



HAL
open science

Uncertainty quantification in presence of missing values

Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, Yaniv Romano

► **To cite this version:**

Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, Yaniv Romano. Uncertainty quantification in presence of missing values. ICSDS 2022 - IMS International Conference on Statistics and Data Science, Dec 2022, Florence, Italy. hal-03896384v1

HAL Id: hal-03896384

<https://hal.science/hal-03896384v1>

Submitted on 13 Dec 2022 (v1), last revised 9 Nov 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Uncertainty quantification in presence of missing values

Margaux Zaffran^[1,2,3] Aymeric Dieuleveut^[3] Julie Josse^[2] Yaniv Romano^[4]

^[1]Electricité De France, Paris, France ^[2]INRIA, Montpellier, France ^[3]Ecole Polytechnique, Paris, France ^[4]Technion - Israel Institute of Technology, Haifa, Israel



Motivations and setting

Objectives

- Characterize the **impact of missing values** on **uncertainty** of the outcome.
- Propose a **methodology** outputting **predictive intervals** with **conditional coverage guarantees** with respect to **each pattern of missing values**.

• $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables.

• **Missing pattern** (mask) $M \in \{0, 1\}^d$: there are 2^d **patterns**.

$$X = (1, \text{NA}, 2) \Rightarrow M = (0, 1, 0) \text{ and } X_{\text{obs}(M)} = (1, 2).$$

• **Missing mechanism**: Missing Completely At Random (**MCAR**)

$$\text{for all } m \in \{0, 1\}^d, \mathbb{P}(M = m | X) = \mathbb{P}(M = m), \text{ i.e. } M \perp\!\!\!\perp X.$$

• **Framework**: learn Y given $X_{\text{obs}(M)}$ and M .

• Most popular strategies to deal with missing values: **imputation**.

ϕ denotes an imputation function (e.g. replaces **NA** by a constant, the empirical mean, etc).

Exchangeability after imputation

Let $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^n$ be i.i.d.. Then, for any missing mechanism, for almost all imputation function ϕ : $(\phi(X_{\text{obs}(M^{(k)})}, M^{(k)}), M^{(k)}, Y^{(k)})_{k=1}^n$ is **exchangeable**.

Infinite data

Consider **Impute-then-Regress** procedures, e.g. $g \circ \phi$. Define $g_{\delta, \phi}^* \in \operatorname{argmin}_{g: \mathbb{R}^d \rightarrow \mathbb{R}} \mathbb{E} [\rho_{\delta}(Y - g \circ \phi(X_{\text{obs}(M)}, M))]$, where ρ_{δ} is the **pinball loss** associated to the quantile of level δ .

