of $\mathcal{P}_{\text{MCAR}}$, $\mathcal{P}_{\text{MAR}}$, $\mathcal{P}$, etc. The data distribution thus belongs to $\left\{Q^{\otimes(n+1)}, Q \in \mathcal{Q}\right\}$, which we denote $\mathcal{Q}^{\otimes(n+1)}$. Furthermore, we also consider here a relaxation of i.i.d., namely *exchangeability*, which is often sufficient to obtain guarantees in distribution-free predictive inference.

Assumption A2 (exchangeability)**.** The random variables $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are exchangeable, i.e., their distribution does not change when we permute them. We denote $\mathcal{Q}^{\text{exch}(n+1)} = \left\{Q^{\text{exch}(n+1)}, Q \in \mathcal{Q}\right\}$ the set of distributions of exchangeable random variables, with marginal distribution in $\mathcal{Q}$.

An i.i.d. sequence is a fortiori exchangeable, while the reverse is not true (for example, sampling without replacement leads to a sequence that is exchangeable but not i.i.d.).

Remark 1.5. We thus have that for any $\mathcal{Q}$, $\mathcal{Q}^{\otimes(n+1)} \subset \mathcal{Q}^{\text{exch}(n+1)}$.

Predictive uncertainty quantification under missing covariates. When features are missing, Equation (1) extends with $\mathcal{C}_\alpha$ a function of two arguments: X and M. Specifically, $\mathcal{C}_\alpha$ is a *marginally valid* predictive set for the test response Y given its corresponding covariates X and the mask M if:

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(X, M)) \geq 1 - \alpha. \tag{MV}$$

However, marginal validity (MV) is not enough from an equity stand point and might result in discriminating between observations depending on their missing pattern (Zaffran et al., 2023). Indeed, missing values create heteroskedasticity in the resulting distribution of Y given $X_{\text{obs}(M)}$. Therefore, they argue that when facing missing values one should aim at *mask-conditional-validity* (MCV) even in the MCAR setting, i.e.:

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(X, M)|M) \geq 1 - \alpha. \tag{MCV}$$

Equation (MCV) is similar in spirit and motivation than Equation (2) but on a discrete space. Hence the impossibility results on X-conditional coverage do not hold anymore. However, (MCV) is a challenging goal as it requires the coverage to be controlled on *any* mask $m \in \mathcal{M}$, even those rarely observed at train time.

In the sequel, to highlight the underlying dependencies and randomness, any estimator of $\mathcal{C}_\alpha(\cdot, \cdot)$ fitted on a data set $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^n$ is denoted as $\widehat{C}_{n,\alpha}(\cdot, \cdot)$. We call a *method* a function that, for any $\alpha \in [0, 1]$, takes as input $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^n$ and outputs an estimator $\widehat{C}_{n,\alpha}(\cdot, \cdot)$. Table 1 reminds the notations.

1.1 Literature's background

Very recent papers have investigated uncertainty quantification with missing values. Both Gui et al. (2023) and Shao and Zhang (2023) consider the question of distribution-free uncertainty quantification for matrix completion tasks. While the former considers building predictive sets for all of the missing entries, the latter focuses on what they call *matrix prediction* where predictive sets are required only for the last "individual" of the data set. Seedat et al. (2023) addresses the related but

Name	Definition	Comment
$\#A$	Cardinal of the set A	
$\mathscr{P}(A)$	Power set of A	
d	Number of features	
$\mathcal{X}$	Features space	$\mathcal{X} \subseteq \mathbb{R}^d$
$\mathcal{Y}$	Label space	
$\mathcal{M}$	Missing values pattern space	$\mathcal{M} \subseteq \{0, 1\}^d$
NA	Not Available (or missing value)	
$\text{obs}(m)$	Indices of the observed components for mask $m \in \mathcal{M}$ (there are $\lvert\text{obs}(m)\rvert := \sum_{i=1}^d m_i$ of them)	$\text{obs}(m) \in \mathbb{N}^{\lvert\text{obs}(m)\rvert}$
$\text{mis}(m)$	Indices of the missing components for mask $m \in \mathcal{M}$ (there are $\lvert\text{mis}(m)\rvert := d - \sum_{i=1}^d m_i$ of them)	$\text{mis}(m) \in \mathbb{N}^{\lvert\text{mis}(m)\rvert}$
$\mathcal{P}$	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$	
$\mathcal{P}_{\text{MAR}}$	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$ such that X is Missing At Random	
$\mathcal{P}_{\text{MCAR}}$	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$ such that X is Missing Completely At Random	
$\mathcal{P}_{Y \perp\!\!\!\perp M \mid X}$	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$ such that $Y \perp\!\!\!\perp M \mid X$	
n	Number of training observations	$n + 1$ is the test index
$P^{\otimes(n+1)}$	Product distribution of P with itself $n + 1$ times (i.e., distribution of $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ drawn i.i.d. with marginal P)	$P \in \mathcal{P}$
$\mathcal{Q}^{\otimes(n+1)}$	$\left\{Q^{\otimes(n+1)}, Q \in \mathcal{Q}\right\}$	$\mathcal{Q} \subseteq \mathcal{P}$
$P^{\text{exch}(n+1)}$	Exchangeable distribution of $n + 1$ random variables of distribution P	$P \in \mathcal{P}$
$\mathcal{Q}^{\text{exch}(n+1)}$	$\left\{Q^{\text{exch}(n+1)}, Q \in \mathcal{Q}\right\}$	$\mathcal{Q} \subseteq \mathcal{P}$
α	Miscoverage rate	$\alpha \in [0, 1]$
$\mathcal{C}_\alpha(\cdot, \cdot)$	Predictive set function aiming at $1 - \alpha$ coverage	$\mathcal{C}_\alpha : \mathcal{X} \times \mathcal{M} \longrightarrow \mathscr{P}(\mathcal{Y})$
$\widehat{C}_{n,\alpha}(\cdot, \cdot)$	Estimator for $\mathcal{C}_\alpha(\cdot, \cdot)$ based on $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^n$, through a *method*	
MV	Marginal validity	
MCV	Mask-conditional-validity	

Table 1: Summary of notations.

distinct problem of missing values in the responses, which is generally known as the semi-supervised setting. They introduce a self-supervised learning approach for incorporating unlabeled training data into the conformalization process. In the same framework, Lee et al. (2024) leverages tools from the causal inference literature to achieve stronger guarantees such as feature and outcome's missingness conditional coverage, which are, in spirit, close to our focus (yet in a different framework).

Closer to our work of predictive uncertainty quantification under missing covariates is Zaffran et al. (2023), as they focus on the same setting (i.e., to predict Y given X, where X might suffer from missing values both at train time and test time). After showing that *impute-then-predict+conformalization* is marginally valid (MV) for any missing mechanism and imputation, they introduce the harder goal of *mask-conditional-validity* (MCV), motivated by an illustration on the heteroskedasticity generated by the missing values on a Gaussian Linear Model. They present a novel methodology, *Missing Data Augmentation* (MDA), which combines with conformal prediction (CP, Vovk et al., 2005) in order to produce MCV sets. CP-MDA includes two algorithms, `CP-MDA-Exact` and `CP-MDA-Nested`, the former requiring a strict subsampling step on the training set, while the latter allows to keep the whole training data, which in turns induce large predictive sets. Zaffran et al. (2023) provide theoretical guarantees on the MCV of `CP-MDA-Exact` and on a technical minor modification of `CP-MDA-Nested`, under MCAR and $Y \perp\!\!\!\perp M \,|X$ assumptions.

1.2 Overview of our contributions (and outline)

In short, our objective is to tackle the following question: **when and how is it possible to achieve MCV?** Notably, we are interested in understanding $i)$ what assumptions are necessary to ensure MCV, $ii)$ how to design a tailored methodology, and $iii)$ what happens when these assumptions are not satisfied.

We start by proving hardness results on distribution-free MCV in Section 2. Notably, for a MCV method outputting $\widehat{C}_{n,\alpha}(\cdot,\cdot)$ with no assumption except from having access to n i.i.d. draws, we prove that the predictive interval is most often uninformative: for any $m \in \mathcal{M}$ the probability that, say, $\widehat{C}_{n,\alpha}(\cdot, m) \equiv \mathcal{Y}$ is higher than $1 - \alpha - \Delta_{m,n}$, where $\Delta_{m,n}$ gets negligible when the mask m is nearly not observed in a sample of size n. In other words, a method that is distribution-free MCV will output uninformative intervals on any mask that does not represent a high enough proportion of the training data. We go further and show that the exact same trade-off still holds for a method that is MCV only on distributions that are MAR, or MCAR, or similarly on distributions such that $Y \perp\!\!\!\perp M \,|X$, i.e., restricting an algorithm to be MCV only when $Y \perp\!\!\!\perp M \,|X$ would still output uninformative sets on rarely observed masks: it is necessary to add another assumption on the dependence between X and M (such as MCAR) to allow for informative MCV on any mask. Importantly, this theoretical analysis brings new insights on the achievability of X-group-conditional validity, beyond MCV[1].

This motivates the investigation of the quantile regression and missing values interplay in Section 3, so as to provide guidelines for practical design of probabilistic prediction with missing covariates. This interplay is hard to characterize in general but becomes explicit under missingness assumptions', or a multivariate Gaussian setting or linear model. Our key findings are (i—Section 3.1) that the uncertainty often increases with more missing values: we analyze different mathematical statements of this main idea (in terms of conditional variance, inter-quantile distance, or predictive interval length) and evaluate theoretically under which distributional assumptions they are satisfied,

[1]Precisely, we provide a rigorous quantification of Vladmir Vovk's comment on X-conditional validity: "of course, the condition that x be a non-atom is essential: if $P_X(x) > 0$, an inductive conformal predictor that ignores all examples with objects different from x will have $1 - \alpha$ object conditional validity and can give narrow predictions if the training set is big enough to contain many examples with x as their object" rephrased from Vovk (2012) to match our notations.

in particular under MCAR and $Y \perp\!\!\!\perp M \mid X$, motivating our methodological design of Section 4; (ii—Section 3.2) if the goal is to estimate quantiles, it is essential to be able to retrieve the mask to construct intervals, in contrast to classic mean regression where the mask is not as crucial.

In Section 4, we propose a unified framework, `CP-MDA-Nested`*, building predictive sets with missing covariates for both regression and classification tasks. Precisely, it bridges the gap between `CP-MDA-Exact` and `CP-MDA-Nested` introduced in Zaffran et al. (2023), by encapsulating these two algorithms as well as any in between with more flexible subsampling schemes, allowing to fix the trade-off between coverage variance (`CP-MDA-Exact`) and overly conservative predictive sets (`CP-MDA-Nested`). Leveraging the similarity between `CP-MDA-Nested`* and leave-one-out conformal approaches (Vovk, 2013; Barber et al., 2021b; Gupta et al., 2022) we provide theory on the marginal validity of `CP-MDA-Nested`* without subsampling, which holds regardless of the missingness distribution (without any assumption on the dependence between M and X, but also without any assumption on the relationship between M and Y conditionally on X). Moreover, we also establish that `CP-MDA-Nested`* is MCV for a wide range of subsampling schemes under MCAR and $Y \perp\!\!\!\perp M \mid X$.

Finally, in Section 5 we conduct synthetic experiments beyond the MCAR and $Y \perp\!\!\!\perp M \mid X$ assumptions. Precisely, we generate missingness that is either MAR (5 different settings), MNAR (11 different settings) or such that $Y \not\perp\!\!\!\perp M \mid X$. `CP-MDA-Nested`* empirically maintains MCV under MAR and MNAR missingness. When $Y \perp\!\!\!\perp M \mid X$ is not satisfied, `CP-MDA-Nested`* does not ensure MCV experimentally, unless the imputation is accurate enough. Overall, these numerical experiments showcase the robustness of `CP-MDA-Nested`* beyond its theoretical scope of validity.

In the following Table 2, we summarize and organize our main contributions. We report the theoretical results on the possibility to achieve informative MCV, either positive results (✓) or negative hardness results (✗), along with our more general result on marginal validity. Moreover, we locate experimental results by indicating the figures that align with particular setups. In particular, we distinguish two kinds of experiments: *Numerical extension* of results beyond the conditions where the theory is applicable, which demonstrates promising outcomes in more challenging configurations, and *Numerical confirmation* of results anticipated by theoretical analysis, that is the outcomes of the experiment either i) were already expected based on the theory or ii) confirm that the theoretical assumptions can not be relaxed to the corresponding distributional setting.

	$\mathcal{P}_{\text{MCAR}}$	$\mathcal{P}_{\text{MAR}}$	$\mathcal{P}_{\text{MNAR}} = \mathcal{P}$	
$\mathcal{P}_{Y \perp\!\!\!\perp M \mid X}$	`CP-MDA-Nested`*: ✓ *Theorem 4.3*	?	Hardness: ✗ *Proposition 2.8*	*Theory*
		Figures 5a and 5b	Figures 6a, 6b, 7a and 7b	*Num. extension*
	Figure 4		Remark 5.1	*Num. confirmation*
$\mathcal{P}$	Hardness: ✗ *Proposition 2.6*	Hardness: ✗ *Proposition 2.6*	Hardness: ✗ *Theorem 2.3* `CP-MDA-Nested`*: MV *Theorem 4.2*	*Theory*
	Figure 8a			*Num. extension*
	Figure 8b	Remark 5.1	Remark 5.1	*Num. confirmation*

Table 2: Summary of the main theoretical results.

2 When is Mask-Conditional-Validity (MCV) a too lofty goal?

We will show in this section that purely distribution-free MCV guarantees are often uninformative. As a consequence, we will have to impose some non-parametric assumption on the underlying data distribution. We thus have to define the concept of MCV with respect to a class of distributions $\mathcal{D}$ (MCV-$\mathcal{D}$), and to study the sets $\mathcal{D}$ that allow for informative MCV-$\mathcal{D}$.

Definition 2.1 (MCV-$\mathcal{D}$)**.** Let $\mathcal{D}$ be a set of distributions on $(\mathcal{X} \times \mathcal{M} \times \mathcal{Y})^{n+1}$. A method outputting $\widehat{C}_{n,\alpha}(\cdot, \cdot)$ based on $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ for any $\alpha \in [0, 1]$ is MCV-$\mathcal{D}$ if *for any distribution* $D \in \mathcal{D}$ and any $\alpha \in [0, 1]$, we have:

$$\mathbb{P}_D\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right) | M^{(n+1)}\right) \overset{a.s.}{\geq} 1 - \alpha,$$

i.e., for any $m \in \mathcal{M}$ such that $\mathbb{P}\left(M^{(n+1)} = m\right) > 0$, it holds:

$$\mathbb{P}_D\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m\right) | M^{(n+1)} = m\right) \geq 1 - \alpha.$$

If $\mathcal{D} = \mathcal{P}^{\text{exch}(n+1)}$ we recover the holy grail of being MCV for any exchangeable distribution, i.e., the most distribution-free result we could target. If $\mathcal{D}$ is not specified thereon, it will refer to MCV-$\mathcal{P}^{\text{exch}(n+1)}$. An easier goal is to aim at MCV-$\mathcal{P}^{\otimes(n+1)}$, that is MCV on i.i.d. distributions.

Remark 2.2. For any sets $\mathcal{D} \subseteq \mathcal{D}'$, a method that is MCV-$\mathcal{D}'$ is also MCV-$\mathcal{D}$, i.e., MCV-$\mathcal{D}' \Rightarrow$ MCV-$\mathcal{D}$.

A naive idea to ensure MCV is to split the data set into $\#\mathcal{M}$ sub data sets, one for each mask, and run any marginally valid method on each of the data sets independently. However, as $\#\mathcal{M}$ grows exponentially in the dimension, this is not practical as it will generate small (or even empty) data sets for some masks. In particular, as long as $\mathbb{P}(M = m)$ is low with respect to n for a given $m \in \mathcal{M}$, estimation on the sub data set is hard, and even finite sample method such as conformal prediction (Vovk et al., 2005) will suffer from important statistical variability or uninformativeness. Therefore, in practice, we usually need to go beyond this solution if we aim to achieve MCV for any mask, even those rarely observed at train times. Nevertheless, the task appears challenging without restricting the link between M and (X, Y), precisely due to the lack of information available in the data set. The question we tackle in this section is the following: **is it possible to achieve *distribution-free* MCV in an informative way for any mask in $\mathcal{M}$, even those occurring with low probability?**

Link with group conditional coverage. More generally, the question is that of finding on which subspace of the features it is possible to obtain meaningful conditional guarantees. Thus, the results demonstrated in this section give some answers to the broader question of when is *group-feature-conditional validity* achievable (a relaxation of Equation (2)), which has attracted considerable interest lately (see e.g., Romano et al., 2020; Barber et al., 2021a; Guan, 2022; Jung et al., 2023; Gibbs et al., 2023, to name just a few).

Our hardness results shed light on X-group-conditional coverage.

In the rest of this section, M can be interpreted as any additional random variable, that may (or may not) depend on X, on which we aim at achieving distribution-free conditional validity. For example, $\mathcal{M}$ could represent subgroups of $\mathcal{X}$, eventually overlapping. Specifically, assume $\mathcal{M} = \{0,1\}^{|\mathcal{G}|}$ for $\mathcal{G}$ a collection of groups on $\mathcal{X}$, then M is an indicator vector on whether X belongs to each group of $\mathcal{G}$ or not.
A particular case of this generalization is $\mathcal{G} = \left\{\{X \in \mathcal{X} : X_j \text{ is missing}\}_{j=1}^{d}\right\}$, recovering our missing covariates setting with M the missing pattern. While our discussion in this section is written towards the missing covariates setting, the interested reader might replace "missing pattern" or "mask" by "groups" whenever it makes sense[2], and the corresponding result will hold without further restriction or assumptions on the way the groups are designed.

2.1 Fully distribution-free result

Our first result, Theorem 2.3, confirms the previous intuition: any MCV-$\mathcal{P}^{\otimes(n+1)}$ method typically does output the whole set $\mathcal{Y}$ with high probability for any distribution, on low probability masks.

Theorem 2.3 (Trade-off set size and mask probability)**.** *Suppose that a method outputting $\widehat{C}_{n,\alpha}$ is MCV-$\mathcal{P}^{\otimes(n+1)}$. Then* for any $P \in \mathcal{P}$ *and any $m \in \mathcal{M}$ such that $P_M(m) > 0$, it holds:*

$$\begin{cases} \text{if } \mathcal{Y} \subseteq \mathbb{R} \text{ (regression)} : \mathbb{P}_{P^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(X,m)\right) = \infty\right) \geq 1-\alpha-\Delta_{m,n}, \\ \text{if } \mathcal{Y} \subseteq \mathbb{N} \text{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{P^{\otimes(n+1)}}\left(y \in \widehat{C}_{n,\alpha}(X,m)\right) \geq 1-\alpha-\Delta_{m,n}, \end{cases}$$

with $\Delta_{m,n} := \sqrt{2\left(1-\left(1-\frac{P_M(m)^2}{2}\right)^{n+1}\right)}$.

Since for any $x > 0$ and $n \in \mathbb{N}^*$, it holds $1-(1-x)^n < nx$, Theorem 2.3 implies that:

$$\begin{cases} \text{if } \mathcal{Y} \subseteq \mathbb{R} \text{ (regression)} : \mathbb{P}_{P^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(X,m)\right) = \infty\right) \geq 1-\alpha-P_M(m)\sqrt{(n+1)}, \\ \text{if } \mathcal{Y} \subseteq \mathbb{N} \text{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{P^{\otimes(n+1)}}\left(y \in \widehat{C}_{n,\alpha}(X,m)\right) \geq 1-\alpha-P_M(m)\sqrt{(n+1)}. \end{cases}$$

Theorem 2.3 provides a lower bound on the probability that the predictive set is uninformative for any $m \in \mathcal{M}$ (i.e., typically $\Lambda(\widehat{C}_{n,\alpha}(\cdot,m)) = \infty$ or $\#\widehat{C}_{n,\alpha}(\cdot,m) \geq \#\mathcal{Y}(1-\alpha)$).

Remark 2.4 (MCV-$\mathcal{P}^{\otimes(n+1)}$ implies uninformative sets even on simple distributions). Crucially, this lower bound holds for *any* distribution in $\mathcal{P}$. This implies that the hardness result applies also to smooth, nonpathological, distributions. Particularly, it means that any method that is fully distribution-free MCV (i.e., MCV-$\mathcal{P}^{\otimes(n+1)}$) will be subject to the lower bound even when applied to data whose actual distribution is as simple as possible (e.g., MCAR and $Y \perp\!\!\!\perp M \mid X$).

Remark 2.5 (Informative sets implies the method is not MCV-$\mathcal{P}^{\otimes(n+1)}$). Conversely, for a given method constructing predictive sets $\widehat{C}_{n,\alpha}$, assume that there exists a distribution $P \in \mathcal{P}$ and a mask m such that $P_M(m) > 0$ and $\Delta_{m,n} < \frac{1-\alpha}{2}$ and under which $\widehat{C}_{n,\alpha}$ is consistently of finite measure

[2]The only result that does not extend is Proposition 2.6 for $\mathcal{P}_{\text{MAR}}$, as by construction it relies on the missingness structure.

for different random draws from $P^{\otimes(n+1)}$. Then, this method is not MCV-$\mathcal{P}^{\otimes(n+1)}$, as otherwise under $P^{\otimes(n+1)}$ the predictive set would be of infinite measure with probability at least 0.25 for $\alpha \leq 0.5$ according to Theorem 2.3 (since $1 - \alpha - \Delta_{m,n} \geq \frac{1-\alpha}{2} \geq 0.25$).

Interpretation of the lower bound. Let us now decompose the lower bound. The first term, $1 - \alpha$, is an "irreducible term". Indeed, the estimator outputting $\mathcal{Y}$ with probability $1 - \alpha$ and the empty set $\emptyset$ with probability α (where the probability corresponds to an exogenous Bernoulli random variable) is valid conditionally on everything, thus a fortiori on M. Hence, the lower bound has to be smaller than $1 - \alpha$ as the set of MCV estimators includes this naive one.

