Summary of *Conformal Prediction with Missing Values*

Margaux Zaffran* Aymeric Dieuleveut Julie Josse Yaniv Romano

Context

By leveraging increasingly large data sets, machine learning methods can be used to support high-stakes decision-making problems. To ensure the safe deployment of predictive models it is crucial to quantify the uncertainty of the resulting predictions, communicating the limits of predictive performance. The emergent field of conformal prediction (CP, Vovk et al., 2005) is a promising framework for distribution-free Uncertainty Quantification (UQ). CP provides controlled predictive regions for any underlying predictive algorithm (e.g., neural networks and random forests), in finite samples with no assumption on the data distribution except for the exchangeability of the train and test data. For a *miscoverage rate* $\alpha \in [0, 1]$, CP outputs a *marginally valid* prediction interval $\widehat{C}_\alpha$ for Y:

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

Split CP (Papadopoulos et al., 2002) meets Eq. (1) by keeping a hold-out set, the *calibration set*, used to evaluate the performance of a fixed predictive model.

At the same time, as the volume of data increases, the volume of missing values (`NA`) also increases. One of the most popular strategies to deal with missing values suggests imputing the missing entries with plausible values to get completed data, on which any analysis can be performed (Le Morvan et al., 2021).

CP is marginally valid with `NA`

We study CP with missing covariates. Specifically, we study downstream Quantile Regression (QR) based CP, like CQR (Romano et al., 2019), on impute-then-predict strategies. Still, the proposed approaches also encapsulate other regression basemodels, and even classification.

We show that CP on impute-then-predict is *marginally* valid regardless of the model, missingness distribution, and imputation function.

Interplay between `NA` and UQ

We describe how different masks (i.e. the set of observed covariates) introduce additional heteroskedasticity: *the predictive uncertainty strongly depends on the set of covariates observed.* We therefore focus on achieving valid coverage *conditionally on the mask*, coined MCV – Mask-Conditional-Validity. MCV is desirable in practice, as occurrence of missing values are linked to important attributes.

Traditional approaches such as QR and CQR fail to achieve MCV because they do not account for this core connection between missing values and uncertainty. Figure 1 shows on a toy example with only 3 features – thus $2^3 - 1 = 7$ possible masks – how the coverage of QR and CQR varies depending on the mask. Both methods dramatically undercover when the most important variable (X_2) is missing, and the loss of coverage worsens when additional features are missing.

*`margaux.zaffran@inria.fr`

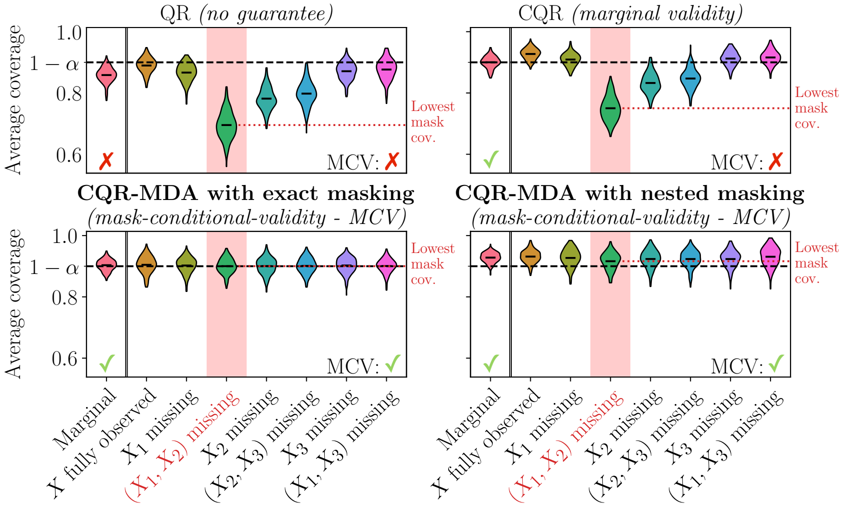

Figure 1: Coverage of the predictive intervals depending on which features are missing, among the 3 features. Evaluation over 200 runs.

A novel method: CP-MDA

We show how to form prediction intervals that are MCV, by suggesting two conformal methods sharing the same core idea of missing data augmentation (MDA): the calibration data is artificially masked to match the mask of the test point.

The first one, *CP-MDA with exact masking*, relies on building an ideal calibration set whose data points have the exact same mask as of the test point. We show its MCV under exchangeability and Missing Completely At Random.

The second one, *CP-MDA with nested masking*, does not require such an ideal calibration set. Instead, it builds a calibration set in which the data points have *at least* the same mask as the test point, i.e., this artificial masking results in calibration points having possibly more missing values than the test point. We show the latter method also achieves MCV, at the cost of an additional assumption: stochastic domination of the quantiles.

Fig. 1 illustrates CP-MDA's MCV, as their lowest mask coverage is above $1-\alpha$.

Discover more in the paper (experiments, theory, etc.)!

Paper ⟶
Poster ⟶
Code ⟶

Many thanks to Aaditya Ramdas for the idea of creating a short summary as gift.

References

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

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *ECML*.

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized quantile regression. In *NeurIPS*.

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