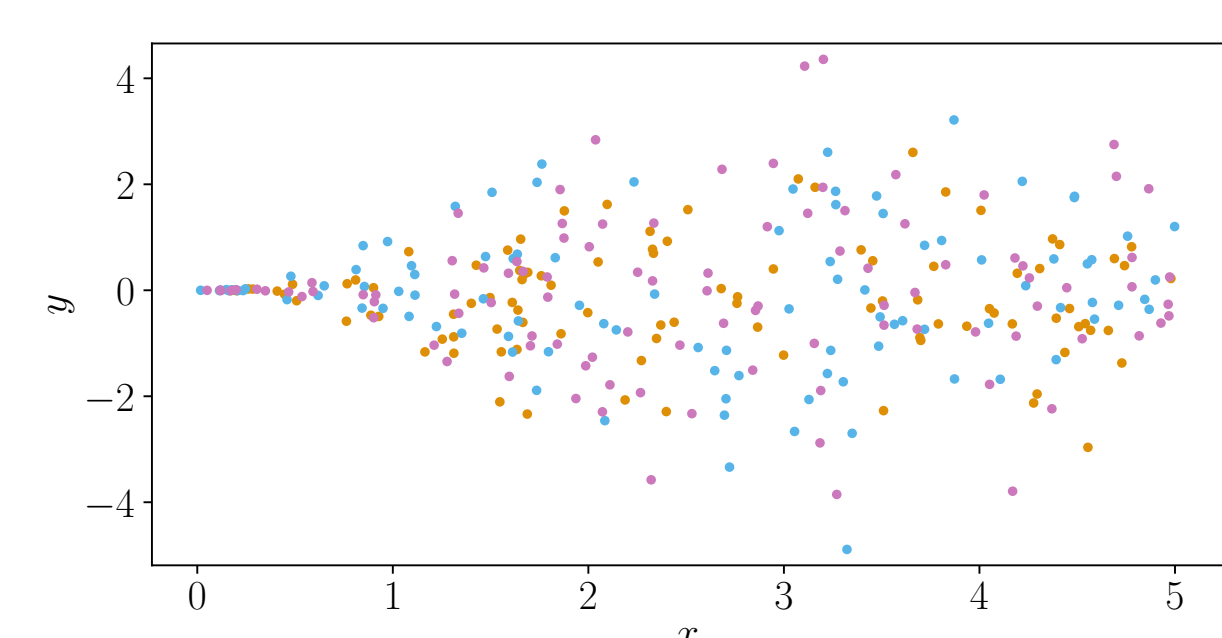
Theorem

For almost all functions $\phi \in \mathcal{F}_{\infty}^I$, $g_{\delta, \phi}^* \circ \phi$ is Bayes optimal for the pinball-risk of level δ .

A universally consistent learner trained on deterministically imputed data set will be Bayes optimal.

\Rightarrow it will reach conditional coverage with respect to the missing data pattern.