For a given distribution P, the second term, $\Delta_{m,n}$, becomes negligible on any $m \in \mathcal{M}$ such that $P_M(m)$ is small with respect to n, making the lower bound be nearly $1 - \alpha$. This reflects the intuition that it is impossible to achieve informative conditional coverage when conditioning on events whose effective sample size is limited. In other words, the smaller the probability of the event occurring, the larger the training size must be to compensate and make "sure" that enough observations are drawn from that event.

Note that as $\mathcal{P}^{\otimes(n+1)} \subset \mathcal{P}^{\text{exch}(n+1)}$, any MCV-$\mathcal{P}^{\text{exch}(n+1)}$ estimator is MCV-$\mathcal{P}^{\otimes(n+1)}$ by Remark 2.2. Thus, the conclusion of Theorem 2.3 extends to any MCV-$\mathcal{P}^{\text{exch}(n+1)}$ estimator, on any $P^{\otimes(n+1)}$ with $P \in \mathcal{P}$.[3]

Proof sketch. For any given distribution $P \in \mathcal{P}$, and a given mask $m \in \mathcal{M}$ such that $P_M(m) > 0$, the idea of the proof is the following. Build another distribution $Q \in \mathcal{P}$, which equals P whenever $M \neq m$, and that "admits" an arbitrary spread on Y when $M = m$ (in short, Q is meant to be pathological yet close to P). By doing so, two statements can be made. First, $Q^{\otimes(n+1)}$ belongs to $\mathcal{P}^{\otimes(n+1)}$, therefore, as $\widehat{C}_{n,\alpha}$ is MCV-$\mathcal{P}^{\otimes(n+1)}$, under $Q^{\otimes(n+1)}$ the probability of $\widehat{C}_{n,\alpha}$ being uninformative is $1 - \alpha$ since Y can typically be anywhere. Second, as P and Q are the same everywhere except on $\{M = m\}$, the total variation distance between them is smaller than $P_M(m)$. This leads to the total variation distance between $P^{\otimes(n+1)}$ and $Q^{\otimes(n+1)}$ being smaller than $\Delta_{m,n}$. Combining these two observations, it finally leads to the probability of $\widehat{C}_{n,\alpha}$ being uninformative under $P^{\otimes(n+1)}$ which is greater than $1 - \alpha - \Delta_{m,n}$. The complete proof is given in Appendix A.2.

A familiar reader will note the similarity with the proofs given by Lei and Wasserman (2014); Vovk (2012). The difference is that, on the one hand, Vovk (2012) proof leverages an "reductio ad absurdum" that does not allow to explicitly build the set on which $P \neq Q$. On the other hand, Lei and Wasserman (2014) is constructive. Nonetheless, it relies on a crucial step that implicitly assumes that conditional-validity holds conditionally on the n data points, leading to an inexact statement: the lower bound obtained becomes 1. As we discussed, as well as Vovk (2012), the lower bound can not be bigger than $1 - \alpha$. We provide an alternate proof to this well-known X-conditional impossibility result that is constructive. Another improvement is that our expression of $\Delta_{m,n}$ comes from a tighter inequality than the ones used in Lei and Wasserman (2014) and Vovk (2012). Indeed, for the original impossibility result, the lower bound does not really matter as we then take its limit when the ball around x_0 shrinks, which is 0. But in our case, this ball is fixed to the event $\{M = m\}$.

2.2 Restricting the class of admissible missingness distributions

Interestingly, the proof of Theorem 2.3 adapts to MCV-$\mathcal{P}_{\text{MAR}}^{\otimes(n+1)}$ or MCV- $\mathcal{P}_{\text{MCAR}}^{\otimes(n+1)}$.

[3]The same is true for the subsequent Proposition 2.6 and Proposition 2.8.

Proposition 2.6 (Trade-off set size and mask probability on $\mathcal{P}_{\mathrm{MAR}}$ or $\mathcal{P}_{\mathrm{MCAR}}$)**.** *Let $\mathcal{Q}$ be either $\mathcal{P}_{\mathrm{MAR}}$ or $\mathcal{P}_{\mathrm{MCAR}}$. Suppose than an estimator $\widehat{C}_{n,\alpha}$ is MCV-$\mathcal{Q}^{\otimes(n+1)}$ at the level α. Then* for any $Q \in \mathcal{Q}$ *and any $m \in \mathcal{M}$ such that $Q_M(m) > 0$, it holds:*

$$\begin{cases} \textit{if } \mathcal{Y} \subseteq \mathbb{R} \textit{ (regression)} : \mathbb{P}_{Q^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(X,m)\right) = \infty\right) \geq 1 - \alpha - \Delta_{m,n}, \\ \textit{if } \mathcal{Y} \subseteq \mathbb{N} \textit{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{Q^{\otimes(n+1)}}\left(y \in \widehat{C}_{n,\alpha}(X,m)\right) \geq 1 - \alpha - \Delta_{m,n}, \end{cases}$$

with $\Delta_{m,n}$ given in Theorem 2.3.

Remark 2.7 (no direct implication between results). Proposition 2.6 for $\mathcal{Q} = \mathcal{P}_{\mathrm{MAR}}$ does not imply Proposition 2.6 for $\mathcal{Q} = \mathcal{P}_{\mathrm{MCAR}}$, nor the contrary. Indeed, on the one hand, as $\mathcal{P}_{\mathrm{MCAR}}^{\otimes(n+1)} \subseteq \mathcal{P}_{\mathrm{MAR}}^{\otimes(n+1)}$, any method that is MCV-$\mathcal{P}_{\mathrm{MAR}}^{\otimes(n+1)}$ is MCV-$\mathcal{P}_{\mathrm{MCAR}}^{\otimes(n+1)}$ (Remark 2.2). However, on the other hand, Proposition 2.6 (or Theorem 2.3) provides a uniform statement over $Q \in \mathcal{Q}$ (Remark 2.4): as $\mathcal{P}_{\mathrm{MCAR}}^{\otimes(n+1)} \subseteq \mathcal{P}_{\mathrm{MAR}}^{\otimes(n+1)}$, the final statement holds on more distributions for $\mathcal{Q} = \mathcal{P}_{\mathrm{MAR}}$ than for $\mathcal{Q} = \mathcal{P}_{\mathrm{MCAR}}$. Thereofore, Proposition 2.6 for $\mathcal{Q} = \mathcal{P}_{\mathrm{MAR}}$ provides a *stronger statement* over *fewer methods* than Proposition 2.6 for $\mathcal{Q} = \mathcal{P}_{\mathrm{MCAR}}$.

For the same reason, Proposition 2.6 is not deduced directly from Theorem 2.3, but from a careful consideration of the construction in its proof: the adversarial distribution built therein does not make any assumption on the relationship between X and M, which can be as simple as desired.

In fact, the key point for the proof of Theorem 2.3 is that the algorithm achieves MCV also on distributions under which Y and M can be dependent even conditionally on X: thus, it allows us to construct an adversarial distribution under which Y is equally likely to be anywhere on the label space for a given $m \in \mathcal{M}$.

In view of this, one could think that in order to break Theorem 2.3, and therefore to ensure that MCV is achievable in an informative way even on low probability masks, we have to *at least* assume $Y \perp\!\!\!\perp M \mid X$ (A1). However, in Proposition 2.8, we show that even estimators that are only MCV-$\mathcal{P}_{\mathrm{Y}\perp\!\!\!\perp\mathrm{M}\mid\mathrm{X}}^{\otimes(n+1)}$ suffer from the same trade-off on efficiency.

Proposition 2.8 (Trade-off set size and mask probability on $\mathcal{P}_{\mathrm{Y}\perp\!\!\!\perp\mathrm{M}\mid\mathrm{X}}$)**.** *Suppose that an estimator $\widehat{C}_{n,\alpha}$ is MCV-$\mathcal{P}_{\mathrm{Y}\perp\!\!\!\perp\mathrm{M}\mid\mathrm{X}}^{\otimes(n+1)}$ at the level α. Then for any $P \in \mathcal{P}_{\mathrm{Y}\perp\!\!\!\perp\mathrm{M}\mid\mathrm{X}}$ and for any $m \in \mathcal{M}$ such that $\frac{1}{\sqrt{2}} \geq P_M(m) > 0$, it holds:*

$$\begin{cases} \textit{if } \mathcal{Y} \subseteq \mathbb{R} \textit{ (regression)} : \mathbb{P}_{P^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(X,m)\right) = \infty\right) \geq 1 - \alpha - \Delta_{m,n}, \\ \textit{if } \mathcal{Y} \subseteq \mathbb{N} \textit{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{P^{\otimes(n+1)}}\left(y \in \widehat{C}_{n,\alpha}(X,m)\right) \geq 1 - \alpha - \Delta_{m,n}, \end{cases}$$

with $\Delta_{m,n} := \sqrt{2\left(1 - (1 - 2P_M(m)^2)^{n+1}\right)}$.

All in all, Proposition 2.8 demonstrates that even the simplest relationship between Y and M does not allow informative predictive sets. This reveals that to ensure that it is possible to obtain informative sets even on low probability masks (or events), one has to design a method that will be conditionally valid *only* on distributions with a constrained structure of dependence between Y and M given X, but also between M and X. In particular, trying to ensure MCV-$\mathcal{P}_{\mathrm{MCAR},\mathrm{Y}\perp\!\!\!\perp\mathrm{M}\mid\mathrm{X}}^{\otimes(n+1)}$ (where $\mathcal{P}_{\mathrm{MCAR},\mathrm{Y}\perp\!\!\!\perp\mathrm{M}\mid\mathrm{X}}^{\otimes(n+1)} := \mathcal{P}_{\mathrm{MCAR}}^{\otimes(n+1)} \cap \mathcal{P}_{\mathrm{Y}\perp\!\!\!\perp\mathrm{M}\mid\mathrm{X}}^{\otimes(n+1)}$) as done in Zaffran et al. (2023) appears as a natural way to approach the minimal set of assumptions.

Remark 2.9. In Figure 4, we illustrate that, on a distribution $P \in \mathcal{P}^{\otimes(n+1)}_{\text{MCAR},Y \perp\!\!\!\perp M \mid X}$, a provably MCV-$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR},Y \perp\!\!\!\perp M \mid X}$ method (introduced in Section 4) consistently outputs finite length predictive intervals (regression case). Therefore, we can conclude that obtaining a hardness result on $\mathcal{P}^{\otimes(n+1)}_{\text{MCAR},Y \perp\!\!\!\perp M \mid X}$ appears impossible, as such it would induce Remark 2.5 (with $\mathcal{P}^{\otimes(n+1)}_{\text{MCAR},Y \perp\!\!\!\perp M \mid X}$ instead of $\mathcal{P}^{\otimes(n+1)}$).

3 Link between missing covariates and predictive uncertainty

In light of the previous section, MCV appears hard to achieve. Thus, the problem that we aim to address now is to **find ways to model properly the missing covariates' influence on predictive uncertainty**. To understand the relationship between missing values and predictive uncertainty, this section explores simplified distributions on (X, M, Y)—such as MCAR and $Y \perp\!\!\!\perp M \mid X$—and/or on (X, Y)—such as linearity, Gaussianity. We consider the regression case with $\mathcal{Y} = \mathbb{R}$. This exploration aims to facilitate the development of suitable frameworks for probabilistic inference when covariates are missing—i.e., models that are as close as possible to achieving MCV.

3.1 Increasing uncertainty with nested masks

The hardness results of Section 2 induce that MCV cannot be (efficiently) achieved without structural assumptions on the links between the predictive distributions conditional on each missing pattern. In this subsection, we gain insights into the underlying reasons for this phenomenon: the predictive uncertainty depends on the missing pattern, a form of *heteroskedasticity*. In summary, we explore the following idea, which is a natural modelization attempt in that direction:

Idea: *The predictive uncertainty increases when less covariates are observed.*

In technical words, the aforementioned heteroskedasticity takes the form of an *isotonicity* (monotony) with respect to the mask, with the inclusion order given by Definition 3.1 below. In short: the more missing values, the more uncertainty there is.

Definition 3.1 (Included masks)**.** Let $(m, m') \in \mathcal{M}^2$, $m \subset m'$ if for any $j \in [\![1, d]\!]$ such that $m_j = 1$ then $m'_j = 1$, i.e., m' includes at least the same missing values than m.

Hereafter, we formally quantify such a statement, in particular in terms of conditional variance, inter-quantile distance, and predictive interval length. We demonstrate that some of those statements are valid, to different extent, under distributional assumptions, either generic or on specific model or examples. To that end, we introduce several properties, that can be considered as non-parametric assumptions on the underlying distributions. We put together some results of this section in the following Table 3, that can be used as a reading guide throughout the section.

Property \ Setup	Model 3.4	Model 3.3	$\mathcal{P}_{\text{MCAR},Y \perp\!\!\!\perp M \mid X}$
Variance	Var-1	~~Var-1~~ Var-2	Var-2
Inter-quantile	IQ-1	IQ-2	
Length of Oracle PI	Len-1	Len-2	Len-2

Table 3: Summary of the results from Section 3.1.

3.1.1 Conditional Variance Isotony w.r.t. the missing data patterns

We start by focusing on the link between M and the *conditional variance* of $Y|X_{\text{obs}(M)}$, that constitutes a natural proxy on the predictive uncertainty. Denote $V(X_{\text{obs}(M)}, M) := \text{Var}\left(Y|X_{\text{obs}(M)}, M\right)$ the conditional variance of Y given $\left(X_{\text{obs}(M)}, M\right)$. We introduce two properties regarding its ordering with respect to M: (Var-1) and (Var-2).

$$V(X_{\text{obs}(m)}, m) \overset{a.s.}{\leq} V(X_{\text{obs}(m')}, m') \qquad \text{for any } m \subset m', \qquad \text{(Var-1)}$$

$$\mathbb{E}\left[V(X_{\text{obs}(M)}, M)|M = m\right] \leq \mathbb{E}\left[V(X_{\text{obs}(M)}, M)|M = m'\right] \qquad \text{for any } m \subset m'. \qquad \text{(Var-2)}$$

Property Var-1 is stronger than Property Var-2 as it is an almost sure result w.r.t. the covariates X. The following proposition ensures that (Var-2) is satisfied under $\mathcal{P}_{\text{MCAR,Y⊥⊥M |X}}$ (that is, assumptions for which no hardness result can exist).

Proposition 3.2. *Under* $\mathcal{P}_{\text{MCAR,Y⊥⊥M |X}}$, (Var-2) *is valid.*

The proof of this result is given in Appendix B.1. This is a first significant result: under general assumptions—i.e., strong assumption on the relation between the mask and both the response and the features, but no assumptions on their distribution—, the averaged variance is always smaller on smaller masks. This establishes the existence of a link between the uncertainties *on patterns that can be compared*, that is patterns that are nested in one another. Note that the order given by Definition 3.1 is only a partial order: the average variance ordering is only enforced w.r.t. that partial order.

It is possible that the predictive uncertainty increases on average with the mask (Equation (Var-2)) but not almost surely on X (Equation (Var-1)), as illustrated by Model 3.3 below:

Model 3.3 (Unidimensional heteroskedasticity). Consider the following one-dimensional model:

- $X \sim \mathcal{N}(0, \sigma^2)$, $\sigma \in \mathbb{R}_+$;
- $\xi \sim \mathcal{N}(0, \tau^2)$, $\tau \in \mathbb{R}_+$, such that $\xi \perp\!\!\!\perp X$;
- $Y = \beta X + X\xi$, with $\beta \in \mathbb{R}$;
- $M \sim \mathcal{B}(\rho)$, with $\rho \in [0, 1]$, and $M \perp\!\!\!\perp (X, Y)$.

Under this model, we obtain that $M \perp\!\!\!\perp X$ (MCAR) and $Y \perp\!\!\!\perp M\ |X$, and

$$\begin{cases} \text{Var}(Y|X, M = 0) = \tau^2 X^2 \\ \text{Var}(Y|M = 1) = (\beta^2 + \tau^2)\sigma^2 \end{cases} \Rightarrow \begin{cases} \mathbb{E}\left[\text{Var}(Y|X, M = 0)\right] = \tau^2\sigma^2 \\ \mathbb{E}\left[\text{Var}(Y|M = 1)\right] = (\beta^2 + \tau^2)\sigma^2 \end{cases}.$$

Thus Equation (Var-2) is verified but Equation (Var-1) is only satisfied for X such that $X^2 \leq \left(1 + \frac{\beta^2}{\tau^2}\right)\sigma^2$. This is illustrated in Figure 1. The first subplot represents Y depending on X, while the third subplot displays $Y - \beta X$ depending on X, that is an illustration of the uncertainty of the distribution of $Y|X$. For any X outside the vertical dashed lines (corresponding to $\pm(1+\beta^2/\tau^2)\sigma^2$), the conditional variance of Y given X is larger than the overall variance when X is missing. Yet, the average variance of Y when X is missing is indeed higher than the average variance of Y when X is observed: this can be seen on the two histograms on subplots 2 and 4.

Finally, while Model 3.3 shows that (Var-1) is not always true, even under the assumptions of Proposition 3.2, we now show that it can be achieved in the following Gaussian linear model, a particular case of Gaussian pattern mixture model.

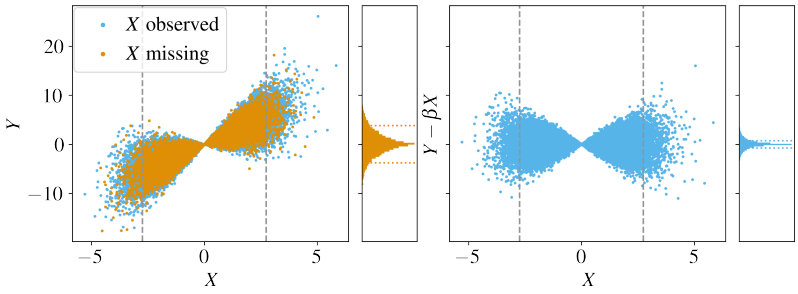


Figure 1: Visualisation of a random draw from the data distribution of Model 3.3, with 100000 i.i.d. samples, $\rho = 0.2$, $\sigma^2 = 1.5$, $\tau^2 = 1$ and $\beta = 2$. The colors indicate whether X is observed or missing. The first subplot represents Y depending on X, while the third subplot displays $Y - \beta X$ depending on X only for observed X, that is an illustration of the uncertainty of $Y|X$. The second subplot is an histogram of Y when X is missing, while the forth subplot is an histogram of $Y - \beta X$ when X is observed, i.e., they represent the predictive distribution of Y depending on whether X is observed or missing.

Model 3.4 (Gaussian linear model (GLM)). The data is generated according to a linear model and the covariates are Gaussian conditionally to the pattern:

- $Y = \beta^T X + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp\!\!\!\perp (X, M), \beta \in \mathbb{R}^d$.
- for all $m \in \mathcal{M}$, there exist $\mu^m \in \mathbb{R}^d$ and $\Sigma^m \in \mathbb{R}^{d \times d}$ such that $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$.

Such a model results in a MCAR distribution when $\Sigma^m \equiv \Sigma$. Indeed under Model 3.4 the resulting predictive distribution is given by $Y|(X_{\text{obs}(m)}, M = m) \sim \mathcal{N}\left(\tilde{\mu}^m, \tilde{\sigma}^m\right)$ for any $m \in \mathcal{M}$, with:

$$\tilde{\mu}^m = \beta_{\text{obs}(m)}^T X_{\text{obs}(m)} + \beta_{\text{mis}(m)}^T \mu_{\text{mis|obs}}^m,$$
$$\tilde{\sigma}^m = \beta_{\text{mis}(m)}^T \Sigma_{\text{mis|obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2,$$

with $\mu_{\text{mis|obs}}^m$ and $\Sigma_{\text{mis|obs}}^m$ defined in Appendix B.1.2 (Le Morvan et al., 2020b; Ayme et al., 2022; Zaffran et al., 2023). Crucially, $\tilde{\sigma}^m$ depends on m in a non-linear fashion, even in MCAR. That is, even in MCAR and a homoskedastic model for $Y|X$, the predictive distribution of $Y|X_{\text{obs}(M)}$ is heteroskedastic: basically, the distribution of Y is a mixture of various distributions with the mask being the latent variable. This simple example already illustrates that missing values generate strong heteroskedasticity: in Proposition 3.5, we show that under this Model 3.4 and $\mathcal{P}_{\text{MCAR}}$, the variance of the conditional distribution of Y increases when the missing pattern increases (in the sense of Definition 3.1).

Proposition 3.5 (Conditional variance increases with the mask under MCAR GLM)**.** *Under Model 3.4 and $\mathcal{P}_{\text{MCAR}}$, if the covariance matrix Σ is positive definite, Equation* (Var-1) *is satisfied.*

To prove that the variance increases with the pattern, we prove that for any $m \subset m'$, $\Sigma_{\text{mis|obs}}^{m'} \succcurlyeq \begin{pmatrix} \Sigma_{\text{mis|obs}}^m & 0 \\ 0 & \mathbf{0} \end{pmatrix}$. This is proved in Appendix B.1.2.

Next, in order to go beyond variances, we focus on inter-quantile distances as a measure of uncertainty, and establish a general result on the expected length of oracle predictive intervals.

3.1.2 Conditional Inter-quantile Isotony w.r.t. the missing data patterns

Ideally, we would like to access the oracle predictive interval (the interval satisfying Equation (MCV) with minimal expected length). Thus, in this section we are interested in characterizing its behavior with respect to M, in order to be able to mimic it. We denote this interval $\mathcal{C}_\alpha^{*,P}$, that is formally defined for any $m \in \mathcal{M}$ as:

$$\mathcal{C}_\alpha^{*,P}(\cdot, m) := \underset{\substack{\mathcal{C}_\alpha : \mathcal{X} \times \mathcal{M} \to \mathscr{P}(\mathbb{R}) \\ \text{s.t.} \mathbb{P}_P(Y \in \mathcal{C}_\alpha(X,m) | M=m) \geq 1-\alpha,}}{\arg\min} \mathbb{E}_P[\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(m)}, m))|M = m].$$

In fact, under Model 3.4, the oracle predictive interval is uniquely defined by the quantiles $\alpha/2$ and $1-\alpha/2$ of the $\mathcal{N}(\tilde{\mu}^m, \tilde{\sigma}^m)$. More importantly, this oracle interval even achieves X-conditional coverage. Proposition 3.5 shows that under $\mathcal{P}_{\text{MCAR}}$ and Model 3.4, increasing the number of missing values (in a nested way) induces an increase in the predictive uncertainty of Y: the oracle intervals, that are given by inter-quantiles intervals, are nested. Notably, this is true almost surely on X_{obs} and not only marginally.

To generalize this property beyond the Gaussian case, we introduce the inter-quantile distance, that encodes the uncertainty for conditional predictive distribution. For all $\beta \leq \frac{1}{2} \leq \gamma$, we define the inter-quantile space for quantile distributions:

$$\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M) = q_\gamma(\mathbb{P}_{Y|X_{\text{obs}(M)},M}) - q_\beta(\mathbb{P}_{Y|X_{\text{obs}(M)},M}).$$

And the following two assumptions, that are similar in spirit to (Var-1) and (Var-2)

$$\text{IQ}_{\beta,\gamma}(X_{\text{obs}(m)}, m) \overset{a.s.}{\leq} \text{IQ}_{\beta,\gamma}(X_{\text{obs}(m')}, m') \qquad \text{for any } m \subset m', \tag{IQ-1}$$

$$\mathbb{E}\left[\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M)|M = m\right] \leq \mathbb{E}\left[\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M)|M = m'\right] \qquad \text{for any } m \subset m'. \tag{IQ-2}$$

The assumptions on the quantiles and the variance are equivalent for Gaussian (conditional) distributions. As a consequence, (IQ-2) is satisfied under Model 3.4 and $\mathcal{P}_{\text{MCAR}}$ as well as under Model 3.3, while (IQ-1) is satisfied only under Model 3.4 and $\mathcal{P}_{\text{MCAR}}$. Inter-quantile assumptions are related to predictive intervals: for any distribution P such that $P_{Y|X_{\text{obs}(M)},M}$ is a.s. unimodal, the oracle predictive interval $\mathcal{C}_\alpha^{*,P}$ writes as an inter-quantile interval almost surely, that is there exist functions $\beta, \gamma : \mathcal{X} \times \mathcal{M} \to [0,1]$ such that

$$\mathcal{C}_\alpha^{*,P}(X_{\text{obs}(M)}, M) \overset{a.s.}{=} \left[q_{\beta(X_{\text{obs}(M)},M)}(P_{Y|X_{\text{obs}(M)},M}); q_{\gamma(X_{\text{obs}(M)},M)}(P_{Y|X_{\text{obs}(M)},M})\right]$$
$$\mathbb{E}_P[\gamma(X_{\text{obs}(M)}, M) - \beta(X_{\text{obs}(M)}, M)|M] \overset{a.s.}{=} 1-\alpha.$$

Indeed, to minimize the average length, the best oracle solution consists in minimizing the length conditionally to $(X_{\text{obs}(M)}, M)$, which is achieved by an inter-quantile interval, under the unimodality assumption. The quantity $\gamma(X_{\text{obs}(M)}, M) - \beta(X_{\text{obs}(M)}, M)$ corresponds to the $(X_{\text{obs}(M)}, M)$–conditional coverage, that is on average, conditionally to $M = m$, the targeted $1-\alpha$.

Yet, in practice, the constructed intervals are not the oracle ones. Therefore, a natural question is whether (IQ-2) extends to a non-oracle $\mathcal{C}_\alpha$. As generally $\mathcal{C}_\alpha$ is not based on the underlying true conditional quantiles, we focus on $\mathcal{C}_\alpha$ length instead, a quantity similar in spirit to the inter-quantile.

We consider the two following assumptions:

$$\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(m)}, m)) \overset{a.s.}{\leq} \Lambda(\mathcal{C}_\alpha(X_{\text{obs}(m')}, m')) \qquad \text{for any } m \subset m', \tag{Len-1}$$

$$\mathbb{E}\left[\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(M)}, M))|M = m\right] \leq \mathbb{E}\left[\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(M)}, M))|M = m'\right] \qquad \text{for any } m \subset m'. \tag{Len-2}$$

We have the following Theorem 3.6 on isotonicity (Len-2) under $\mathcal{P}_{\text{MCAR,YIIM}\,|\text{X}}$.

Theorem 3.6. *Let $\mathcal{C}_\alpha$ be an MCV-$\mathcal{P}_{\text{MCAR,YIIM}\,|\text{X}}$ predictive interval. There exists a predictive interval $\widetilde{\mathcal{C}_\alpha}$ which*

i) is MCV-$\mathcal{P}_{\text{MCAR,YIIM}\,|\text{X}}$,

ii) has conditional length smaller or equal to that of $\mathcal{C}_\alpha$ on each pattern,

iii) is averaged-length-isotonic w.r.t. the patterns, i.e., satisfies (Len-2).

The proof of Theorem 3.6 exploits the fact that under $\mathcal{P}_{\text{MCAR,YIIM}\,|\text{X}}$, a strategy to ensure conditional coverage w.r.t. a pattern m, is to transform $(X_{\text{obs}}(m), m)$ into $(X_{\text{obs}}(m'), m')$ by additionally masking some entries, and using the predictive interval given on pattern m'. For $m \subset m'$, we denote $X_{\text{obs}(\max(m,m'))}$ the point in which we additionally mask elements of m' in X. We have that under $\mathcal{P}_{\text{MCAR,YIIM}\,|\text{X}}$, the distribution of the data *post-masking* is equal to that of the data with more missing entries: $\mathbb{P}_{Y|X_{\text{obs}(\max(M,m'))},\max(M,m')} = P_{Y|X_{\text{obs}(m')},M=m'}$. We can leverage this observation to build intervals: if the averaged length of the predictive interval conditionally to a pattern $m \subset m'$ is larger than that conditionally to a pattern $m \subset m'$, we can replace $\mathcal{C}_\alpha(X_{\text{obs}(m)}, m)$ by $\mathcal{C}_\alpha(X_{\text{obs}(m')}, m')$, ensuring both that new interval length is smaller and that we satisfy (Len-2). Formally, we proceed by induction: starting from the pattern $m' = (1, \dots, 1)$ (no data observed), we first consider all patterns $m = (1, \dots, 1, 0, 1, \dots)$ with a single observed value, and define $\widetilde{\mathcal{C}_\alpha}(X_{\text{obs}(M)}, M)$, conditionally to $M = m$, as either $\mathcal{C}_\alpha(X_{\text{obs}(M)}, M)$ or $\mathcal{C}_\alpha(X_{\text{obs}(\max(M,m'))}, \max(M, m'))$ (depending on which expected length is smaller). We then repeat the same reasoning inductively. For a pattern m, we pick for $\widetilde{\mathcal{C}_\alpha}$ either $\mathcal{C}_\alpha(\cdot, m)$ or the minimal-length interval among all $\mathcal{C}_\alpha(\cdot, m')$ for all patterns m' that have one more missing data than m, and artificially mask on of the components of $X_{\text{obs}(m)}$ when predicting.

Interpretation: we leverage towards predictive interval construction the fact that we can transform an observed point, by removing some covariates, and recover the same distribution as the one with more missing entries. This idea will be one of the key techniques leveraged in Section 4.

As consequence of Theorem 3.6 is the following corollary, that is obtained by a minimality argument for the oracle interval (i.e., knowing that applying the aforedmentioned transformation to the oracle does not change it, as it already has minimal-expected length on each pattern):

Corollary 3.7. *Let $P \in \mathcal{P}_{\text{MCAR,YIIM}\,|\text{X}}$. Then the oracle interval $\mathcal{C}_\alpha^{*,P}$ is averaged-length-isotonic w.r.t. the patterns, i.e., satisfies* (Len-2).

Overall, (Len-2) is thus satisfied by our target sets under $\mathcal{P}_{\text{MCAR,YIIM}\,|\text{X}}$, and thus appears as a reasonable constraint to impose on our predictive sets. Indeed, it seems to be close to the minimal set of assumptions required in order to ensure that no hardness result exists (Section 2) while inducing a leverageable structure between patterns that can be compared (Theorem 3.6).

3.2 Guidelines for practitioners: which information through imputation for quantile regression?

In this section, we highlight specifities of predictive uncertainty quantification under missing covariates with respect to mean regression, and provide generic guidelines usable in practice.

Impute-then-predict. Most predictive algorithms can not cope directly with missing covariates. To bypass this, the most common approach is to impute the incomplete data via an imputation function Φ, that maps observed values to themselves and missing values to a function of the observed values. Using notations from Le Morvan et al. (2021) we note $\phi^m : \mathbb{R}^{|\text{obs}(m)|} \to \mathbb{R}^{|\text{mis}(m)|}$ the imputation function which, given a mask $m \in \mathcal{M}$, takes as input observed values and outputs imputed values, i.e., plausible values. Then, the overall imputation function Φ belongs to $\mathcal{F}^I := \left\{\Phi : \mathcal{X} \times \mathcal{M} \to \mathcal{X} : \forall j \in [\![1, d]\!], (\Phi(X, M))_j = X_j \mathbb{1}_{M_j=0} + \left(\phi^M\left(X_{\text{obs}(M)}\right)\right)_j \mathbb{1}_{M_j=1}\right\}$. The imputed data set becomes the n random variables $(\Phi(X, M), M, Y)$. In practice, Φ is the result of an algorithm $\mathcal{I}$ trained on $\left\{\left(X^{(k)}, M^{(k)}\right)\right\}_{k=1}^{n+1}$. The impact of imputation has been studied for mean regression tasks (in particular in Le Morvan et al., 2021; Ayme et al., 2023, 2024).

How to account for the missingness when imputing? Given the impact of missing covariates on the shape of prediction uncertainty discussed in Section 3.1, impute-then-predict raises a specific concern: is there a way to impute which incorporates the necessary information on the missing values?

Hereafter, we show that the answer is significantly different if we restrict ourselve to mean regression. Specifically, we show that incorporating the mask (e.g., by concatenating the mask to the features) is more critical for quantile regression. To that end, we provide in Proposition 3.8 simple models showcasing that unbiased imputation choices are sufficient to obtain an optimal model for regression but fail for quantile regression. For mean regression, the efficiency of such imputation methods have been established in practice (see e.g., Josse et al., 2024; Le Morvan et al., 2021) and Proposition 3.8 support those findings.

Proposition 3.8. *Assume* $\mathcal{P}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$ *and* $Y = \beta^{*T} X + \varepsilon$ *with* ε *s.t.* $\mathbb{E}\left[\varepsilon | X_{\text{obs}(M)}, M\right] = 0$.

i) Mean regression

- *if the covariates* $(X_j)_{j=1}^d$ *are independent, then the optimal linear model taking* $\Phi_{\text{mean}}(X, M)$ *as input is Bayes optimal, with* Φ_{mean} *the imputation by the mean;*
- *the optimal linear model taking* $\Phi_{\text{conditional mean}}(X, M)$ *as input is Bayes optimal, with* $\Phi_{\text{conditional mean}}$ *the imputation by the conditional mean;*

ii) Any quantile linear model taking unbiased imputed data as input (i.e., $\mathbb{E}[\Phi(X, M)|M] \overset{a.s.}{=} \mathbb{E}[X]$*) leads to intervals of constant expected length across patterns, thus is not Bayes optimal when* $Y \not\perp\!\!\!\perp X$.

Point i) of Proposition 3.8 illustrates that if the learner was able to retrieve the true underlying regression coefficients and the data were imputed by their mean or conditional mean, then the learned model would be the best possible at the task of predicting the conditional expectation, i.e., all necessary information is preserved by using only the imputed data set and not leveraging the associated mask. Although the non-necessity of using the mask in the conditional expectation estimation and MCAR framework does not systematically extend when the data is more complex than linear, it is insightful as even in the linear setting, the same does not hold for quantile regression.

Indeed, point ii) of the same Proposition 3.8 highlights that on the contrary a learner accessing the true underlying regression coefficients with the very same unbiased imputed data would not lead to an optimal model, as a method whose resulting predictive interval have constant lengths across the missing patterns does not retrieve the underlying heteroskedasticity induced by the missing values (Section 3.1), and thereby cannot be MCV. Precisely, the assumption on the imputation for this result corresponds for example to imputing by the feature's expectation (i.e., Φ_{mean}), the feature's conditional expectation (i.e., $\Phi_{\text{conditional mean}}$), or a random draw from a distribution whose expectation is the feature's expectation, under $\mathcal{P}_{\text{MCAR}}$. This includes MICE (van Buuren and Groothuis-Oudshoorn, 2011), which consists in imputing by random draws from the conditional distribution hence the imputed data have the same expectation than the features themselves.

Overall, Proposition 3.8 tells that $i)$ the state-of-the-art imputation method MICE is not the best choice for predictive uncertainty quantification, $ii)$ by contrast to mean regression, in the linear case imputing by the expectation or the conditional expectation is detrimental. In fact, data-independent constant imputation would result in more adaptive intervals. This is because quantile regression needs to retrieve the information on the patterns to adapt its structure to it. Therefore, when using such imputations, **a natural idea is to give the information of the mask to the model by concatenating the mask to the features**.

4 Principled unified Missing Data Augmentation (MDA) framework: `CP-MDA-Nested`*

In this section, we go beyond generic guidelines and we introduce a general framework, coined `CP-MDA-Nested`*, to generate predictive sets that achieve MCV under $\mathcal{P}_{\text{MCAR,YIM}\,|\text{X}}$. Our approach is applicable to both classification and regression tasks, by building upon any conformal score function (Vovk et al., 2005). It combines over-masking ideas introduced in Section 3, with subsampling techniques, and similar machinery than leave-one-out conformal prediction methods (Barber et al., 2021b; Gupta et al., 2022).

4.1 Presentation of `CP-MDA-Nested`*

We start by reminding the necessary concepts of split Conformal Prediction (CP) in the complete case, without missing values, before diving into the details of our unified framework `CP-MDA-Nested`*.

4.1.1 Background on split CP

Introduced in Papadopoulos et al. (2002); Vovk et al. (2005); Lei et al. (2018), split CP builds predictive regions by first splitting the n points of the training set into two disjoint sets $\text{Tr}, \text{Cal} \subset [\![1, \text{n}]\!]$, to create a *proper training set*, Tr, and a *calibration set*, Cal, of sizes $\#\text{Tr} = (1-\rho)n$ and $\#\text{Cal} = \rho n$ with $\rho \in]0, 1]$. On the proper training set, a model $\hat{f}$ (chosen by the user) is fitted, and then used to predict on the calibration set. *Conformity scores* $\mathcal{S} = \left\{ \left(s\left(X^{(k)}, Y^{(k)}; \hat{f}\right)\right)_{k \in \text{Cal}} \right\} \cup \{+\infty\}$ are computed to assess how well the fitted model $\hat{f}$ predicts the response values of the calibration points. In regression, usually the absolute value of the residuals is used, i.e. $s(x, y; \hat{\mu}) = |\hat{\mu}(x) - y|$. In classification, the simplest score is $s(x, y; \hat{p}) = 1 - \hat{p}(x)_y$ (where $\hat{p} : \mathcal{X} \mapsto [0, 1]^{\mathcal{Y}}$ outputs a vector of estimated probabilities for each class). Finally, the $(1-\alpha)$-th quantile of these scores $q_{1-\alpha}(\mathcal{S})$ (i.e., their $\lceil (1-\alpha)(\#\text{Cal}+1) \rceil$ smallest value) is computed to define the predictive region: $\widehat{C}_{n,\alpha}(x) := \{y \in \mathcal{Y} \text{ such that } s(x, y; \hat{f}) \leq q_{1-\alpha}(\mathcal{S})\}$. In regression with the absolute values of the residual score, this reduces to $\widehat{C}_{n,\alpha}(x) := [\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$.

This procedure satisfies Equation (1) for any $\hat{f}$, any (finite) sample size n, as long as the data points are exchangeable.[4] For more details on split CP, we refer to Angelopoulos and Bates (2023); Vovk et al. (2005), as well as to Manokhin (2022).

4.1.2 `CP-MDA-Nested`*

From an high level perspective, the idea is to apply split CP on top of an impute-then-predict pipeline (of imputation function Φ), and to modify the calibration step in order to ensure MCV. This is called CP-MDA, for *conformal prediction with missing data augmentation*. Generally, for a given test point $\left(X^{(n+1)}, M^{(n+1)}\right)$, CP-MDA works by artificially masking covariates in the calibration set so as to match *at least* the mask of the test point, by creating a new mask $\widetilde{M}^{(k)} = \max\left(M^{(k)}, M^{(n+1)}\right)$ for each $k \in \mathrm{Cal}$. In other words, it corresponds to discarding from the calibration set the covariates that are missing in the test point. This leads to $M^{(n+1)} \subseteq \widetilde{M}^{(k)}$, i.e., all over-masked (or *augmented*) points $\left(X^{(k)}, \widetilde{M}^{(k)}, Y^{(k)}\right)_{k\in\mathrm{Cal}}$ have at least the missing entries of $\left(X^{(n+1)}, M^{(n+1)}\right)$. The points such that $\widetilde{M}^{(k)} = M^{(n+1)}$ can be used directly as under distributional assumptions ($\mathcal{P}^{\otimes(n+1)}_{\mathrm{MCAR},\mathbb{YM}\,|\mathbb{X}}$), they now have the same mask and distribution as the test point. Yet for many calibration points it remains that $\widetilde{M}^{(k)} \neq M^{(n+1)}$ (precisely, for all the $k \in \mathrm{Cal}$ such that $M^{(k)} \not\subseteq M^{(n+1)}$). This means that those over-masked points follow another conditional distribution than the one of the test point, and MCV can not be directly ensured.

An idea is to subsample the calibration set so that the effective calibration set contains only $k \in \mathrm{Cal}$ such that $M^{(k)} \not\subseteq M^{(n+1)}$ (i.e., $\widetilde{M}^{(k)} = M^{(n+1)}$) (this is the approach followed in `CP-MDA-Exact`, Zaffran et al., 2023). However, this can lead to overly small calibration set size in high dimension, resulting in a large variance (on the coverage level and thus set size). Therefore, two questions naturally arise:

- How to build the calibration set?
- How to leverage the test point so as to account for the different distributions present in the over-masked calibration set—and with many of them not matching the test mask conditional distribution—when constructing the predictive set?

The answers we suggest define our generalized framework `CP-MDA-Nested`*, whose pseudo-code is available in Algorithm 1, and are illustrated in Figure 2.

Construction of the calibration set. `CP-MDA-Nested`* includes a subsampling step: it calibrates on the set of indices $\widetilde{\mathrm{Cal}} \subseteq \mathrm{Cal}$ provided by the user, where $\widetilde{\mathrm{Cal}}$ can be obtained with any subsampling strategy, that might even be stochastic, as long as the randomness is independent of the covariates and outputs, $\left(X^{(k)}, Y^{(k)}\right)_{k\in\mathrm{Cal}\cup\{n+1\}}$ (it can still depend on the masks). The following strategies work if the data distribution belongs to $\mathcal{P}^{\otimes(n+1)}_{\mathrm{MCAR},\mathbb{YM}\,|\mathbb{X}}$ (which is an assumption we make anyway when using `CP-MDA-Nested`* since, as we show precisely in Theorem 4.3, `CP-MDA-Nested`* is typically MCV-$\mathcal{P}^{\otimes(n+1)}_{\mathrm{MCAR},\mathbb{YM}\,|\mathbb{X}}$):

i) subsampling only the indices $\left\{k \in \mathrm{Cal} : M^{(k)} \subseteq M^{(n+1)}\right\} := \widetilde{\mathrm{Cal}}$ (this is the strategy of `CP-MDA-Exact`, Zaffran et al., 2023);

ii) no subsampling, $\widetilde{\mathrm{Cal}} := \mathrm{Cal}$ (this is the path taken by `CP-MDA-Nested`, Zaffran et al., 2023);

iii) subsampling only the indices $\left\{k \in \mathrm{Cal} : M^{(k)} \subseteq m'\right\} := \widetilde{\mathrm{Cal}}$, for some $m' \supseteq M^{(n+1)}$;

[4] Only the calibration and test data points need to be exchangeable.

iv) obtain $\widetilde{\text{Cal}}$ by subsampling from the indices $\{k \in \text{Cal} : M^{(k)} \subseteq m'\}$, for some $m' \supseteq M^{(n+1)}$, using a mixture distribution, whose weights only depend on $\left(M^{(k)}\right)_{k \in \text{Cal} \cup \{n+1\}}$.

Then, for any $k \in \widetilde{\text{Cal}}$, the over-mask is constructed, defining $\widetilde{M}^{(k)} = \max\left(M^{(k)}, M^{(n+1)}\right)$. This is schematized in Figure 2.

Leveraging temporary test points. After the subsampling step aforedmentioned, the over-masked calibration points and the test point do not necessarily have the same conditional distribution conditionally to the mask, as $M^{(n+1)} \subseteq \widetilde{M}^{(k)}$ without equality in general. In order to match those distributions, an idea is to create temporary test points (one for each calibration point) and to apply $\widetilde{M}^{(k)}$ to it. This is illustrated in green in Figure 2. `CP-MDA-Nested`* evaluates the number of over-masked calibration points that have a conformity score smaller than that of the *corresponding over-masked test point* for a given $y \in \mathcal{Y}$. Then, the predictive set includes only the $y \in \mathcal{Y}$ such that this number is small enough (a threshold that depends on α and the effective calibration size). This careful treatment of the test point allows to compare scores obtained from identical distributions conditionally on their associated mask.