Finite sample: Conformalized Quantile Regression (CQR, Romano et al., 2019)



Randomly split the data to obtain a **proper training set** and a **calibration set**. Keep the **test set**.

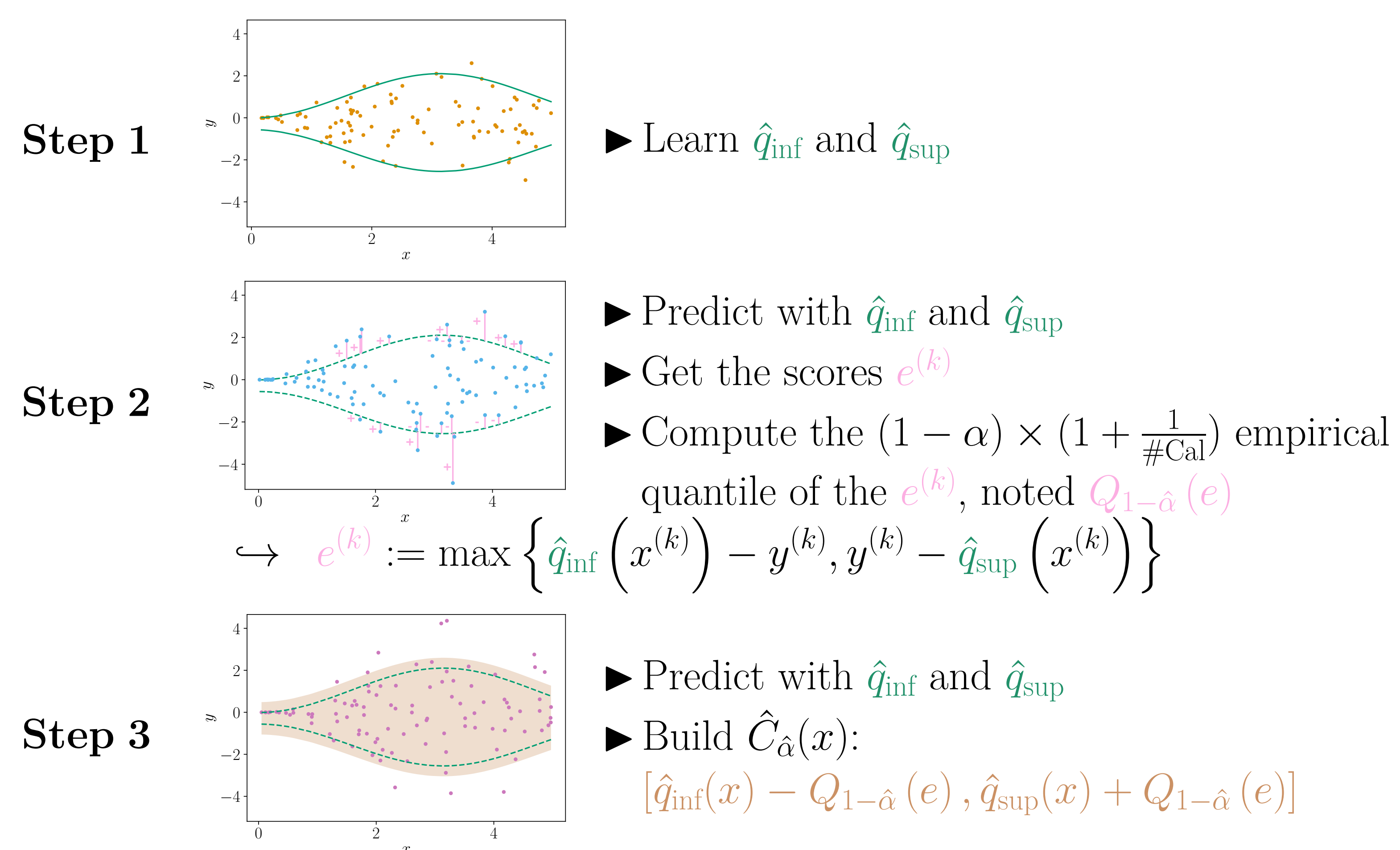
• Given any quantile regression functions \hat{q}_{inf} and \hat{q}_{sup}