4.1.3 Key comments on `CP-MDA-Nested`*

In summary, `CP-MDA-Nested`* bridges the gap between `CP-MDA-Exact` and `CP-MDA-Nested` by proposing a tighter generalized framework. On the one hand, `CP-MDA-Exact` comes with a potentially small calibration set, thus with increased variability. On the other hand, by leveraging all

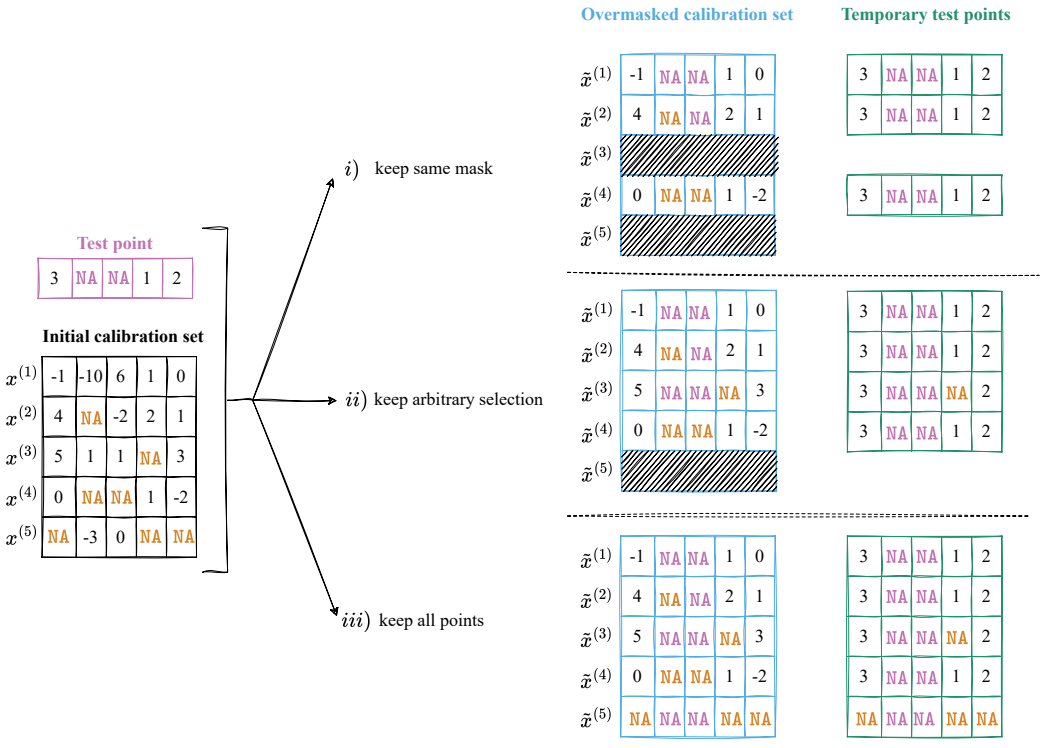


Figure 2: `CP-MDA-Nested`* illustration. Different subsampling strategies are shown, with their associated over-masked calibration set and temporary test points according to one test point.

Algorithm 1 `CP-MDA-Nested`*

Input: Training set $\left\{\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)\right\}_{k=1}^{n}$, imputation algorithm $\mathcal{I}$, learning algorithm $\mathcal{A}$ taking its values in $\mathcal{F} := \mathcal{Y}^{\mathcal{X}\times\mathcal{M}}$, calibration proportion $\rho \in]0,1]$, $\left\{\mathrm{Tr}, \mathrm{Cal}, \Phi, \hat{A}\right\}$ the output of the splitting Algorithm 2 ran on $\left\{\left\{\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)\right\}_{k=1}^{n}, \mathcal{I}, \mathcal{A}, \rho\right\}$, conformity score function $s\left(\cdot,\cdot; f\right)$ for $f \in \mathcal{F}$, significance level α, test point $\left(X^{(n+1)}, M^{(n+1)}\right)$, subsampled set of calibration indices $\widetilde{\mathrm{Cal}} \subseteq \mathrm{Cal}$

Output: Prediction set $\widehat{C}_{n,\alpha}^{\text{MDA-Nested}^\star}\left(X^{(n+1)}, M^{(n+1)}\right)$

// Generate an over-masked calibration set:

for $k \in \widetilde{\mathrm{Cal}}$ **do** Additional nested masking

$\quad \widetilde{M}^{(k)} = \max(M^{(k)}, M^{(n+1)})$

end for Over-masked calibration set generated. //

$$\widehat{C}_{n,\alpha}^{\text{MDA-Nested}^\star}\left(X^{(n+1)}, M^{(n+1)}\right) := \left\{y \in \mathcal{Y} : (1-\alpha)(1+\#\widetilde{\mathrm{Cal}}) > \sum_{k\in\mathrm{Cal}} \mathbb{1}\left\{s\left(\left(X^{(k)}, \widetilde{M}^{(k)}\right), Y^{(k)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right) < s\left(\left(X^{(n+1)}, \widetilde{M}^{(k)}\right), y; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right)\right\}\right\}$$

Algorithm 2 Split and train

Input: Imputation algorithm $\mathcal{I}$, learning algorithm $\mathcal{A}$ taking its values in $\mathcal{F} := \mathcal{Y}^{\mathcal{X}\times\mathcal{M}}$, training set $\left\{\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)\right\}_{k=1}^{n}$, calibration proportion $\rho \in]0,1]$

Output: Splitted sets of indices Tr and Cal, imputation function Φ, fitted predictor $\hat{A}$

1: Randomly split $\{1,\ldots,n\}$ into 2 disjoint sets Tr & Cal of sizes $\#\mathrm{Tr} = (1-\rho)n$ and $\#\mathrm{Cal} = \rho n$

2: Fit the imputation function: $\Phi(\cdot,\cdot) \leftarrow \mathcal{I}\left(\left\{\left(X^{(k)}, M^{(k)}\right), k \in \mathrm{Tr}\right\}\right)$

3: Fit the learning algorithm $\mathcal{A}$: $\hat{A}\left(\cdot,\cdot\right) \leftarrow \mathcal{A}\left(\left\{\left(\Phi\left(X^{(k)}, M^{(k)}\right), M^{(k)}\right), k \in \mathrm{Tr}\right\}\right)$

calibration points, including those with very few observed covariates, the average interval length produced by `CP-MDA-Nested` is typically larger than that of `CP-MDA-Exact` (cf. (Len-2)). Furthermore, `CP-MDA-Nested` is less generic than CP in the sense that it is specific to predictive *intervals* (unlike `CP-MDA-Exact` which is as generic as CP and can be plugged with any score function, including classification). Overall, `CP-MDA-Nested`* unifies this framework for any score function and provides high flexibility in the trade-offs between *efficiency* and *variability*:

- At the extreme of no subsampling at all, we obtain a generalization of `CP-MDA-Nested` which encapsulates the classification setting;
- This generalization provides tighter sets than that of `CP-MDA-Nested` in the particular case of interval-based scores (see Remark 4.1);
- At the other extreme of the strictest subsampling procedure, we retrieve `CP-MDA-Exact`;
- Any other less restrictive subsampling (possibly with a random selection between various augmented mask) belongs to this framework, providing more flexibility in the trade-offs between exact validity and statistical variability.

This overview is summarized in Table 4.

In the case where the nested predictive sets are intervals and $\widetilde{\mathrm{Cal}} = \mathrm{Cal}$, then the final predictive sets obtained through `CP-MDA-Nested`* are included in the ones of `CP-MDA-Nested`.

Remark 4.1. When $\widetilde{\mathrm{Cal}} = \mathrm{Cal}$, and using absolute value of the residuals scores or conformalized quantile regression scores (Romano et al., 2019), or any score such that $\{y \in \mathcal{Y}$ such that $s(x, y; \hat{f}) \leq$

$b\}$ for some b is an interval, then $\widehat{C}^{\text{MDA-Nested}^\star}_{n,\alpha}(\cdot) \subseteq \widehat{C}^{\text{MDA-Nested}}_{n,\alpha}(\cdot)$ (see Appendix D).

This unification allows us to provide theoretical guarantees, stated in Section 4.2, leveraging the deep connections between `CP-MDA-Nested`* and leave-one-out conformal methods (such as Barber et al., 2021b; Gupta et al., 2022). Indeed, the rationale for predicting on masked test points, using the augmented calibration masked, is that we want to treat the test and calibration points in a symmetric way. We summarize them in the following Table 5.

4.2 Theoretical guarantees on CP-MDA-Nested and `CP-MDA-Nested`* leveraging their connection to leave-one-out CP

Hereafter, we give our theoretical results on the coverage of our `CP-MDA-Nested`* algorithm.

Theorem 4.2 (Marginal validity of `CP-MDA-Nested`*)**.** *`CP-MDA-Nested`* with* $\widetilde{\text{Cal}} = \text{Cal}$ *(and thus `CP-MDA-Nested`) is MV-*$\mathcal{P}^{\text{exch}(n+1)}$ *at the level* $1 - 2\alpha$.

Theorem 4.2 provides a lower bound on `CP-MDA-Nested`* and `CP-MDA-Nested` coverage as $1 - 2\alpha$. This result is important as it equips `CP-MDA-Nested`* with $\widetilde{\text{Cal}} = \text{Cal}$ and `CP-MDA-Nested` with controlled coverage on any exchangeable distribution: they are marginally valid even on MNAR distributions or when $Y \not\perp\!\!\!\perp M \,|X$. It means that despite modifying the data set independently from X and Y and breaking the structure of (X, M, Y), the obtained estimator makes reliable predictions including when X, M, and Y are strongly dependent. This originates from the fact that the whole data set has been treated equally, including the test point.

Theorem 4.3 (Conditional validity of `CP-MDA-Nested`*)**.** *`CP-MDA-Nested`* with* $\widetilde{\text{Cal}}$ *independent of the data set* $\left(X^{(k)}, Y^{(k)}\right)_{k\in\text{Cal}\cup\{n+1\}}$ *(and thus `CP-MDA-Nested`) is MCV-*$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,YIM}\,|\text{X}}$ *at the level* $1 - 2\alpha$.

The proofs of Theorems 4.2 and 4.3 are deferred to Appendix D.1 and Appendix D.2 respectively. They are heavily based on the deep connections between `CP-MDA-Nested`* with Jackknife+ and general leave-one-out or k-fold CP (Barber et al., 2021b; Vovk, 2013; Gupta et al., 2022). Indeed, one can observe that, for each $k \in \text{Cal}$, we evaluate the conformity score of the test point $(X^{(n+1)}, M^{(n+1)}, Y^{(n+1)})$ using the k-th augmented mask, as the equivalent of evaluating the conformity score of the test point with the fitted model that has left-out the k-th calibration point. This connection between `CP-MDA-Nested`* and leave-one-out conformal approaches directly stems from the same core motivations: i) both enforce a design that use all the observations of the training or calibration sets to handle small sample sizes, ii) both need to avoid invalid designs that arise naturally when keeping all these points, such as comparing scores obtained with different predictors.

Method	`CP-MDA-Exact`	`CP-MDA-Nested`* **(new)**	`CP-MDA-Nested`
Size of actual calibration set	# points in Cal with $M \subseteq M^{(n+1)}$	Any	#Cal
Mask of the points used for calibration	exactly $M^{(n+1)}$		all, leading to $\widetilde{M}$ s.t. $M^{(n+1)} \subseteq \widetilde{M}$
Overall behavior	Too few Cal points → high coverage variance	Flexible	Too large intervals (cf. (Len-2))
Applies to classification	✓	✓(new)	✗
Outputs non-interval sets	✓	✓(new)	✗
Marginal guarantee (MV)	✓	✓(new)	✓(new)
Conditional guarantee (MCV)	✓	✓(new)	✓(new)

Table 4: Summary of the high-level characteristics of MDA algorithms, coming from the literature, as well as our novel contributions indicated by "(new)". Characteristics are given for a test point $\left(X^{(n+1)}, Y^{(n+1)}, M^{(n+1)}\right)$. Details regarding guarantees are given in Table 5.

Guarantees	MV	MCV
`CP-MDA-Exact` i.e., `CP-MDA-Nested`* with subsampling only $k \in \text{Cal}$ s.t. $M^{(k)} \subseteq M^{(n+1)}$	$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,YIM}\,|\text{X}}$, level α, with upper bound, from Zaffran et al. (2023)	$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,YIM}\,|\text{X}}$, level α, with upper bound, from Zaffran et al. (2023)
`CP-MDA-Nested`*	$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,YIM}\,|\text{X}}$, level 2α	$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,YIM}\,|\text{X}}$, level 2α
`CP-MDA-Nested`* without subsampling	$\mathcal{P}^{\text{exch}(n+1)}$, level 2α	$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,YIM}\,|\text{X}}$, level 2α

Table 5: Theoretical guarantees of `CP-MDA-Nested`* depending on the subsampling choice.

On the factor 2 and link with empirical quantile aggregation. Despite the coverage guarantee being of $1-2\alpha$ instead of the desired $1-\alpha$, in practice, our experiments in Section 5 show that `CP-MDA-Nested`* without subsampling (i.e., `CP-MDA-Nested`) tends to over-cover. This aligns with Figure 2 of Barber et al. (2021b), that illustrates the fact that leave-one-out conformal methods achieve empirically exactly $1-\alpha$ coverage, while K-fold conformal approaches over-cover. The reason behind this phenomenon is still unclear in the community, and is likely to be the same than the reason for `CP-MDA-Nested`* over-coverage, as one can see `CP-MDA-Nested`* as having access to many folds of calibration points, since each augmented calibration mask typically appears many times in the calibration set. In particular, Zaffran et al. (2023) provide MCV-$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,YIM}\,|\text{X}}$ guarantees at the level $1-\alpha$ on a modified version of `CP-MDA-Nested` which leverages this folding point of view by calibrating only on one (arbitrarily) chosen such fold. Similarly than for K-fold and leave-one-out conformal methods, we can look at `CP-MDA-Nested`* as a way to aggregate many valid empirical quantiles or p-values, one for each fold, i.e., one for each augmented mask. Due to the strong dependencies between these random variables, such an aggregation does not lead to a valid aggregated quantile or p-value, and induces a loss of coverage.

Theorem 4.3 proof approach: coupling our algorithm with a leave-one-out conformal method on a virtual complete data set. We work conditionally to the mask of the test point, $M^{(n+1)}$. Then, we introduce a randomized predictor, for which "training" consists in randomly picking one individual predictor among a bag of individual predictors, each of them corresponding to an augmented calibration mask. This bag contains exactly $2^{|\text{obs}(M^{(n+1)})|}$ possible individual predictors, where $|\text{obs}(M^{(n+1)})|$ is the number of 1s in $M^{(n+1)}$, i.e., the number of observed features in the test point. Each individual predictor in the bag is thus parametrized by a *super/over-mask* of $M^{(n+1)}$. We call such a predictor a mixture-predictor, as it basically consists in picking randomly one individual predictor in a mixture of individual predictors. That sampling has to be made independently of the data the mixture predictor is applied to, but non necessarily uniformly. Furthermore, we ensure that the individual predictor indexed by a mask M only relies on the covariates $X_{\text{obs}(M)}$ for the prediction, in order for this algorithm to be applicable in practice (e.g., an invalid design would be individual predictors that require the knowledge of some of the $X_{\text{mis}(M)}$, unobserved in practice, in order to make predictions).

We then show that our algorithm `CP-MDA-Nested`*, applied to the data set with missing entries $\left(X^{(k)}_{\text{obs}\left(M^{(k)}\right)}, Y^{(k)}, M^{(k)}\right)_{k=1}^{n+1}$, has the same guarantees in expectation as the leave-one-out conformal that uses the mixture predictor, applied onto a virtual complete data set $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$, if we make some assumptions on the missingness distribution. More specifically, we show that

there exists a coupling between the two algorithms, that ensures that the output (and thus coverage) have the same distribution. This ultimately enables us to reuse existing guarantees for leave-one-out conformal estimators.

5 A practical glimpse on the impacts of breaking the distribution's assumptions

In this concluding section, we investigate the numerical performances of `CP-MDA-Nested`* mainly outside its theoretical set of assumptions. Experiments under $\mathcal{P}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$ are provided in Section 5.1, then Section 5.2 presents numerical results when the data distribution either belongs to $\mathcal{P}_{\text{MAR}}$ or $\mathcal{P}_{\text{MNAR}}$, and finally Section 5.3 reports empirical performances when $Y \not\perp\!\!\!\perp M \,|X$.

In all experiments, the data are imputed using iterative regression (`iterative ridge` implemented in Scikit-learn, Pedregosa et al. (2011)). The predictive models are fitted on the imputed data concatenated with the mask. The prediction algorithm is a neural network, fitted to minimize the pinball loss (Sesia and Romano, 2021). For the vanilla QR, we use both the training and calibration sets for training. The training set contains 500 data points, and the calibration set 250 data points. To evaluate the marginal coverage, a test set is generated with missing values following the same distribution as on the training and calibration sets. Then, to estimate mask-conditional coverage (i.e., $\mathbb{P}(Y \in \widehat{C}_{n,\alpha}(X,m)|M=m)$ for each $m \in \mathcal{M}$), we generate another test set by imposing that the number of observations per pattern is fixed to 100, in order to ensure that the variability is not impacted by $\mathbb{P}(M=m)$. Each experiment is repeated 100 times (unless stated otherwise).

5.1 Experiments under $\mathcal{P}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$

Data generation. The data is generated with $d = 10$ according to Model 3.4 (regression), $Y = \beta^T X + \varepsilon$ with $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, \cdots, 1)^T$ and $\Sigma = \varphi(1,\cdots,1)^T(1,\cdots,1) + (1-\varphi)I_d$, $\varphi \in \{0, 0.8\}$ depending on the experiment, Gaussian noise $\varepsilon \sim \mathcal{N}(0,1) \perp\!\!\!\perp (X, M)$ and the following regression coefficients $\beta = (1, 2, -1, 3, -0.5, -1, 0.3, 1.7, 0.4, -0.3)^T$. Each of these 10 features is missing with probability 0.2 independently from anything else.

5.1.1 `CP-MDA-Nested`* provides flexibility

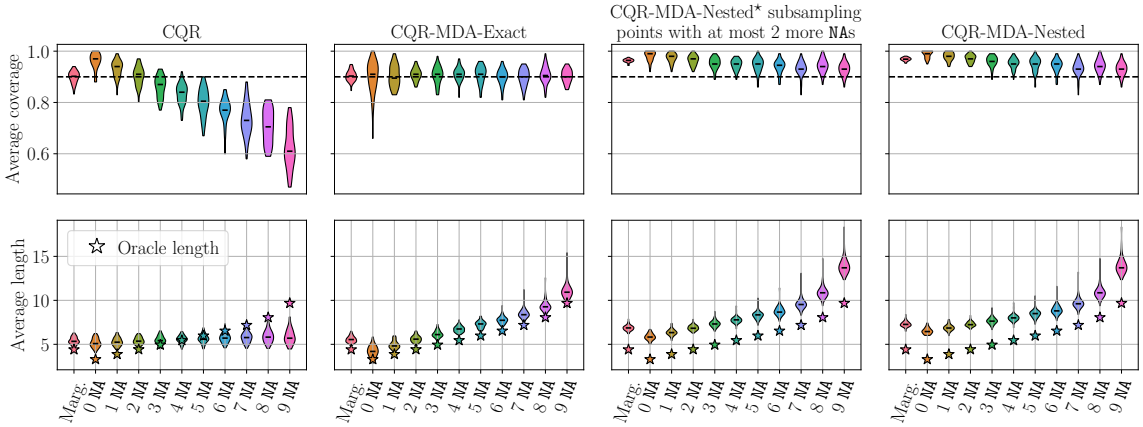
In our first experiments, we compare CQR to `CP-MDA-Exact` and `CP-MDA-Nested`, as well as `CP-MDA-Nested`* where we subsample all the calibration points that have at most two features that are missing among the observed features of the test point. As $d = 10$, there are 1024 different patterns, avoiding to display the performances of the algorithms on each of the patterns. Therefore, instead, we represent the coverage and the length of the predictive intervals depending on the mask size, a proxy for mask-conditional coverage. For each pattern size, 100 observations are drawn according to the distribution of $M|\text{size}(M)$ in the test set. In this subsection only, the number of repetition is of 50.

Figure 3a displays the results of this experiment. As noticed in Zaffran et al. (2023), CQR is not MCV-$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$ as its intervals undercover or overcover depending on the number of missing values. The three versions of `CP-MDA-Nested`* ensure that the coverage is at least $1-\alpha$ for any pattern size, as supported by our theory (Section 4.2)[5] Comparing `CP-MDA-Exact` and `CP-MDA-Nested`, we observe that `CP-MDA-Exact` is more efficient as it produces smaller

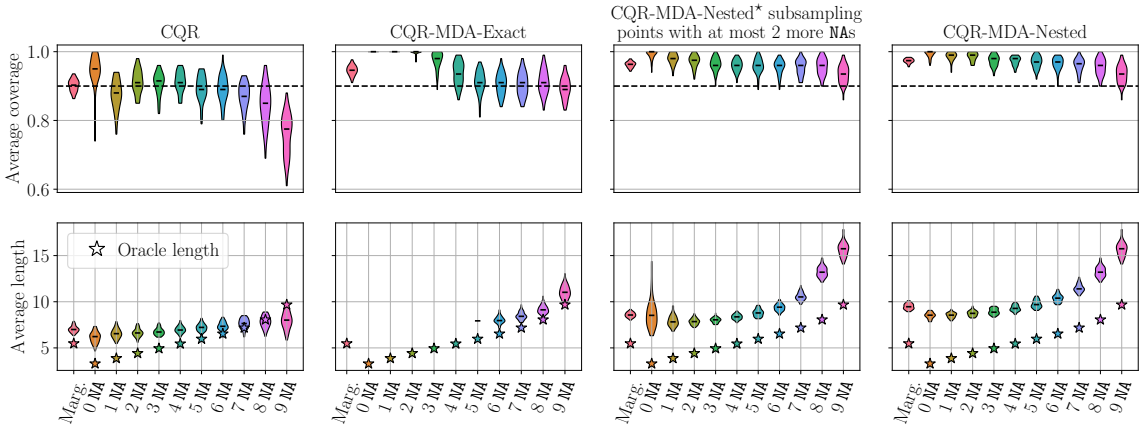
[5]Note that MCV implies validity on any mask size, but not the contrary.

intervals and its coverage is exactly of $1 - \alpha$ on average, while suffering for more variability than `CP-MDA-Nested`. The intermediate version of `CP-MDA-Nested`* allows to reduce `CP-MDA-Exact` variability while improving the efficiency of the intervals by 5.5% marginally (the comparison consists in computing the difference between `CP-MDA-Nested`* and `CP-MDA-Nested` interals' median length, and normalize it by `CP-MDA-Nested` intervals' median length), reaching nearly 10% of improvement on the test points that have no missing values. For 7 to 9 missing values, this `CP-MDA-Nested`* is equivalent to `CP-MDA-Nested` as the subsampling scheme of `CP-MDA-Nested`* boils down to keeping all the calibration points on these patterns.

`CP-MDA-Nested` reveals all its interest over `CP-MDA-Exact` in settings where the exact subsampled calibration set contains really few points for some masks (e.g., in high dimension or when the probability of missing values is high). In Figure 3b, the probability of each features being missing is increased to 0.4. We observe that `CP-MDA-Exact` outputs infinite intervals more than half of the time on the marginal test, as well as on the test sets containing between 0 and 4 missing



(a) Each features is missing with probability 0.2.



(b) Each features is missing with probability 0.4.

Figure 3: Validity and efficiency with **MCAR missing values** on dependent Gaussian features, with $\varphi = 0.8$, and such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{M} \,|\mathbf{X}$. Average coverage (top) and length (bottom) as a function of the missing pattern sizes. The black horizontal line in each violin plot corresponds to the median. The first violin plot shows the marginal coverage. The marginal test set includes 2000 observations. The mask-conditional test set includes 100 individuals for each missing data pattern size.

values. This is particularly unpractical. On the contrary, `CP-MDA-Nested` produces finite length intervals on any test point, at the cost of being overly conservative. The improvements brought by `CP-MDA-Nested`* with subsampling only the calibration points with at most 2 additional missing values are more stringent. In particular, the efficiency is improved by nearly 9.5% marginally, and is in between 8.5% and 10% on test points that have between 1 and 6 missing values.

Note that this is only one example of `CP-MDA-Nested`* for a given subsampling strategy, and that in practice the choice of strategy is highly dependent on the settings and could lead to even better performances. From now on, we restrict the subsequent experiments with `CP-MDA-Nested`* to the two extremes—`CP-MDA-Exact` and `CP-MDA-Nested`—as their main goal is to investigate the robustness beyond $\mathcal{P}_{\text{MCAR},\mathbf{Y}\perp\!\!\!\perp\mathbf{M}\,|\mathbf{X}}$. For the same reason, we do not want to restrict ourselves to the mask-size conditional coverage, as it does not imply mask conditional coverage. Therefore, we use another visualization approach that was introduced in Zaffran et al. (2023). As an appetizer, Figure 4 presents the results under $\mathcal{P}^{\otimes(n+1)}_{\text{MCAR},\mathbf{Y}\perp\!\!\!\perp\mathbf{M}\,|\mathbf{X}}$ for QR, CQR, `CP-MDA-Exact` and `CP-MDA-Nested`, using this visualization. The x-axis represents the average coverage and the average length is in the y-axis. The marker colors are associated to the different methods. A method is MCV if all the markers of its color are at the right of the vertical dotted line (90%). The design of Figure 4, and the following figures, requires a cautious interpretation. For each method we report, for the pattern having the highest (or lowest) coverage, its length and coverage. However, as this pattern may depend on the method, the length for the highest/lowest should not be directly compared between methods.

This Figure 4 illustrates that `CP-MDA-Nested`* is MCV-$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR},\mathbf{Y}\perp\!\!\!\perp\mathbf{M}\,|\mathbf{X}}$. Our hardness results of Section 2 provide a new perspective on these results:

Remark 5.1. If `CP-MDA-Nested`* was MCV on a broader class of distributions than $\mathcal{P}^{\otimes(n+1)}_{\text{MCAR},\mathbf{Y}\perp\!\!\!\perp\mathbf{M}\,|\mathbf{X}}$ for which a hardness result exists, then it would produce uninformative intervals on any distribution within this class, including $\mathcal{P}^{\otimes(n+1)}_{\text{MCAR},\mathbf{Y}\perp\!\!\!\perp\mathbf{M}\,|\mathbf{X}}$. Therefore, the fact that `CP-MDA-Nested`* obtain finite length intervals in this experiment (Figure 4) tends to confirm (with high probability) that the theory on the `CP-MDA-Nested`* MCV can not be extended to $\mathcal{P}^{\otimes(n+1)}_{\mathbf{Y}\perp\!\!\!\perp\mathbf{M}\,|\mathbf{X}}$ or $\mathcal{P}^{\otimes(n+1)}_{\text{MAR}}$ nor $\mathcal{P}^{\otimes(n+1)}_{\text{MCAR}}$. This analysis is included in Table 2, as a numerical confirmation on `CP-MDA-Nested`* theory.

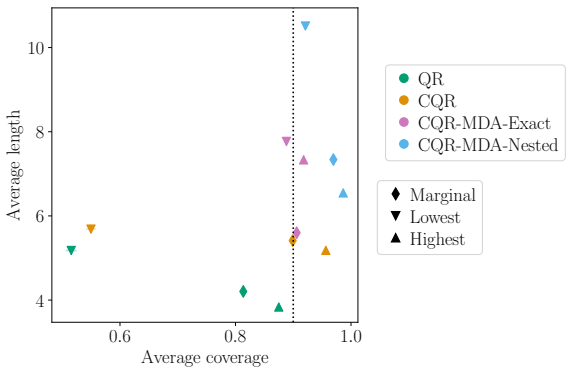


Figure 4: Validity and efficiency with **MCAR missing values** on dependent Gaussian features, with $\varphi = 0.8$, and such that $\mathbf{Y}\perp\!\!\!\perp\mathbf{M}\,|\mathbf{X}$. Colors represent the methods. Diamonds (♦) represent marginal coverage while the patterns giving the lowest and highest coverage are represented with triangles (▼ and ▲). Vertical dotted lines represent the target coverage of 90%. Experimental details: $\#\text{Tr} = 500$; $\#\text{Cal} = 250$; the marginal test set includes 2000 observations; the mask-conditional test set includes 100 individuals for each missing data pattern.

5.2 Beyond MCAR

Beyond MCAR experiments. To generate missing values under MAR or MNAR distribution, we select 6 variables (denote this set X_{missing}) out of 10 that can be missing (the 4 others form the set X_{observed}). Especially, $X_{\text{missing}} = \{X_1, X_2, X_3, X_5, X_8, X_9\}$ in order to include different range of associated regression coefficients. We used the GitHub repository associated to Muzellec et al. (2020) in order to introduce missing values in X_{missing} according to the following mechanisms, fixing the proportion of missing entries to be 20%. For each of these following settings, we run two sets of experiments: one in which the correlation between the features is high ($\varphi = 0.8$) and therefore imputing through iterative regression allows to recover quite accurately the missing values, and one in which the features are independent ($\varphi = 0$) leading to an imputation that can not be better than the marginal expectation of the features.

▶ MAR experiments (Figure 5). Missing values in X_{missing} are introduced under a MAR mechanism. To do so, a logistic model of arguments X_{observed} determines the probability of the variables in X_{missing} to be missing. This setting is declined 5 times, with different weights for the logistic model. Within each one, the experiments are repeated 100 times to assess for the variability.

▶ MNAR self-masked (Figure 6). Missing values in X_{missing} are introduced under a MNAR self masked mechanism. To do so, a logistic model determines the probability of each variable in X_{missing} to be missing by taking as input the exact same variable. This setting is declined 5 times, with different weights for the logistic model. Within each one, the experiments are repeated 100 times to assess for the variability.

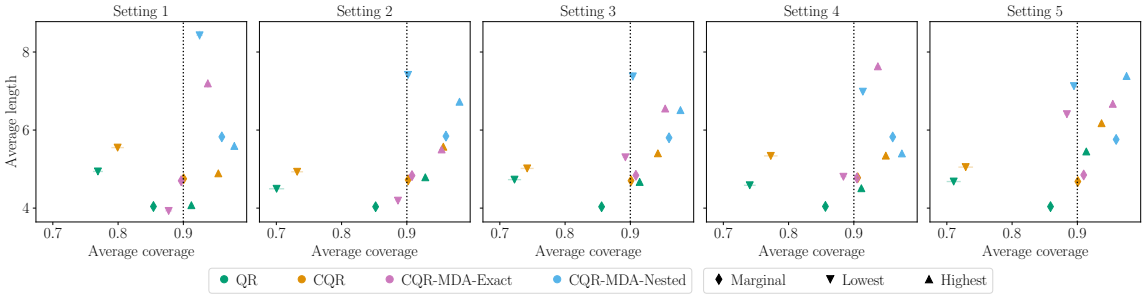
▶ MNAR quantile censorship (Figure 7). Missing values in X_{missing} are introduced under a quantile censorship MNAR mechanism. In particular, missing values are introduced at random in each q-quantile of the variables in X_{missing}. q varies between 0.5, 0.75, 0.8, 0.85, 0.9 and 0.95 and this way we obtain 6 different settings. Within each one, the experiments are repeated 100 times to assess for the variability.

These experiments show that impute-then-CQR is marginally valid even under $\mathcal{P}_{\text{MAR}}$ and $\mathcal{P}_{\text{MNAR}}$. This is expected due to Proposition 3.3 of Zaffran et al. (2023), that demonstrates that vanilla impute-then-SplitCP is marginally valid for any missing mechanism as long as the initial data set is exchangeable. However, it is not MCV, which is also expected for the same reason that the fact that it is not MCV under $\mathcal{P}_{\text{MCAR,YIM}\,|\text{X}}$. Most importantly, `CP-MDA-Nested`*, through `CP-MDA-Exact` and `CP-MDA-Nested`, is both marginally valid and MCV, despite the MCAR assumption not being satisfied, even when the imputation can not retrieve more information than the features expectation (i.e., when $\varphi = 0$). This is a positive empirical result that hints robustness of `CP-MDA-Nested`* on more complex relationships between X and M than independence.

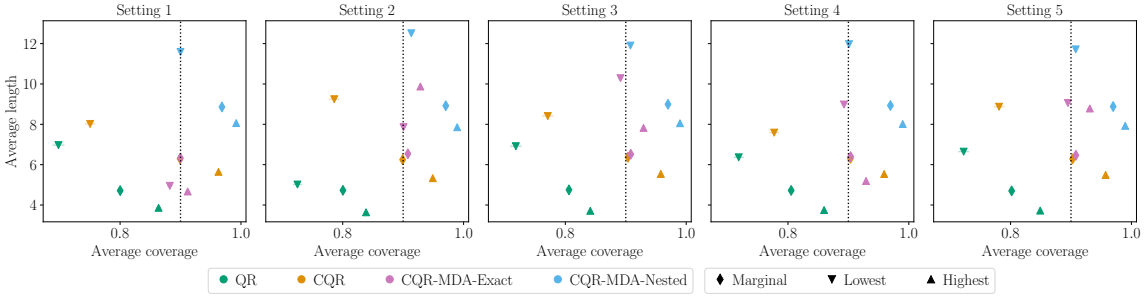
5.3 Breaking $Y \perp\!\!\!\perp M \,|X$ Assumption

Our last set experiments aim at breaking the $Y \perp\!\!\!\perp M \,|X$ assumption. We focus on $d = 3$ to be able to display all of the patterns and thus better illustrate the phenomenon. We generate data with $\varepsilon \sim \mathcal{N}(0,1) \perp\!\!\!\perp (X, M)$, $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1,1,1)^T$, $\Sigma = \varphi(1,1,1)^T(1,1,1) + (1-\varphi)I_d$, $\varphi \in \{0, 0.8\}$ depending on the experiment, and $M_i \sim \mathcal{B}(0.2)$ for any $i \in [\![1,3]\!]$, independently from X and ε. Finally: $Y = X_1 \mathbb{1}\{M_1 = 0\} + 2X_1 \mathbb{1}\{M_1 = 1\} + 3X_2 \mathbb{1}\{M_2 = 1, M_3 = 1\} + \varepsilon$. Note that according to this data generation process, the masks for which at least X_1 is missing, and the mask where X_2 and X_3 are missing, have important predictive power. As there are only 3 features that can be missing in this setting, Figures 8a and 8b represent the 7 different missing patterns.

These figures highlight that in the easiest setting where the conditional expectation imputation is able to reconstruct the missing values quite accurately ($\varphi = 0.8$, Figure 8a) `CP-MDA-Nested`*

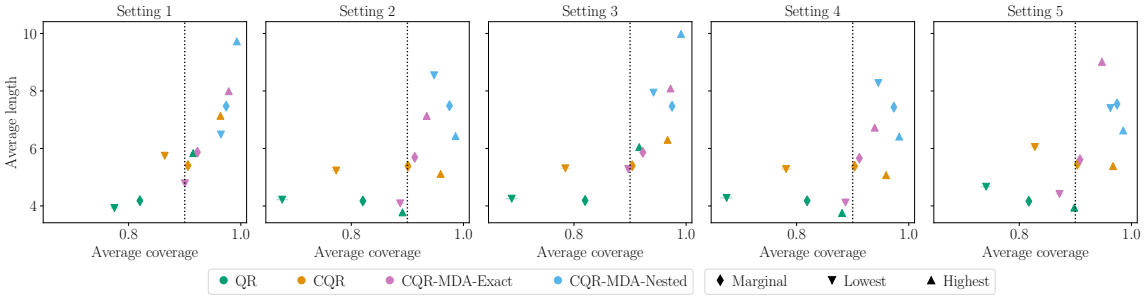


(a) Dependent Gaussian features, with $\varphi = 0.8$.

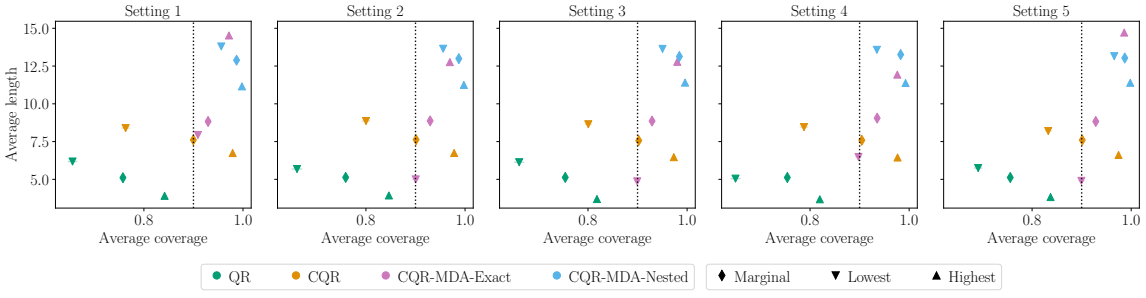


(b) Independent Gaussian features.

Figure 5: Same caption than Figure 4, for **MAR missing values**, each panel representing a different setting (set of parameters) for the missingness distribution.

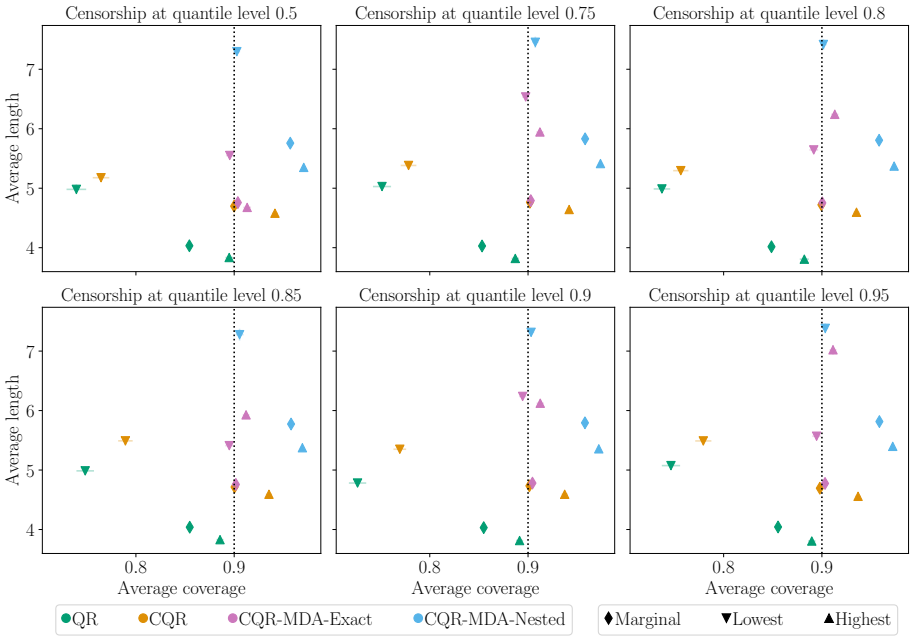


(a) Dependent Gaussian features, with $\varphi = 0.8$.

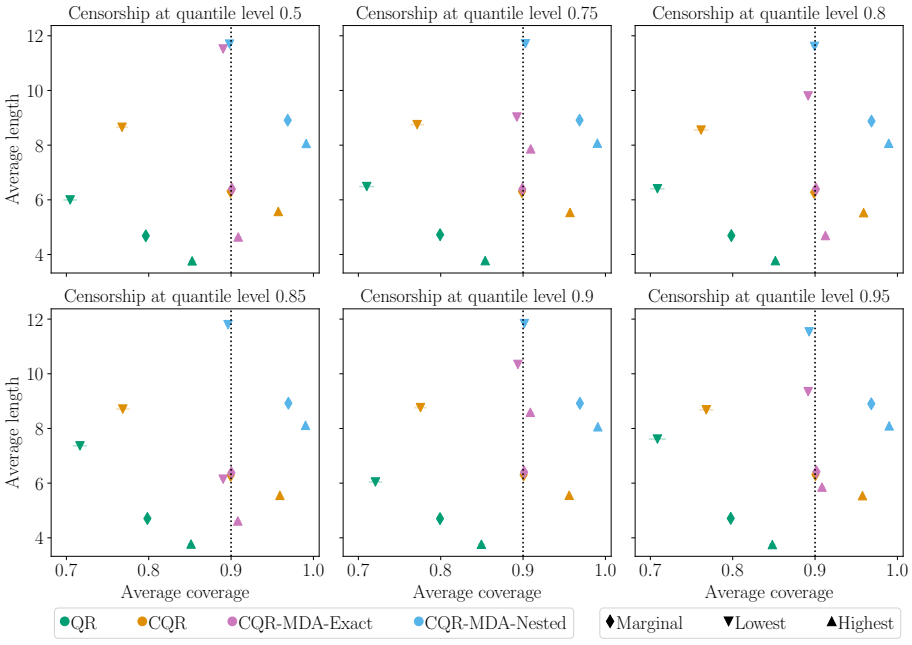


(b) Independent Gaussian features.

Figure 6: Same caption than Figure 5, for **MNAR self masked missing values**.

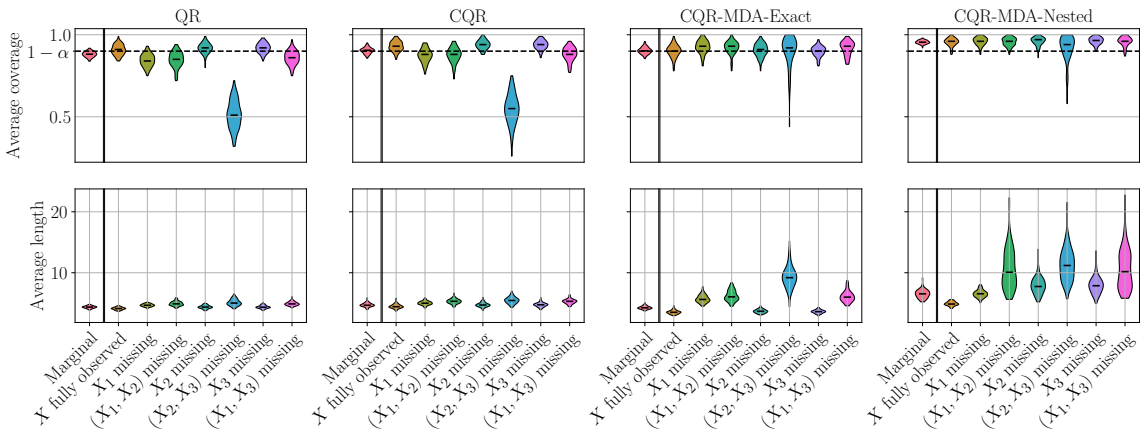


(a) Dependent Gaussian features, with $\varphi = 0.8$.

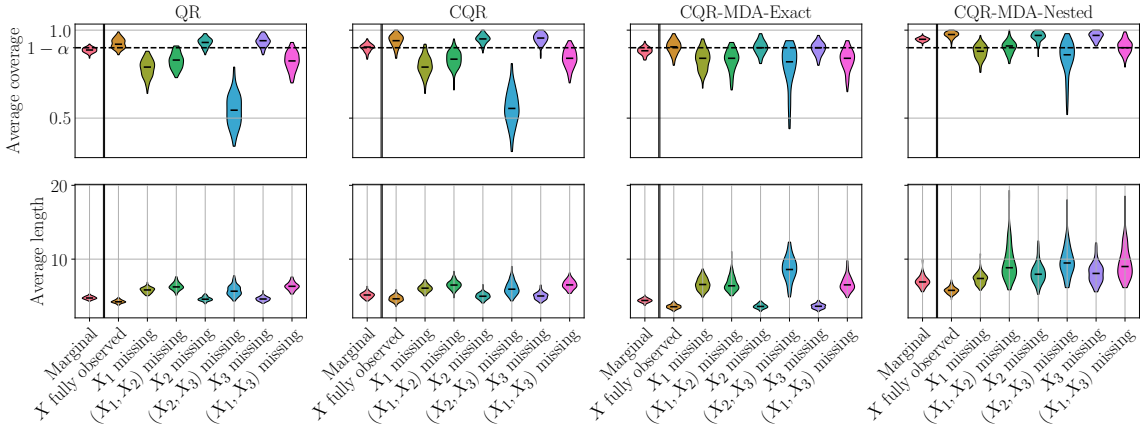


(b) Independent Gaussian features.