• For any (**finite**) sample size n

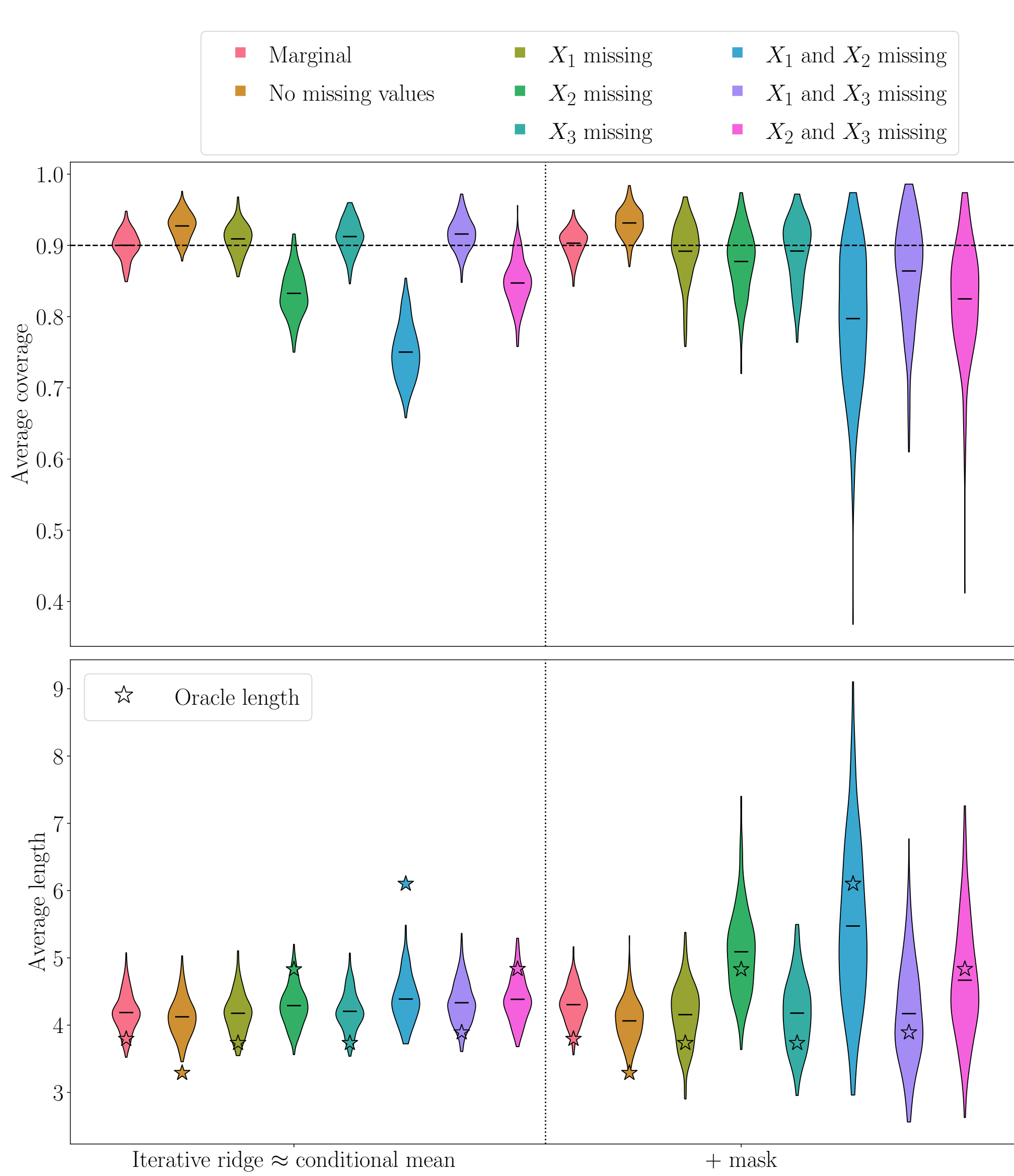
• If the $(X^{(k)}, Y^{(k)})$ are **exchangeable**