Figure 7: Same caption than Figure 5, for **MNAR quantile censorship missing values**.



(a) Dependent Gaussian features, with $\varphi = 0.8$.



(b) Independent Gaussian features.

Figure 8: Y and M are not independent given X, and the features are Gaussian dependent with $\varphi = 0.8$. Average coverage (top) and length (bottom) as a function of the missing patterns. The first violin plot shows the marginal coverage. The marginal test set includes 2000 observations. The mask-conditional test set includes 100 individuals for each missing data pattern.

manages to maintain MCV. However, in the hardest case of uncorrelated features ($\varphi = 0$, Figure 8b), it does not achieve MCV as it undercovers on the masks that have predictive power. Yet, `CP-MDA-Nested`* still improves upon vanilla impute-then-predict+CQR, and in particular `CP-MDA-Nested` is slightly more robust than `CP-MDA-Exact`.

Acknowledgements

This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. M. Zaffran has been awarded the 2022 Scholarship for Mathematics granted by the Séphora Berrebi Foundation which she gratefully thanks for its support. The work of J. Josse is partially supported by ANR-16-IDEX-0006. Y. Romano was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 729/21). He also thanks the Career Advancement Fellowship, Technion, for providing additional research support. The work of A. Dieuleveut is partially supported by ANR-19-CHIA-0002-01/chaire SCAI and Hi! Paris.

Appendices

The appendices are organized as follows.

Appendix A provides a the proofs for the hardness results presented in Section 2.
Appendix B contains the proofs of the Section 3 results.
Appendix C reminds the proof of leave-one-out CP in the case of randomized algorithms.
Appendix D derives `CP-MDA-Nested`* theoretical validities proofs, marginal and conditional.

A Hardness results

A.1 Reminders on object conditional impossibility results

Let's start with some reminders on the impossibility result on X-conditional coverage (also known as object conditional coverage), originally stated in Lei and Wasserman (2014), Lemma 1, and re-written more accurately and more generally in Vovk (2012), Proposition 4.

Definition A.1 (X-conditional coverage)**.** An estimator $\widehat{C}_{n,\alpha}$ achieves X-conditional coverage at the level α if for any distribution P and any x:

$$\mathbb{P}_{P^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(x\right) | X^{(n+1)} = x\right) \geq 1-\alpha.$$

Lemma A.2 (X-conditional coverage is not achievable in an informative way (Vovk, 2012))**.** *Suppose that an estimator $\widehat{C}_{n,\alpha}$ achieves X-conditional coverage at the level α. Then, for any distribution P and any x_0 such that x_0 is a non-atomic point of P:*

$$\begin{cases} \mathbb{P}_{P^{\otimes(n)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}\left(x_0\right)\right) = +\infty\right) \geq 1-\alpha, & \text{if } \mathcal{Y} \subseteq \mathbb{R} \text{ (regression)}, \\ \forall y \in \mathcal{Y}: \ \mathbb{P}_{P^{\otimes(n)}}\left(y \in \widehat{C}_{n,\alpha}\left(x_0\right)\right) \geq 1-\alpha, & \text{if } \mathcal{Y} \subseteq \mathbb{N} \text{ (classification)}. \end{cases}$$

where Λ is the Lebesgue measure.

Proof. Assume $\widehat{C}_{n,\alpha}$ be X-conditionally valid, as defined in Definition A.1.

Let P a distribution on $\mathcal{X} \times \mathcal{Y}$, and let $x_0 \in$ non-atom (P_X).

Let $\varepsilon > 0$. Let $\varepsilon_n = \sqrt{2\left(1-\left(1-\frac{\varepsilon^2}{8}\right)^{1/n}\right)}$.

Let $E \subseteq \mathcal{X}$ such that $x_0 \in E$ and $0 < P_X(E) \leq \varepsilon_n$ (this is possible as a non-atom of a distribution P_X belongs to its support).

Before diving in the details of the proof, let us define the total variation distance between two distributions P and Q on $\mathcal{Z}$, denoted $TV(P,Q)$:

$$TV(P,Q) := \sup_{Z \in \mathcal{Z}} |P(Z) - Q(Z)|.$$

$\hookrightarrow$ Classification case.

Let $y \in \mathcal{Y}$.

Define Q another distribution on $\mathcal{X} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$Q\left(A \times B\right) = P\left(A \cap E^c \times B\right) + P_X\left(A \cap E\right) S_y(B),$$

with S_y defined on $\mathcal{Y}$, which is a dirac on y.

On the one hand, exactly as in the regression case, by construction, $TV(P,Q) \leq P_X(E) \leq \varepsilon_n$. Hence, using Lemma A.3, $TV\left(P^{\otimes(n)}, Q^{\otimes(n)}\right) \leq \varepsilon$. Therefore, for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n)}(A \times B) \geq Q^{\otimes(n)}(A \times B) - \varepsilon. \tag{3}$$

On the other hand, let $x \in E$. As $\widehat{C}_{n,\alpha}$ is distribution-free X-conditionally valid, it satisfies:

$$\begin{aligned}
1-\alpha &\leq \mathbb{P}_{Q^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}(x) | X^{(n+1)} = x\right) \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\mathbb{1}\left\{Y^{(n+1)} \in \widehat{C}_{n,\alpha}(x)\right\} | X^{(n+1)} = x\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\mathbb{1}\left\{y \in \widehat{C}_{n,\alpha}(x)\right\} | X^{(n+1)} = x\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{1}\left\{y \in \widehat{C}_{n,\alpha}(x)\right\}\right] \\
&= \mathbb{P}_{Q^{\otimes(n)}}\left(y \in \widehat{C}_{n,\alpha}(x)\right).
\end{aligned}$$

Combining with Equation (3), we finally get:

$$\mathbb{P}_{P^{\otimes(n)}}\left(y \in \widehat{C}_{n,\alpha}(x)\right) \geq 1 - \alpha - \varepsilon,$$

which concludes the proof for the classification case by letting $\varepsilon \to 0$.

$\hookrightarrow$ Regression case.

Let $D > 0$.

Define Q another distribution on $\mathcal{X} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$Q(A \times B) := P(A \cap E^c \times B) + P_X(A \cap E) R(B),$$

with R defined on $\mathcal{Y}$, uniform on $[-D; D]$.

On the one hand, by construction, $TV(P,Q) \leq P_X(E) \leq \varepsilon_n$. Hence, using Lemma A.3, $TV\left(P^{\otimes(n)}, Q^{\otimes(n)}\right) \leq \varepsilon$. Therefore, for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n)}(A \times B) \geq Q^{\otimes(n)}(A \times B) - \varepsilon. \tag{3}$$

On the other hand, let $x \in E$. As $\widehat{C}_{n,\alpha}$ is distribution-free X-conditionally valid, it satisfies:

$$\begin{aligned}
1-\alpha &\leq \mathbb{P}_{Q^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}(x) | X^{(n+1)} = x\right) \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\int_{\widehat{C}_{n,\alpha}(x)} q(y|x)\mathrm{d}y\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\Lambda\left(\widehat{C}_{n,\alpha}(x) \cap [-D; D]\right) \times \frac{1}{2D}\right].
\end{aligned}$$

Note that $\Lambda\left(\widehat{C}_{n,\alpha}(x) \cap [-D; D]\right) \times \frac{1}{2D} \leq 1$. Therefore, using Lemma A.4, for any $t > 0$:

$$\begin{aligned}
\mathbb{P}_{Q^{\otimes(n)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(x) \cap [-D; D]\right) \times \frac{1}{2D} \geq 1 - t\right) &\geq 1 - \frac{\alpha}{t} \\
\mathbb{P}_{Q^{\otimes(n)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(x) \cap [-D; D]\right) \geq (1-t)2D\right) &\geq 1 - \frac{\alpha}{t} \\
\Rightarrow \mathbb{P}_{Q^{\otimes(n)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(x)\right) \geq (1-t)2D\right) &\geq 1 - \frac{\alpha}{t}.
\end{aligned}$$

Let $t = 1 - \frac{1}{\sqrt{D}}$ and obtain $\mathbb{P}_{Q^{\otimes(n)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(x)\right) \geq 2\sqrt{D}\right) \geq 1 - \frac{\alpha}{1-\frac{1}{\sqrt{D}}}$.

Combining with Equation (3), we finally get:

$$\mathbb{P}_{P^{\otimes(n)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}(x)\right) \geq 2\sqrt{D}\right) \geq 1 - \frac{\alpha}{1-\frac{1}{\sqrt{D}}} - \varepsilon.$$

Letting $\varepsilon \to 0$ and $D \to +\infty$, the result is proven for the regression case.

□

A.2 Proofs of Section 2

A.2.1 Most general distribution-free result: Theorem 2.3

Proof. Let $n \in \mathbb{N}^*$ the total training size (proper training and calibration).

Let $\alpha \in]0, 1[$.

Let $\widehat{C}_{n,\alpha}$ be MCV, as defined in Definition 2.1.

Let P a distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$.

Let $m_0 \in \mathcal{M}$.

Denote by $\rho := P_M(\{m_0\})$.

$\hookrightarrow$ Regression case.

Let $D > 0$.

Define Q another distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$Q\left(A \times L \times B\right) := P\left(A \times L \setminus \{m_0\} \times B\right) + P_{(X,M)}\left(A \times \{m_0\}\right) R\left(B\right),$$

with R defined on $\mathcal{Y}$, uniform on $[-D; D]$.

Recall that the total variation distance between two probability distributions on $\mathcal{Z}$, say P and Q, is defined as: $TV(P, Q) := \sup_{Z \in \mathcal{Z}} |P(Z) - Q(Z)|$.

On the one hand, by construction, $TV(P, Q) \leq P_M(\{m_0\}) = \rho$. Hence, using Lemma A.3: $TV(P^{\otimes(n+1)}, Q^{\otimes(n+1)}) \leq \sqrt{2\left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}$. Therefore, for any $A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n+1)}\left(A \times L \times B\right) \geq Q^{\otimes(n+1)}\left(A \times L \times B\right) - \sqrt{2\left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}. \tag{4}$$

On the other hand, as $\widehat{C}_{n,\alpha}$ is MCV, it satisfies:

$$\begin{aligned}
1-\alpha \leq & \mathbb{P}_{Q^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) | M^{(n+1)}=m_0\right) \\
&= \mathbb{E}_{Q^{\otimes(n+1)}}\left[\mathbb{1}\left\{Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right\} | M^{(n+1)}=m_0\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\mathbb{1}\left\{Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right\}\right.\right. \\
&\qquad\qquad \left.\left. | M^{(n+1)}=m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\mathbb{E}_Q\left[\mathbb{1}\left\{Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right\}\right.\right.\right. \\
&\qquad\qquad\qquad \left. | X^{(n+1)}, M^{(n+1)}=m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}\right] \\
&\qquad\qquad\qquad \left.\left. | M^{(n+1)}=m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\int_{\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)} q\left(y | X^{(n+1)}, m_0\right) \mathrm{d}y\right.\right. \\
&\qquad\qquad \left.\left. | M^{(n+1)}=m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) \cap[-D ; D]\right) \times \frac{1}{2D}\right.\right. \\
&\qquad\qquad \left.\left. | M^{(n+1)}=m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n+1)}}\left[\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) \cap[-D ; D]\right) \times \frac{1}{2D} | M^{(n+1)}=m_0\right]
\end{aligned}$$

Note that $\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) \cap[-D ; D]\right) \times \frac{1}{2D} \leq 1$ almost surely. Therefore, using Lemma A.4, for any $t>0$:

$$\begin{aligned}
\mathbb{P}_{Q^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) \cap[-D ; D]\right) \times \frac{1}{2D} \geq 1-t\right) & \geq 1-\frac{\alpha}{t} \\
\mathbb{P}_{Q^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) \cap[-D ; D]\right) \geq (1-t) 2D\right) & \geq 1-\frac{\alpha}{t} \\
\Rightarrow \mathbb{P}_{Q^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right) \geq (1-t) 2D\right) & \geq 1-\frac{\alpha}{t}.
\end{aligned}$$

Let $t=1-\frac{1}{\sqrt{D}}$ and obtain $\mathbb{P}_{Q^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right) \geq 2\sqrt{D}\right) \geq 1-\frac{\alpha}{1-\frac{1}{\sqrt{D}}}$.

Combining with Equation (4), we finally get:

$$\mathbb{P}_{P^{\otimes(n+1)}}\left(\Lambda\left(\widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right) \geq 2\sqrt{D}\right) \geq 1-\frac{\alpha}{1-\frac{1}{\sqrt{D}}}-\sqrt{2\left(1-\left(1-\frac{\rho^2}{2}\right)^{n+1}\right)}.$$

Letting $D \to +\infty$, the result is proven.

$\hookrightarrow$ Classification case.

Let $y \in \mathcal{Y}$.

Define Q another distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$Q(A \times L \times B) := P(A \times L \setminus \{m_0\} \times B) + P_{(X,M)}(A \times \{m_0\}) S(B),$$

with S defined on $\mathcal{Y}$, being null everywhere except on y (a dirac in y).

On the one hand, exactly as in the regression case, by construction, $TV(P,Q) \leq P_X(E) \leq P_M(m_0) = \rho$. $TV(P^{\otimes(n+1)}, Q^{\otimes(n+1)}) \leq \sqrt{2\left(1-\left(1-\frac{\rho^2}{2}\right)^{n+1}\right)}$. Therefore, for any $A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n+1)}\left(A \times L \times B\right) \geq Q^{\otimes(n+1)}\left(A \times L \times B\right) - \sqrt{2\left(1-\left(1-\frac{\rho^2}{2}\right)^{n+1}\right)}. \tag{4}$$

On the other hand, as $\widehat{C}_{n,\alpha}$ is MCV, it satisfies:

$$\begin{aligned}
1-\alpha &\leq \mathbb{P}_{Q^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) | M^{(n+1)} = m_0\right) \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\mathbb{1}\left\{Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right\}\right.\right. \\
&\qquad\qquad \left.\left. | M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n)}}\left[\mathbb{E}_Q\left[\mathbb{1}\left\{y \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right\} | M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}\right]\right] \\
&= \mathbb{E}_{Q^{\otimes(n+1)}}\left[\mathbb{1}\left\{y \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right\}\right] \\
&= \mathbb{P}_{Q^{\otimes(n+1)}}\left(y \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right)\right).
\end{aligned}$$

Combining with Equation (3), we finally get:

$$\mathbb{P}_{P^{\otimes(n+1)}}\left(y \in \widehat{C}_{n,\alpha}(X^{(n+1)}, m_0)\right) \geq 1-\alpha - \sqrt{2\left(1-\left(1-\frac{\rho^2}{2}\right)^{n+1}\right)}$$

which concludes the proof for the classification case.

□

The proof of Theorem 2.3 relied on the following Lemmas A.3 and A.4.

Lemma A.3. *For P and Q two probability distributions, and $n \in \mathbb{N}^*$, it holds:*

$$TV(P^n, Q^n) \leq \sqrt{2\left(1-\left(1-\frac{TV(P,Q)^2}{2}\right)^n\right)}.$$

Proof. The proof of this lemma is based on the relationship between the total variation distance and the Hellinger distance between two probability distributions denoted by $H(\cdot,\cdot)$ (see Tsybakov, 2009).

Let $n \in \mathbb{N}^*$ and let P and Q be two probability distributions.

On the one hand, note that:

$$TV(P,Q) \leq H(P,Q). \tag{5}$$

On the other hand, observe that:

$$H^2(P^n, Q^n) = 2\left(1-\left(1-\frac{H^2(P,Q)}{2}\right)^n\right). \tag{6}$$

Therefore, by combining Equations (5) and (6) (that can be found in Tsybakov, 2009), we obtain the desired result. □

Lemma A.4. *Let W be a random variable such that $0 \leq W \leq 1$ and $\mathbb{E}[W] \geq \beta$ with $\beta \in [0,1]$. Then, for any $t > 0$, it holds $\mathbb{P}(W \geq 1 - t) \geq 1 - \frac{1-\beta}{t}$.*

Proof. Let $t > 0$.

As $W \leq 1$, $1 - W \geq 0$. Therefore, using Markov's inequality:

$$\mathbb{P}(1 - W \geq t) \leq \frac{\mathbb{E}[1 - W]}{t} = \frac{1 - \mathbb{E}[W]}{t} \leq \frac{1 - \beta}{t}$$

Noting that:

$$\mathbb{P}(1 - W \geq t) = \mathbb{P}(W \leq 1 - t) = 1 - \mathbb{P}(W \geq 1 - t),$$

we finally get $\mathbb{P}(W \geq 1 - t) \geq 1 - \frac{1-\beta}{t}$. □

A.2.2 Restricting to $\mathcal{P}_{Y \perp\!\!\!\perp M \mid X}$: Proposition 2.8

Proof. The skeleton of the proof is the exactly the same than the one of Theorem 2.3, with a careful attention required in the construction of the adversarial distribution Q.

Let $n \in \mathbb{N}^*$ the total training size (proper training and calibration).

Let $\alpha \in]0,1[$.

Let $\widehat{C}_{n,\alpha}$ be MCV-$\mathcal{P}_{Y \perp\!\!\!\perp M \mid X}^{\otimes(n+1)}$.

Let $P \in \mathcal{P}_{Y \perp\!\!\!\perp M \mid X}$.

Let $(X, M, Y) \sim P$.

Let $m_0 \in \mathcal{M}$ such that $\rho := P_M(\{m_0\}) > 0$.

$\hookrightarrow$ Regression case.

Let $D > 0$.

We will now define Q another distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$ which is:

(i) close in total variation to P with respect to ρ;

(ii) such that Assumption A1 holds (to ensure that $\widehat{C}_{n,\alpha}$ is also MCV under Q);

(iii) such that there exists some subset of $\mathcal{X}$, say F_0, which determines the event of drawing mask m_0 under Q. This allows to remark that

$$\begin{aligned} &\mathbb{P}_{Q^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) \mid M^{(n+1)} = m_0\right) \\ &= \mathbb{P}_{Q^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\left(X^{(n+1)}, m_0\right) \mid X^{(n+1)} \in F_0\right). \end{aligned}$$

Let $(\tilde{X}, \tilde{M}, \tilde{Y}) \sim Q$. Q is built in the following way.

Let $F_0 \subseteq \mathcal{X}$ such that $P_X(F_0) = \rho$.

$$\begin{cases} \text{if } X \notin F_0 \text{ and } M \neq m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) = (X, M, Y), \\ \text{if } X \in F_0 \text{ or } M = m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) \sim \mathcal{U}(F_0) \times \delta_{m_0} \times \mathcal{U}([-D, D]). \end{cases}$$

Using this construction, the proof will follow as in Theorem 2.3. The only "tricky points" to check are (i), (ii), and (iii).

By construction, (iii) is directly satisfied.

Remark that by construction $\mathbb{P}\left((X, M, Y) \neq (\tilde{X}, \tilde{M}, \tilde{Y})\right) \leq 2\delta$ (the worst case scenario being if F_0 has been chosen such that $\mathbb{1}\{X \in F_0\}\mathbb{1}\{M = m_0\} \overset{a.s.}{=} 0$, leading to an equality in the

previous equation). Therefore, using Lemma A.5, we get that $TV(P,Q) \leq 2\delta$, therefore verifying (i).

The remaining task is to show that (ii) is satisfied. Let $B \in \mathcal{Y}$. We have:

$$\begin{aligned}
\mathbb{P}\left(\tilde{Y} \in B | \tilde{X}, \tilde{M}\right) &= \begin{cases} \mathbb{P}\left(Y \in B | X, M\right) \text{ if } \tilde{X} \in F_0 \\ \Lambda\left(B \cap [-D; D]\right) \dfrac{1}{2D} \text{ if } \tilde{X} \notin F_0 \end{cases} \\
&= \begin{cases} \mathbb{P}\left(Y \in B | X\right) \text{ if } \tilde{X} \in F_0 \text{ as } P \text{ satisfies Assumption A1} \\ \Lambda\left(B \cap [-D; D]\right) \dfrac{1}{2D} \text{ if } \tilde{X} \notin F_0 \end{cases} \\
&= \mathbb{P}\left(\tilde{Y} \in B | \tilde{X}\right).
\end{aligned}$$

$\hookrightarrow$ Classification case.

The idea is as previously, except that, as in the other hardness results, we replace the uniform distribution by a Dirac. In particular, let $y \in \mathcal{Y}$.

Let $(\tilde{X}, \tilde{M}, \tilde{Y}) \sim Q$. Q is built in the following way.