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

\Rightarrow CQR is **marginally valid** on imputed data sets.



How conditional coverage fails



$$Y = \beta^T X + \varepsilon$$

$$X \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}\right)$$

$$\beta = (1, 2, -1)^T \quad \varepsilon \sim \mathcal{N}(0, 1)$$

• M is MCAR, of probability 0.2.

• X is imputed by iterative regression.

• CQR based on neural network:

- on the imputed data set;
- on the imputed data set concatenated with the mask.

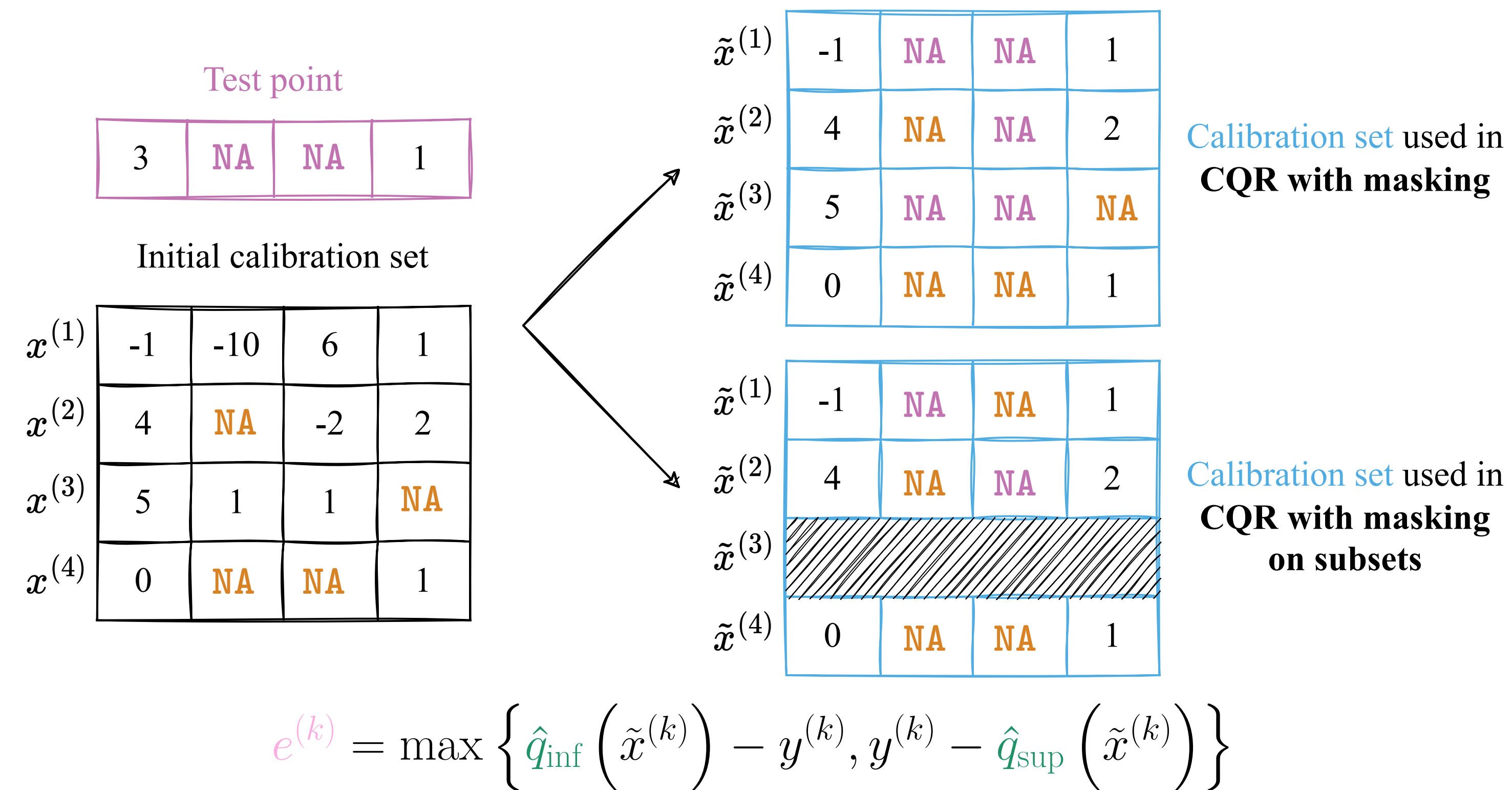
• Marginal validity is achieved.