Let $F_0 \subseteq \mathcal{X}$ such that $P_X(F_0) = \rho$.

$$\begin{cases} \text{if } X \notin F_0 \text{ and } M \neq m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) = (X, M, Y), \\ \text{if } X \in F_0 \text{ or } M = m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) \sim \mathcal{U}(F_0) \times \delta_{m_0} \times \delta_y. \end{cases}$$

The conclusion follows as in Theorem 2.3, since, as shown in the regression case above, Q is such that: (i) $TV(P,Q) \leq 2\rho$, (ii) Assumption A1 and (iii) holds by construction. □

Lemma A.5. *Let $\mathbb{P}_Z$ and $\mathbb{P}_{Z'}$ be two distributions for the random variables X and X' taking their value in $\mathcal{Z}$. $TV(\mathbb{P}_Z, \mathbb{P}_{Z'}) \leq \mathbb{P}(Z \neq Z')$.*

Proof.

$$\begin{aligned}
TV\left(\mathbb{P}_Z, \mathbb{P}_{Z'}\right) &= \sup_{A \subseteq \mathcal{Z}} |\mathbb{P}_Z(A) - \mathbb{P}_{Z'}(A)| \\
&= \sup_{A \subseteq \mathcal{Z}} |\mathbb{E}\left[\mathbb{1}\left\{Z \in A\right\}\right] - \mathbb{E}\left[\mathbb{1}\left\{Z \in A\right\}\right]| \\
&\leq \sup_{A \subseteq \mathcal{Z}} \mathbb{E}\left[|\mathbb{1}\left\{Z \in A\right\} - \mathbb{1}\left\{Z \in A\right\}|\right] \\
&= \sup_{A \subseteq \mathcal{Z}} \mathbb{E}\left[|\mathbb{1}\left\{Z \in A\right\} - \mathbb{1}\left\{Z \in A\right\}| \mathbb{1}\left\{Z \neq Z'\right\}\right] \\
&\leq \sup_{A \subseteq \mathcal{Z}} \mathbb{E}\left[\mathbb{1}\left\{Z \neq Z'\right\}\right] \\
&= \sup_{A \subseteq \mathcal{Z}} \mathbb{P}\left(Z \neq Z'\right)
\end{aligned}$$

□

B Link between missing covariates and uncertainty

B.1 Proofs for Conditional Variance results

B.1.1 Results under $\mathcal{P}_{\text{MCAR,YIIM |X}}$ (Proposition 3.2)

Proof. Under the assumptions, $M \perp\!\!\!\perp (Y, X)$, and thus for any m:

$$\mathbb{E}\left[V(X_{\text{obs}(M)}, M) | M = m\right] = \mathbb{E}\left[V(X_{\text{obs}(m)}, m) | M = m\right]$$

$$= \mathbb{E}\left[V(X_{\text{obs}(m)}, m)\right]$$
$$= \mathbb{E}\left[\text{Var}\left(Y|X_{\text{obs}(m)}\right)\right]$$

Moreover, for any $m \subset m'$,

$$\begin{aligned}\text{Var}\left(Y|X_{\text{obs}(m')}\right) &= \mathbb{E}\left[\text{Var}\left(Y|X_{\text{obs}(m)}\right)|X_{\text{obs}(m')}\right] + \text{Var}(\mathbb{E}\left[Y|X_{\text{obs}(m)}\right]|X_{\text{obs}(m')}).\\ &\geq \mathbb{E}\left[\text{Var}\left(Y|X_{\text{obs}(m)}\right)|X_{\text{obs}(m')}\right].\end{aligned}$$

Thus $\mathbb{E}\left[\text{Var}\left(Y|X_{\text{obs}(m')}\right)\right] \geq \mathbb{E}\left[\text{Var}\left(Y|X_{\text{obs}(m)}\right)\right]$. And finally:

$$\mathbb{E}\left[V(X_{\text{obs}(M)}, M)|M = m'\right] \geq \mathbb{E}\left[V(X_{\text{obs}(M)}, M)|M = m\right].$$

□

B.1.2 Results under Gaussian Linear Model and $\mathcal{P}_{\text{MCAR}}$

Previous works (Le Morvan et al., 2020b; Ayme et al., 2022; Zaffran et al., 2023) have shown that under Model 3.4, $Y|(X_{\text{obs}(m)}, M = m) \sim \mathcal{N}\left(\tilde{\mu}^m, \tilde{\sigma}^m\right)$ for any $m \in \mathcal{M}$, with:

$$\begin{aligned}\tilde{\mu}^m &= \beta^T_{\text{obs}(m)} X_{\text{obs}(m)} + \beta^T_{\text{mis}(m)} \mu^m_{\text{mis|obs}}\\ \mu^m_{\text{mis|obs}} &= \mu^m_{\text{mis}(m)} + \Sigma^m_{\text{mis}(m),\text{obs}(m)}(\Sigma^m_{\text{obs}(m),\text{obs}(m)})^{-1}(X_{\text{obs}(m)} - \mu^m_{\text{obs}(m)}),\\ \widetilde{\sigma}^m &= \beta^T_{\text{mis}(m)} \Sigma^m_{\text{mis|obs}} \beta_{\text{mis}(m)} + \sigma^2_\varepsilon\\ \Sigma^m_{\text{mis|obs}} &= \Sigma^m_{\text{mis}(m),\text{mis}(m)} - \Sigma^m_{\text{mis}(m),\text{obs}(m)}(\Sigma^m_{\text{obs}(m),\text{obs}(m)})^{-1}\Sigma^m_{\text{obs}(m),\text{mis}(m)}.\end{aligned}$$

We now provide the proof of Proposition 3.5.

Proof. Consider Model 3.4 and assume additionally that the missing mechanism is MCAR. Therefore, for any $m \in \mathcal{M}$, $\Sigma^m = \Sigma$. Hence, for any $m \in \mathcal{M}$:

$$\text{Var}\left(Y|X_{\text{obs}(m)}, M = m\right) = \beta^T_{\text{mis}(m)} \Sigma^m_{\text{mis|obs}} \beta_{\text{mis}(m)} + \sigma^2_\varepsilon,$$

with $\Sigma^m_{\text{mis|obs}} = \Sigma_{\text{mis}(m),\text{mis}(m)} - \Sigma_{\text{mis}(m),\text{obs}(m)}(\Sigma_{\text{obs}(m),\text{obs}(m)})^{-1}\Sigma_{\text{obs}(m),\text{mis}(m)}$.

Let $(m, m') \in \mathcal{M}^2$ such that $m \subseteq m'$. Our goal is to show that:

$$\begin{aligned}\text{Var}\left(Y|X_{\text{obs}(m')}, M = m'\right) - \text{Var}\left(Y|X_{\text{obs}(m)}, M = m\right) &\geq 0\\ \beta^T_{\text{mis}(m')} \Sigma^{m'}_{\text{mis|obs}} \beta_{\text{mis}(m')} + \sigma^2_\varepsilon - \beta^T_{\text{mis}(m)} \Sigma^m_{\text{mis|obs}} \beta_{\text{mis}(m)} - \sigma^2_\varepsilon &\geq 0\\ \beta^T_{\text{mis}(m')} \Sigma^{m'}_{\text{mis|obs}} \beta_{\text{mis}(m')} - \beta^T_{\text{mis}(m)} \Sigma^m_{\text{mis|obs}} \beta_{\text{mis}(m)} &\geq 0\\ \beta^T_{\text{mis}(m')} \Sigma^{m'}_{\text{mis|obs}} \beta_{\text{mis}(m')} - \beta^T_{\text{mis}(m')} \begin{pmatrix} \Sigma^m_{\text{mis|obs}} & 0\\ 0 & \mathbf{0}\end{pmatrix} \beta_{\text{mis}(m')} &\geq 0\\ \beta^T_{\text{mis}(m')} \left(\Sigma^{m'}_{\text{mis|obs}} - \begin{pmatrix} \Sigma^m_{\text{mis|obs}} & 0\\ 0 & \mathbf{0}\end{pmatrix}\right) \beta_{\text{mis}(m')} &\geq 0,\end{aligned}$$

holds for any β. Therefore, we have to show that $\Sigma^{m'}_{\text{mis|obs}} - \begin{pmatrix} \Sigma^m_{\text{mis|obs}} & 0\\ 0 & \mathbf{0}\end{pmatrix}$ is semi-definite positive.

The marginal covariance matrix Σ can be rewritten by blocks in the following way:

$$\Sigma = \begin{pmatrix} A & B & C\\ B^T & D & E\\ C^T & E^T & F\end{pmatrix},$$

where:

$$\begin{cases} A = \Sigma_{\text{mis}(m),\text{mis}(m)}, \\ \begin{pmatrix} D & E \\ E^T & F \end{pmatrix} = \Sigma_{\text{obs}(m),\text{obs}(m)}, \\ \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} = \Sigma_{\text{mis}(m'),\text{mis}(m')}, \\ F = \Sigma_{\text{obs}(m'),\text{obs}(m')}. \end{cases}$$

Additionally, assume that $\Sigma > 0$ (that is, Σ is definite positive).

Therefore, $D > 0, F > 0$. Thus F is invertible, of inverse $F^{-1} > 0$. Furthermore, $G := D - EF^{-1}E^T$ is also positive definite, as it is the sum of $D > 0$ and $EF^{-1}E^T \geq 0$, and thus G is invertible.

$\Sigma^m_{\text{mis}|\text{obs}}$ and $\Sigma^{m'}_{\text{mis}|\text{obs}}$ can be rewritten using the previous decomposition.

On the one hand, for m it gives:

$$\begin{aligned} \Sigma^m_{\text{mis}|\text{obs}} =& A - \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} D & E \\ E^T & F \end{pmatrix}^{-1} \begin{pmatrix} B^T \\ C^T \end{pmatrix} \\ =& A - \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} G^{-1} & -G^{-1}EF^{-1} \\ -F^{-1}E^TG^{-1} & F^{-1} + F^{-1}E^TG^{-1}EF^T \end{pmatrix} \begin{pmatrix} B^T \\ C^T \end{pmatrix} \\ =& A - \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} G^{-1}B^T - G^{-1}EF^{-1}C^T \\ -F^{-1}E^TG^{-1}B^T + F^{-1}C^T + F^{-1}E^TG^{-1}EF^TC^T \end{pmatrix} \\ =& A - BG^{-1}B^T + BG^{-1}EF^{-1}C^T \\ & + CF^{-1}E^TG^{-1}B^T - CF^{-1}C^T - CF^{-1}E^TG^{-1}EF^TC^T \\ \text{(rearranging)} =& A - CF^{-1}C^T - BG^{-1}B^T + BG^{-1}EF^{-1}C^T \\ & + CF^{-1}E^TG^{-1}B^T - CF^{-1}E^TG^{-1}EF^TC^T \\ =& A - CF^{-1}C^T - BG^{-1}\left(B^T - EF^{-1}C^T\right) \\ & + CF^{-1}E^TG^{-1}\left(B^T - EF^TC^T\right) \\ =& A - CF^{-1}C^T - \left(B - CF^{-1}E^T\right)G^{-1}\left(B^T - EF^{-1}C^T\right), \end{aligned}$$

and by denoting $H := B - CF^{-1}E^T$, we finally obtain (as F is symmetric):

$$\Sigma^m_{\text{mis}|\text{obs}} = A - CF^{-1}C^T - HG^{-1}H^T.$$

On the other hand, for m':

$$\begin{aligned} \Sigma^{m'}_{\text{mis}|\text{obs}} &= \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} - \begin{pmatrix} C \\ E \end{pmatrix} F^{-1} \begin{pmatrix} C^T & E^T \end{pmatrix} \\ &= \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} - \begin{pmatrix} CF^{-1}C^T & CF^{-1}E^T \\ EF^{-1}C^T & EF^{-1}E^T \end{pmatrix} \\ &= \begin{pmatrix} A - CF^{-1}C^T & B - CF^{-1}E^T \\ B^T - EF^{-1}C^T & D - EF^{-1}E^T \end{pmatrix} \\ &= \begin{pmatrix} A - CF^{-1}C^T & B - CF^{-1}E^T \\ B^T - EF^{-1}C^T & G \end{pmatrix} \\ \Sigma^{m'}_{\text{mis}|\text{obs}} &= \begin{pmatrix} A - CF^{-1}C^T & H \\ H^T & G \end{pmatrix} \end{aligned}$$

Therefore, combining the two terms and rewriting together, we obtain:

$$\Sigma^{m'}_{\text{mis}|\text{obs}} - \begin{pmatrix} \Sigma^{m}_{\text{mis}|\text{obs}} & 0 \\ 0 & \mathbf{0} \end{pmatrix} = \begin{pmatrix} A - CF^{-1}C^T & H \\ H^T & G \end{pmatrix} - \begin{pmatrix} A - CF^{-1}C^T - HG^{-1}H^T & 0 \\ 0 & \mathbf{0} \end{pmatrix}$$
$$= \begin{pmatrix} A - CF^{-1}C^T - A + CF^{-1}C^T + HG^{-1}H^T & H \\ H^T & G \end{pmatrix}$$
$$\Sigma^{m'}_{\text{mis}|\text{obs}} - \begin{pmatrix} \Sigma^{m}_{\text{mis}|\text{obs}} & 0 \\ 0 & \mathbf{0} \end{pmatrix} = \begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix}.$$

Hence, our objective is to show that $\begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix}$ is semi-definite positive.

Let $z = \begin{pmatrix} x & y \end{pmatrix} \in \mathbb{R}^{1\times(\#m+(\#m'-\#m))}$.

$$\begin{aligned} z \begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix} z^T &= \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix} \begin{pmatrix} x^T \\ y^T \end{pmatrix} \\ &= xHG^{-1}H^Tx^T + xHy^T + yH^Tx^T + yGy^T \\ &= xHG^{-1}GG^{-1}H^Tx^T + xHG^{-1}Gy^T + yGG^{-1}H^Tx^T + yGy^T \\ &= xHG^{-1}G\left(G^{-1}H^Tx^T + y^T\right) + yG\left(G^{-1}H^Tx^T + y^T\right) \\ &= \left(xHG^{-1} + y\right)G\left(G^{-1}H^Tx^T + y^T\right) \\ &= \left(xHG^{-1} + y\right)G\left(xHG^{-1} + y\right)^T \\ &\geq 0 \text{ as } G \text{ is positive definite.} \end{aligned}$$

□

B.2 Impact of the imputation under a linear quantile regression model (Proposition 3.8)

To prove Item i) of Proposition 3.8, we prove the following Lemma B.1.

Lemma B.1. *Assume $\mathcal{P}_{\text{MCAR}}$, and $Y = \beta^{*T}X + \varepsilon$ with ε s.t. $\mathbb{E}\left[\varepsilon|X_{\text{obs}(M)}, M\right] = 0$.*

Then $\mathbb{E}\left[Y|X_{\text{obs}(M)}, M\right] = \beta^{*T}\Phi_{\text{conditional mean}}(X, M)$, with $\Phi_{\text{conditional mean}}$ the imputation by the conditional mean. Furthermore, if the covariates are independent, then $\mathbb{E}\left[Y|X_{\text{obs}(M)}, M\right] = \beta^{*T}\Phi_{\text{mean}}(X, M)$, with Φ_{mean} the imputation by the mean.

Proof.

$$\begin{aligned} \mathbb{E}\left[Y|X_{\text{obs}(M)}, M\right] = \mathbb{E}\left[\beta^{*T}X|X_{\text{obs}(M)}, M\right] &= \sum_{i=1}^{d} \beta^*_i \mathbb{E}\left[X_i|X_{\text{obs}(M)}, M\right] \\ &= \sum_{i=1}^{d} \beta^*_i (X_i \mathbb{1}\left\{i \in \text{obs}(M)\right\} \\ &\qquad + \mathbb{E}\left[X_i|X_{\text{obs}(M)}, M\right] \mathbb{1}\left\{i \notin \text{obs}(M)\right\}) \\ \mathcal{P}_{\text{MCAR}} \rightarrow &= \sum_{i=1}^{d} \beta^*_i (X_i \mathbb{1}\left\{i \in \text{obs}(M)\right\} \\ &\qquad + \mathbb{E}\left[X_i|X_{\text{obs}(M)}\right] \mathbb{1}\left\{i \notin \text{obs}(M)\right\}) \end{aligned}$$

$$= \sum_{i=1}^{d} \beta_i^* \left(\Phi_{\text{conditional mean}}(X, M)\right)_i$$

$$\text{if } (X_i)_{i=1}^d \perp\!\!\!\perp, \mathbb{E}\left[X_i | X_{\text{obs}(M)}\right] = \mathbb{E}\left[X_i\right] \rightarrow = \sum_{i=1}^{d} \beta_i^* \left(\Phi_{\text{mean}}(X, M)\right)_i$$

□

To prove Item ii) of Proposition 3.8, we prove the following Proposition B.2. Indeed, the oracle predictive intervals vary at least once in length we respect to the patterns, as, on the one hand, under $\mathcal{P}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$ Equation (Len-2) holds and, on the other hand, when $Y \not\perp\!\!\!\perp X$ the variance of Y given X is different than the overall variance of Y.

Proposition B.2 (Non-adaptivity of the linear quantile regression)**.** *Assume that:*

i) the quantile regression is learned within the class of linear models;

ii) the (random) values used to impute have the same expectation than the feature itself, i.e., $\mathbb{E}\left[\Phi(X, m) | M = m\right] = \mathbb{E}\left[X\right]$ *for any* $m \in \mathcal{M}$ *such that* $\mathbb{P}(M = m) > 0$.

Then the expectation of the predictive intervals length is independent of the missing pattern.

Proof. Since the quantile regression is learned within the class of linear models, the fitted quantile functions (upper and lower) can be written as $\widehat{q}_\delta(z) = \beta_\delta^T z + \beta_\delta^0$, with $\beta \in \mathbb{R}^d$ and $\beta^0 \in \mathbb{R}$. Therefore, the length of the resulting interval L_α at some—imputed—point $\Phi(X_{\text{obs}(M)}, M)$ will be:

$$\begin{aligned} L_\alpha(\Phi(X_{\text{obs}(M)}, M)) &:= \widehat{q}_{\delta_{(u)}}(\Phi(X_{\text{obs}(M)}, M)) - \widehat{q}_{\delta_{(l)}}(\Phi(X_{\text{obs}(M)}, M)) \\ &= \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T\right) \Phi(X_{\text{obs}(M)}, M) + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0, \end{aligned}$$

with $\delta_{(l)}$ and $\delta_{(u)}$ chosen by the user or fixed by the algorithm such that $\delta_{(u)} - \delta_{(l)} = 1 - \alpha$. Thus:

$$\begin{aligned} \mathbb{E}\left[L_\alpha(\Phi(X_{\text{obs}(M)}, M))\right] &= \mathbb{E}\left[\left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T\right) \Phi(X_{\text{obs}(M)}, M) + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0\right] \\ &= \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T\right) \mathbb{E}\left[\Phi(X_{\text{obs}(M)}, M)\right] + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0. \end{aligned}$$

Let $m \in \mathcal{M}$ such that $\mathbb{P}(M = m) > 0$. Conditioning by m:

$$\mathbb{E}\left[L_\alpha(\Phi(X_{\text{obs}(M)}, M)) | M = m\right] = \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T\right) \mathbb{E}\left[\Phi(X_{\text{obs}(M)}, M) | M = m\right] + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0.$$

Given the assumption that $\mathbb{E}\left[\Phi(X_{\text{obs}(M)}, M) | M = m\right] = \mathbb{E}\left[X\right]$, one can conclude that:

$$\mathbb{E}\left[L_\alpha(\Phi(X_{\text{obs}(M)}, M)) | M = m\right] = \sum_{j=1}^{d} \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T\right)_j \mathbb{E}\left[X\right] + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0 \perp\!\!\!\perp M.$$

□

C Leave-one-out predictive sets for randomized algorithms

We provide in this section a more detailed proof of leave-one-out or k-fold cross-conformal (Vovk, 2013) and jackknife+ (Barber et al., 2021b) methods which allows us to highlight where exactly the arguments of data exchangeability and symmetrical algorithm play a role. In particular, by emphasizing these precise influences, we can understand how to include a non-deterministic symmetrical algorithm (such as Random Forest or Stochastic Gradient Descent).

C.1 On the definition of randomized symmetric algorithms

Definition C.1 (Randomized learning algorithm)**.** A randomized learning algorithm is defined as:

$$\mathcal{A} : \left(\bigcup_{n \geq 0} (\mathcal{X} \times \mathcal{Y})^n \right) \times [0,1] \mapsto \mathcal{Y}^{\mathcal{X}}$$
$$\left(X^{(k)}, Y^{(k)} \right)_{k=1}^{n} \times \xi \mapsto \hat{A}(\cdot)$$

where ξ encodes the randomness of $\mathcal{A}$.

Definition C.2 (Randomized symmetric algorithm (Kim and Barber, 2023))**.** A randomized learning algorithm $\mathcal{A}$ is symmetric if for any data set $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n}$, for any permutation σ on $[\![1, n]\!]$, there exists a coupling that maps $\xi \sim \mathcal{U}([0,1])$ to $\xi' \sim \mathcal{U}([0,1])$, which depends only on σ, s.t.:

$$\mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n}; \xi\right) = \mathcal{A}\left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))}\right)_{k=1}^{n}; \xi'\right).$$

C.2 Detailing leave-one-out conformal predictors validity proof

Let $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ be exchangeable, and $\mathcal{A}$ a (possible randomized) symmetric algorithm.

Let s be a conformity score function. For $i \in [\![1, n]\!]$, denote $\hat{A}_{-i}(\cdot) := \mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{\substack{k=1 \\ k \neq i}}^{n}\right)$, that is the fitted left-one-out algorithm, removing data point i.

Consider the leave-one-out conformal estimator defined as:

$$\widehat{C}_{n,\alpha}^{\text{LOO}}(x) := \left\{ y \in \mathcal{Y} : \sum_{k=1}^{n} \mathbb{1}\left\{ s\left(X^{(k)}, Y^{(k)}; \hat{A}_{-k}\right) < s\left(x, y; \hat{A}_{-k}\right) \right\} < (1-\alpha)(n+1) \right\}$$

.

Previous works (Barber et al., 2021b; Gupta et al., 2022) have proven that under exchangeability of $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ and symmetry of $\mathcal{A}$, $\mathbb{P}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}^{\text{LOO}}\left(X^{(n+1)}\right)\right) \geq 1 - 2\alpha$. We recall below the key proof's steps, detailing the last one which uses the exchangeability and symmetry arguments.

Step 1. Remark that:

$$\begin{aligned}
&\left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{LOO}}\left(X^{(n+1)}\right) \right\} \\
&= \left\{ \sum_{k=1}^{n} \mathbb{1}\left\{ s\left(X^{(k)}, Y^{(k)}; \hat{A}_{-k}\right) < s\left(X^{(n+1)}, Y^{(n+1)}; \hat{A}_{-k}\right) \right\} \geq (1-\alpha)(n+1) \right\} \\
&:= \left\{ \sum_{k=1}^{n} \mathbb{1}\left\{ S^{(k),n+1} < S^{(n+1),k} \right\} \geq (1-\alpha)(n+1) \right\} \\
&:= \left\{ \sum_{k=1}^{n} \mathcal{C}_{n+1,k} \geq (1-\alpha)(n+1) \right\}.
\end{aligned}$$

with $S^{(i),j} := s\left(X^{(i)}, Y^{(i)}; \hat{A}_{-(i,j)}\right)$ the score on data point i of the predictor that has been fitted without seeing nor data point i nor data point j, for $(i,j) \in [\![1, n+1]\!]^2$ and extending $\hat{A}_{-i}$ to $\hat{A}_{-(i,j)} := \mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{\substack{k=1 \\ k \notin \{i,j\}}}^{n+1}\right)$, where the $n+1$ data point is added.

Denote by $\mathcal{C}_{\mathcal{A}}$ the function building the comparison matrix $\mathcal{C} \in \{0,1\}^{(n+1)\times(n+1)}$: $\mathcal{C}_{\mathcal{A}}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}\right)_{i,j} = \mathbb{1}\left\{S^{(i),j} > S^{(j),i}\right\} = \mathcal{C}_{i,j}$.

Step 2. Deterministically, Barber et al. (2021b) shows that $\#\{i \in [\![1, n+1]\!] : \sum_{j=1}^{n+1} \mathcal{C}_{i,j} \geq (1-\alpha)(n+1)\} \leq 2\alpha(n+1)$. This is shown for *any* comparison matrix.

Step 3. The last (and crucial) step of leave-one-out conformal predictors is to show that for any permutation σ on $[\![1, n+1]\!]$ it holds: $\left(\mathcal{C}_{\sigma(i),\sigma(j)}\right)_{i,j} \overset{d}{=} (\mathcal{C}_{i,j})_{i,j}$.

$$
\begin{aligned}
\mathcal{C}_{\sigma(i),\sigma(j)} &= \mathcal{C}_{\mathcal{A}}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}\right)_{\sigma(i),\sigma(j)} \\
&= \mathbb{1}\Bigg\{ s\left(Y^{(\sigma(i))}, X^{(\sigma(i))}, \mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1,k\notin\{\sigma(i),\sigma(j)\}}^{n+1}; \xi\right)\right) \\
&\qquad > s\left(Y^{(\sigma(j))}, X^{(\sigma(j))}, \mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1,k\notin\{\sigma(i),\sigma(j)\}}^{n+1}; \xi\right)\right)\Bigg\} \\
&= \mathbb{1}\Bigg\{ s\left(Y^{(\sigma(i))}, X^{(\sigma(i))}, \mathcal{A}\left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))}\right)_{k=1,k\notin\{i,j\}}^{n+1}; \xi'_\sigma\right)\right) \\
&\qquad > s\left(Y^{(\sigma(j))}, X^{(\sigma(j))}, \mathcal{A}\left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))}\right)_{k=1,k\notin\{i,j\}}^{n+1}; \xi'_\sigma\right)\right)\Bigg\} \quad \mathcal{A} \text{ is symmetric} \\
&= \mathcal{C}_{\mathcal{A}}\left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))}\right)_{k=1}^{n+1}\right)_{i,j}
\end{aligned}
$$

Thus, leveraging the fact that $\xi'_\sigma \perp\!\!\!\perp \left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ and that $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are exchangeable, we obtain that:

$$
\left(\mathcal{C}_{\sigma(i),\sigma(j)}\right)_{i,j\in[\![1,n+1]\!]^2} \overset{d}{=} \mathcal{C}_{\mathcal{A}}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}\right) = (\mathcal{C}_{i,j})_{i,j\in[\![1,n+1]\!]^2}\,.
$$

Hence, for any permutation σ on $[\![1, n+1]\!]$ it holds that $\Pi_\sigma^T \mathcal{C} \Pi_\sigma \overset{d}{=} \mathcal{C}$, concluding the proof as then each element of $[\![1, n+1]\!]$ is equally likely to belong to $\{i \in [\![1, n+1]\!] : \sum_{j=1}^{n+1} \mathcal{C}_{i,j} \geq (1-\alpha)(n+1)\}$.

D Theory on `CP-MDA-Nested`* and `CP-MDA-Nested`

Let us first remark that $\widehat{C}_{n,\alpha}^{\text{MDA-Nested}^\star}(\cdot) \subseteq \widehat{C}_{n,\alpha}^{\text{MDA-Nested}}(\cdot)$ when the conformity score function outputs intervals and $\widetilde{\text{Cal}} = \text{Cal}$ (Remark 4.1).

Proof.

$$
\begin{aligned}
\left\{Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}}\left(X^{(n+1)}, M^{(n+1)}\right)\right\} &= \left\{Y^{(n+1)} > \widehat{Q}_{1-\alpha}\left(\mathcal{U}_\alpha\left(X^{(n+1)}\right)\right)\right. \\
&\qquad \left. \text{or } Y^{(n+1)} < \widehat{Q}_\alpha\left(\mathcal{L}_\alpha\left(X^{(n+1)}\right)\right)\right\} \\
&= \left\{(1-\alpha)(\#\text{Cal}+1) \leq \sum_{k=1}^{n} \mathbb{1}\left\{Y^{(n+1)} > u_\alpha^{(k)}\left(X^{(n+1)}\right)\right\}\right.
\end{aligned}
$$

$$
\begin{aligned}
&\text{or } (1-\alpha)(\#\text{Cal}+1) \leq \sum_{k=1}^{n} \mathbb{1}\left\{Y^{(n+1)} < \ell_\alpha^{(k)}\left(X^{(n+1)}\right)\right\}\Bigg\} \\
\subset &\Bigg\{(1-\alpha)(\#\text{Cal}+1) \leq \sum_{k=1}^{n} \mathbb{1}\Big\{Y^{(n+1)} > u_\alpha^{(k)}\left(X^{(n+1)}\right) \\
&\qquad\qquad \text{or } Y^{(n+1)} < \ell_\alpha^{(k)}\left(X^{(n+1)}\right)\Big\}\Bigg\} \\
= &\Bigg\{(1-\alpha)(\#\text{Cal}+1) \\
&\quad \leq \sum_{k=1}^{n} \mathbb{1}\Big\{s\left(\left(X^{(n+1)}, \widetilde{M}^{(k)}\right), Y^{(n+1)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right) \\
&\qquad\qquad > s\left(\left(X^{(k)}, \widetilde{M}^{(k)}\right), Y^{(k)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right)\Big\}\Bigg\} \\
= &\left\{Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^\star}\left(X^{(n+1)}, M^{(n+1)}\right)\right\}
\end{aligned}
$$

□

Therefore, any upper bound on the miscoverage of `CP-MDA-Nested`* extends to `CP-MDA-Nested`.

D.1 Marginal validity of `CP-MDA-Nested`*.

The proof of Theorem 4.2 is highly inspired by the leave-one-out conformal predictors proof, from Barber et al. (2021b) and detailed previously in Appendix C.

Proof. One can see this proof as analogous of the one of leave-one-out conformal predictors, where "predicting on point i with point j left out" corresponds to "predicting on point i when additionally masking it with the mask of point j".

Step 1.

$$
\begin{aligned}
&\left\{Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^\star}\left(X^{(n+1)}, M^{(n+1)}\right)\right\} \\
&\quad = \Bigg\{(1-\alpha)(\#\text{Cal}+1) \\
&\qquad \leq \sum_{k\in\text{Cal}} \mathbb{1}\Big\{s\left(\left(X^{(n+1)}, \widetilde{M}^{(k)}\right), Y^{(n+1)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right) \\
&\qquad\qquad > s\left(\left(X^{(k)}, \widetilde{M}^{(k)}\right), Y^{(k)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right)\Big\}\Bigg\} \\
&\quad := \left\{(1-\alpha)(\#\text{Cal}+1) \leq \sum_{k\in\text{Cal}} \mathbb{1}\left\{S^{(n+1),k} > S^{(k),n+1}\right\}\right\},
\end{aligned}
$$

where we defined $S^{(i),j} := s\left(\left(X^{(i)}, \max\left(M^{(i)}, M^{(j)}\right)\right), Y^{(i)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right)$, that is the score of the point i when the mask of the point j is applied to it, on top of its own mask $M^{(i)}$.

Step 2. Define the comparison matrix $\mathcal{C} \in \{0,1\}^{(\#\text{Cal}+1)\times(\#\text{Cal}+1)}$, s.t. for $(i,j) \in (\text{Cal} \cup \{n+1\})^2$: $\mathcal{C}_{i,j} = \mathbb{1}\left\{S^{(i),j} > S^{(j),i}\right\}$. Hence, we now have (since by definition $\mathcal{C}_{n+1,n+1} = 0$):

$$\left\{Y^{(n+1)} \notin \widehat{C}^{\text{MDA-Nested}^\star}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right)\right\} = \left\{\sum_{k\in\text{Cal}\cup\{n+1\}} \mathcal{C}_{n+1,k} \geq (1-\alpha)(\#\text{Cal}+1)\right\}.$$

Denote $W(\mathcal{C}) = \{i \in \text{Cal}\cup\{n+1\} : \sum_{k\in\text{Cal}\cup\{n+1\}} \mathcal{C}_{i,k} \geq (1-\alpha)(\#\text{Cal}+1)\}$. We can re-write:

$$\left\{Y^{(n+1)} \notin \widehat{C}^{\text{MDA-Nested}^\star}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right)\right\} = \{n+1 \in W(\mathcal{C})\}.$$

Therefore $\mathbb{P}\left\{Y^{(n+1)} \notin \widehat{C}^{\text{MDA-Nested}^\star}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right)\right\} = \mathbb{P}\{n+1 \in W(\mathcal{C})\}$. Thus, we will now bound $\mathbb{P}\{n+1 \in W(\mathcal{C})\}$.

Again, $\#W(\mathcal{C}) \leq 2\alpha(\#\text{Cal}+1)$ deterministically (Barber et al., 2021b).

Step 3. To conclude the proof, observe that the matrix $\mathcal{C}$ can be viewed as the output of a deterministic function $\mathcal{C}$ of the exchangeable (by A2) sequence $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$: $\mathcal{C} = \mathcal{C}\left(\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}\right)$.

Thus, for any permutation σ on $\text{Cal}\cup\{n+1\}$, it holds:

$$\mathcal{C}\left(\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k\in\text{Cal}\cup\{n+1\}}\right) \overset{d}{=} \mathcal{C}\left(\left(X^{(\sigma(k))}, M^{(\sigma(k))}, Y^{(\sigma(k))}\right)_{k\in\text{Cal}\cup\{n+1\}}\right) := \mathcal{C}^\sigma.$$

It follows that for any $k \in \text{Cal}\cup\{n+1\}$, $\mathbb{P}\{k \in W(\mathcal{C})\} = \mathbb{P}\{k \in W(\mathcal{C}^\sigma)\}$ for any permutation σ on $\text{Cal}\cup\{n+1\}$. Therefore $\mathbb{P}\{k \in W(\mathcal{C})\}$ does not depend on k. Finally:

$$\begin{aligned}
\mathbb{P}\left\{Y^{(n+1)} \notin \widehat{C}^{\text{MDA-Nested}^\star}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right)\right\} &= \mathbb{P}\{n+1 \in W(\mathcal{C})\} \\
&= \frac{1}{\#\text{Cal}+1}\sum_{k\text{Cal}\cup\{n+1\}} \mathbb{P}\{k \in W(\mathcal{C})\} \\
&= \frac{1}{\#\text{Cal}+1}\mathbb{E}[\,\#W(\mathcal{C})] \\
&\leq \frac{1}{\#\text{Cal}+1} 2\alpha(\#\text{Cal}+1) = 2\alpha.
\end{aligned}$$

□

D.2 MCV of `CP-MDA-Nested`*

To prove that `CP-MDA-Nested`* and `CP-MDA-Nested` are MCV-$\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$, we leverage again the parallel with leave-one-out conformal predictors, but this time seeing the missing pattern as exogenous randomness, which is possible when working with distributions in $\mathcal{P}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$.

Proof. Under $\mathcal{P}^{\otimes(n+1)}_{\text{MCAR,Y}\perp\!\!\!\perp\text{M}\,|\text{X}}$, it holds that $M^{(n+1)} \perp\!\!\!\perp \left(\left(X^{(k)}, Y^{(k)}\right)_{k\in\text{Cal}}, \left(X^{(n+1)}, Y^{(n+1)}\right)\right)$. Thus the sequence $\left\{\left(X^{(k)}, M^{(n+1)}, Y^{(k)}\right)_{k\in\text{Cal}}, \left(X^{(n+1)}, M^{(n+1)}, Y^{(n+1)}\right)\right\}$ is exchangeable conditionally to $M^{(n+1)}$.

Remark now that for any $(X, M, Y) \in \mathcal{X} \times \mathcal{M} \times \mathcal{Y}$, we can rewrite the score on this point with augmented mask $\widetilde{M} := \max\left(M, M^{(n+1)}\right)$ as:

$$s\left(\left(X, \widetilde{M}\right), Y; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right) := s\left(\left(X, M^{(n+1)}\right), Y; \widetilde{A}\left(\widetilde{\Phi}\left(\cdot,\cdot;M\right),\cdot;M\right)\right),$$

where, for an additional mask $M' \in \mathcal{M}$, $\widetilde{\Phi}\left(X, M; M'\right) := \Phi\left(X, \max\left(M, M'\right)\right)$ and similarly $\widetilde{A}\left(X, M; M'\right) := \hat{A}\left(X, \max\left(M, M'\right)\right)$.

Thus, we can re-write `CP-MDA-Nested`* as:

$$\begin{aligned}
&\left\{Y^{(n+1)} \notin \widehat{C}^{\text{MDA-Nested}^\star}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right)\right\} \\
=&\Bigg\{(1-\alpha)(\#\text{Cal}+1) \\
&\quad \leq \sum_{k\in\text{Cal}} \mathbb{1}\Big\{s\left(\left(X^{(n+1)}, \widetilde{M}^{(k)}\right), Y^{(n+1)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right) \\
&\qquad\qquad > s\left(\left(X^{(k)}, \widetilde{M}^{(k)}\right), Y^{(k)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right)\Big\}\Bigg\} \\
=&\Bigg\{(1-\alpha)(\#\text{Cal}+1) \\
&\quad \leq \sum_{k\in\text{Cal}} \mathbb{1}\Big\{s\left(\left(X^{(n+1)}, M^{(n+1)}\right), Y^{(n+1)}; \widetilde{A}\left(\widetilde{\Phi}\left(\cdot,\cdot;M^{(k)}\right),\cdot;M^{(k)}\right)\right) \\
&\qquad\qquad > s\left(\left(X^{(k)}, M^{(n+1)}\right), Y^{(k)}; \widetilde{A}\left(\widetilde{\Phi}\left(\cdot,\cdot;M^{(k)}\right),\cdot;M^{(k)}\right)\right)\Big\}\Bigg\}.
\end{aligned}$$

Therefore, an equivalent rewriting of `CP-MDA-Nested`* is a specific instance of what is presented in Algorithm 3, where the differences with `CP-MDA-Nested`* (Algorithm 1) are highlighted through green text.

Algorithm 3 MDA based on random masks

Input: Imputation function Φ, fitted predictor $\hat{A}$, conformity score function $s\left(\cdot,\cdot;f\right)$ for $f \in \mathcal{F} := \mathcal{Y}^{\mathcal{X}\times\mathcal{M}}$, level α, calibration set $\left\{\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)\right\}_{k\in\widetilde{\text{Cal}}}$, test point $\left(X^{(n+1)}, M^{(n+1)}\right)$

Output: Prediction set $\widehat{C}^{\text{MDA-RandomMask}}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right)$

Define $\mathcal{G}\left(\nu\right) := \widetilde{A}\left(\widetilde{\Phi}\left(\cdot,\cdot;\nu\right);\nu\right)$ for some $\nu \in \mathcal{M}$

for $k \in \widetilde{\text{Cal}}$ **do** Additional nested masking

  Randomly draw ν_k, independently from $\left(X^{(k)}, Y^{(k)}, X^{(n+1)}, Y^{(n+1)}\right)$

  Fit $\hat{g}_k := \mathcal{G}\left(\nu_k\right) = \widetilde{A}\left(\widetilde{\Phi}\left(\cdot,\cdot;\nu_k\right);\nu_k\right)$

end for

$$\begin{aligned}
&\widehat{C}^{\text{MDA-RandomMask}}_{n,\alpha}\left(X^{(n+1)}, M^{(n+1)}\right) \\
:=&\Bigg\{y \in \mathcal{Y} : (1-\alpha)(1+\#\text{Cal}) > \\
&\qquad \sum_{k\in\widetilde{\text{Cal}}} \mathbb{1}\left\{s\left(\left(X^{(k)}, M^{(k)}\right), Y^{(k)}; \hat{g}_k\right) < s\left(\left(X^{(n+1)}, M^{(n+1)}\right), y; \hat{g}_k\right)\right\}\Bigg\}
\end{aligned}$$

Indeed, conditionally on $M^{(n+1)}$, we can apply Algorithm 3 to the modified data set $\left(X^{(k)}, M^{(n+1)}, Y^{(k)}\right)_{k\in\widetilde{\mathrm{Cal}}}$, by using the $\left(M^{(k)}\right)_{k\in\widetilde{\mathrm{Cal}}}$ as random draw for $(\nu_k)_{k\in\widetilde{\mathrm{Cal}}}$ in line 3. This is legit only when the distribution of $\left(X^{(k)}, M^{(n+1)}, Y^{(k)}\right)_{k\in\widetilde{\mathrm{Cal}}\cup\{n+1\}}$ belongs to $\mathcal{P}^{\otimes(\#\widetilde{\mathrm{Cal}}+1)}_{\mathrm{MCAR,YIM\,|X}}$, as then for any $k \in \widetilde{\mathrm{Cal}}$, it holds that $M^{(k)} \perp\!\!\!\perp \left(X^{(k)}, Y^{(k)}, X^{(n+1)}, Y^{(n+1)}\right)$.

This Algorithm 3 is a special case of leave-one-out CP presented in Appendix C, with a randomized algorithm that only returns a pre-determined function associated with a parameter value, without fitting anything on the $n-1$ data points. Therefore, the validity result of leave-one-out CP extends to Algorithm 3.

In particular, under $\mathcal{P}^{\otimes(n+1)}_{\mathrm{MCAR,YIM\,|X}}$, `CP-MDA-Nested`* corresponds to applying Algorithm 3 to the data set $\left(X^{(k)}, M^{(n+1)}, Y^{(k)}\right)_{k\in\mathrm{Cal}}$ which is exchangeable conditionally on $M^{(n+1)}$, and by using in line 3 the $\left(M^{(k)}\right)_{k\in\mathrm{Cal}}$ as random draw for $(\nu_k)_{k\in\mathrm{Cal}}$. Therefore, `CP-MDA-Nested`* is MCV-$\mathcal{P}^{\otimes(n+1)}_{\mathrm{MCAR,YIM\,|X}}$ at the level $1-2\alpha$. □

The idea in this re-writing is to see that, conditionally on $M^{(n+1)}$, `CP-MDA-Nested`* predicting on the test point $\left(X^{(n+1)}, M^{(n+1)}\right)$ given the data set $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$, is in fact another run of `CP-MDA-Nested`* which predicts on a complete test point $\breve{X}^{(n+1)} \in \breve{\mathcal{X}}$, where $\breve{\mathcal{X}}$ is the set of dimension $|\mathrm{obs}\left(M^{(n+1)}\right)|$ containing only the observed dimensions of $\mathcal{X}$ according to $M^{(n+1)}$, given the cropped data set $\left(\breve{X}^{(k)}, \breve{M}^{(k)}, Y^{(k)}\right)_{k=1}^{n}$, with $\breve{M}^{(k)} \in \breve{\mathcal{M}}$ that, similarly to $\breve{X}$, is the set of dimension $|\mathrm{obs}\left(M^{(n+1)}\right)|$ containing only the observed dimensions of $\mathcal{M}$ according to $M^{(n+1)}$.

References

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591.

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. (2022). Near-optimal rate of consistency for linear models with missing values. In *Proceedings of the 39th International Conference on Machine Learning*.

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. (2023). Naive imputation implicitly regularizes high-dimensional linear models. In *Proceedings of the 40th International Conference on Machine Learning*.

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. (2024). Random features models: a way to study the success of naive imputation.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021b). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507.

Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees.

Guan, L. (2022). Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50.

Gui, Y., Barber, R. F., and Ma, C. (2023). Conformalized matrix completion.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496.

Josse, J., Chen, J. M., Prost, N., Scornet, E., and Varoquaux, G. (2024). On the consistency of supervised learning with missing values.

Josse, J. and Reiter, J. P. (2018). Introduction to the Special Section on Missing Data. *Statistical Science*, 33(2):139 – 141.

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2023). Batch multivalid conformal prediction. In *International Conference on Learning Representations*.

Kim, B. and Barber, R. F. (2023). Black-box tests for algorithmic stability. *Information and Inference: A Journal of the IMA*, 12(4):2690–2719.

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. (2020a). Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*.

Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. (2021). What's a good imputation to predict with missing values? In *Advances in Neural Information Processing Systems*.

Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. (2020b). Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

Lee, Y., Dobriban, E., and Tchetgen, E. T. (2024). Simultaneous conformal prediction of missing outcomes with propensity score ϵ-discretization.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

Little, R. J. A. (2019). *Statistical analysis with missing data, third edition*. John Wiley & Sons.

Manokhin, V. (2022). Awesome conformal prediction.

Mayer, I., Sportisse, A., Josse, J., Tierney, N., and Vialaneix, N. (2022). R-miss-tastic: a unified platform for missing values methods and workflows. *R journal*.

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

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive Confidence Machines for Regression. In *Machine Learning: ECML 2002*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. (2020). With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2).

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*.

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

Seedat, N., Jeffares, A., Imrie, F., and van der Schaar, M. (2023). Improving adaptive conformal prediction using self-supervised learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*.

Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*.

Shao, M. and Zhang, Y. (2023). Distribution-free matrix prediction under arbitrary missing pattern.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer New York.

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

Van Ness, M., Bosschieter, T. M., Halpin-Gregorio, R., and Udell, M. (2022). The missing indicator method: From low to high dimensions.

Vovk, V. (2012). Conditional Validity of Inductive Conformal Predictors. In *Asian Conference on Machine Learning*.

Vovk, V. (2013). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1–2):9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer US.

Zaffran, M., Dieuleveut, A., Josse, J., and Romano, Y. (2023). Conformal prediction with missing values. In *Proceedings of the 40th International Conference on Machine Learning*.