• Not valid conditionally to the missing data pattern.

• Adding the mask improves conditionality.

Proposed algorithms

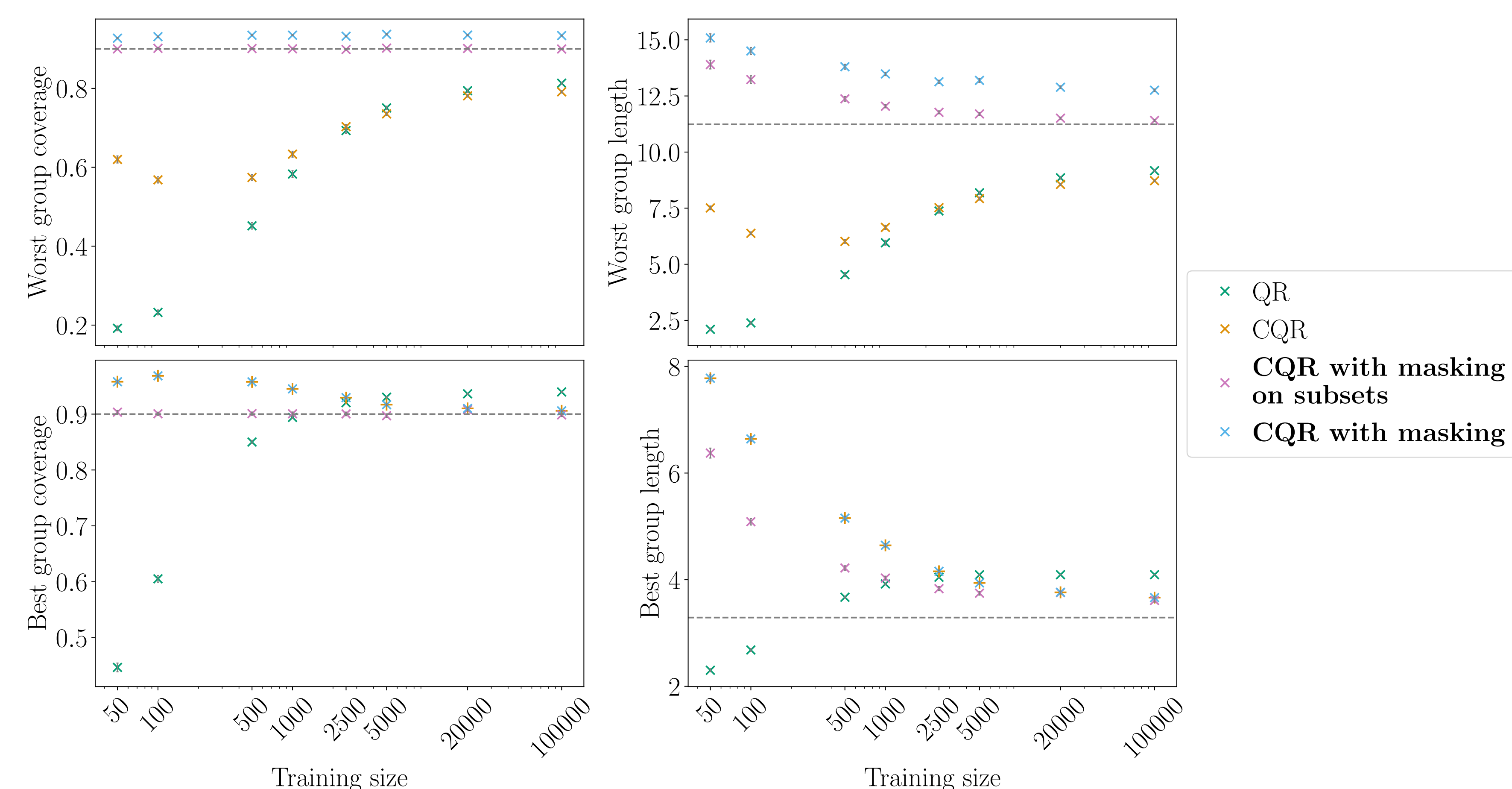
Idea: generate **additional missing values** in the **calibration set**.



$$e^{(k)} = \max \left\{ \hat{q}_{\text{inf}}(\tilde{x}^{(k)}) - y^{(k)}, y^{(k)} - \hat{q}_{\text{sup}}(\tilde{x}^{(k)}) \right\}$$

Appropriate coverage conditionally to the missing patterns

On Gaussian linear data with $d = 10$, focus on **2 extreme missing patterns**: largest and smallest number of missing values.



- As the training size increases, **QR** and **CQR** improve conditional coverage.
- CQR with masking on subsets** is not over-conservative on the easiest group, but requires more calibration data than **CQR with masking**.
- As the training size increases, **CQR with masking on subsets** \rightarrow oracle length.

Insights from the Gaussian linear model

• $Y = \beta^T X + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_{\varepsilon}^2) \perp\!\!\!\perp X$, and $\beta \in \mathbb{R}^d$.

• X conditional on M is Gaussian: for all $m \in \{0, 1\}^d$, there exist μ_m and Σ_m such that $X | (M = m) \sim \mathcal{N}(\mu_m, \Sigma_m)$.

Particular case: $X \sim \mathcal{N}(\mu, \Sigma)$, and M is MCAR. Then, $\mu_m \equiv \mu$ and $\Sigma_m \equiv \Sigma$.

Oracle intervals

Under the Gaussian linear model, for any $m \in \{0, 1\}^d$, the oracle length is given by:

$$\mathcal{L}_{\alpha}^*(m) = 2 \times q_{1-\alpha/2}^{\mathcal{N}(0,1)} \times \sqrt{\beta_{\text{mis}(m)} \Sigma_{\text{mis}(m)} \text{obs}(m) \beta_{\text{mis}(m)}^T + \sigma_{\varepsilon}^2},$$

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

- The oracle intervals depend on the regression coefficients.
- Additional **heteroskedasticity** is generated by the missing values.
- The oracle intervals **depend** on the **mask** in a **non-linear** fashion. \rightarrow even under MCAR data, it is useful to add the mask as feature.