Conformal Prediction with 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 $\mathbf{2^d}$ **patterns**.

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

- One possible 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 $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^n$ be exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi\left(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}\right), M^{(k)}, Y^{(k)}\right)_{k=1}^n$ is **exchangeable**.

Infinite data

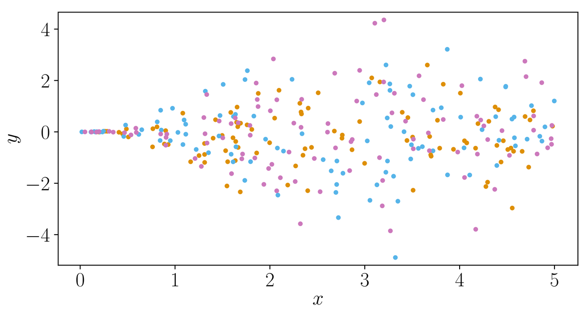
Consider **Impute-then-Regress** procedures, e.g. $g \circ \phi$. Define $g^*_{\delta,\phi} \in \underset{g:\mathbb{R}^d \to \mathbb{R}}{\text{argmin}}\ \mathbb{E}\left[\rho_\delta\left(Y - g \circ \phi(X_{\text{obs}(M)}, M)\right)\right]$, where ρ_δ is the **pinball loss** associated to the quantile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

Theorem

For almost all functions $\phi \in \mathcal{F}^I_\infty$, $g^*_{\delta,\phi} \circ \phi$ 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)



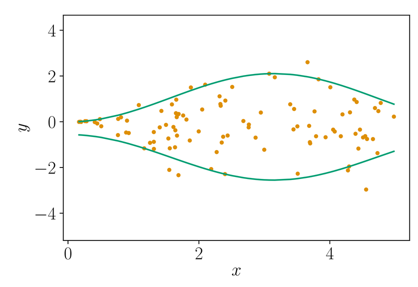
Randomly split the data to obtain a proper training set and a calibration set. Keep the test set.

- Given any quantile regression functions $\hat{q}_{\text{inf}}$ and $\hat{q}_{\text{sup}}$
- For any (**finite**) sample size n
- If the $(X^{(k)}, Y^{(k)})$ are **exchangeable**

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

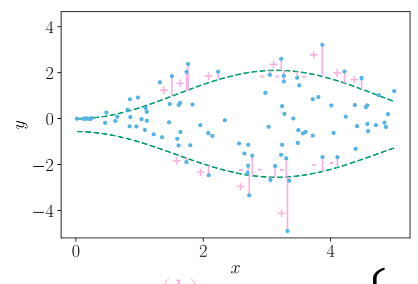
$\Rightarrow$ CQR is **marginally valid** on imputed data sets.

Step 1



▶ Learn $\hat{q}_{\text{inf}}$ and $\hat{q}_{\text{sup}}$

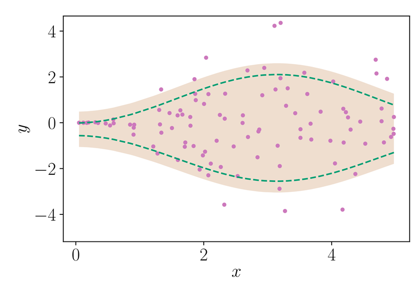
Step 2



▶ Predict with $\hat{q}_{\text{inf}}$ and $\hat{q}_{\text{sup}}$
▶ Get the scores $e^{(k)}$
▶ Compute the $(1-\alpha) \times (1 + \frac{1}{\#\text{Cal}})$ empirical quantile of the $e^{(k)}$, noted $Q_{1-\tilde{\alpha}}(e)$

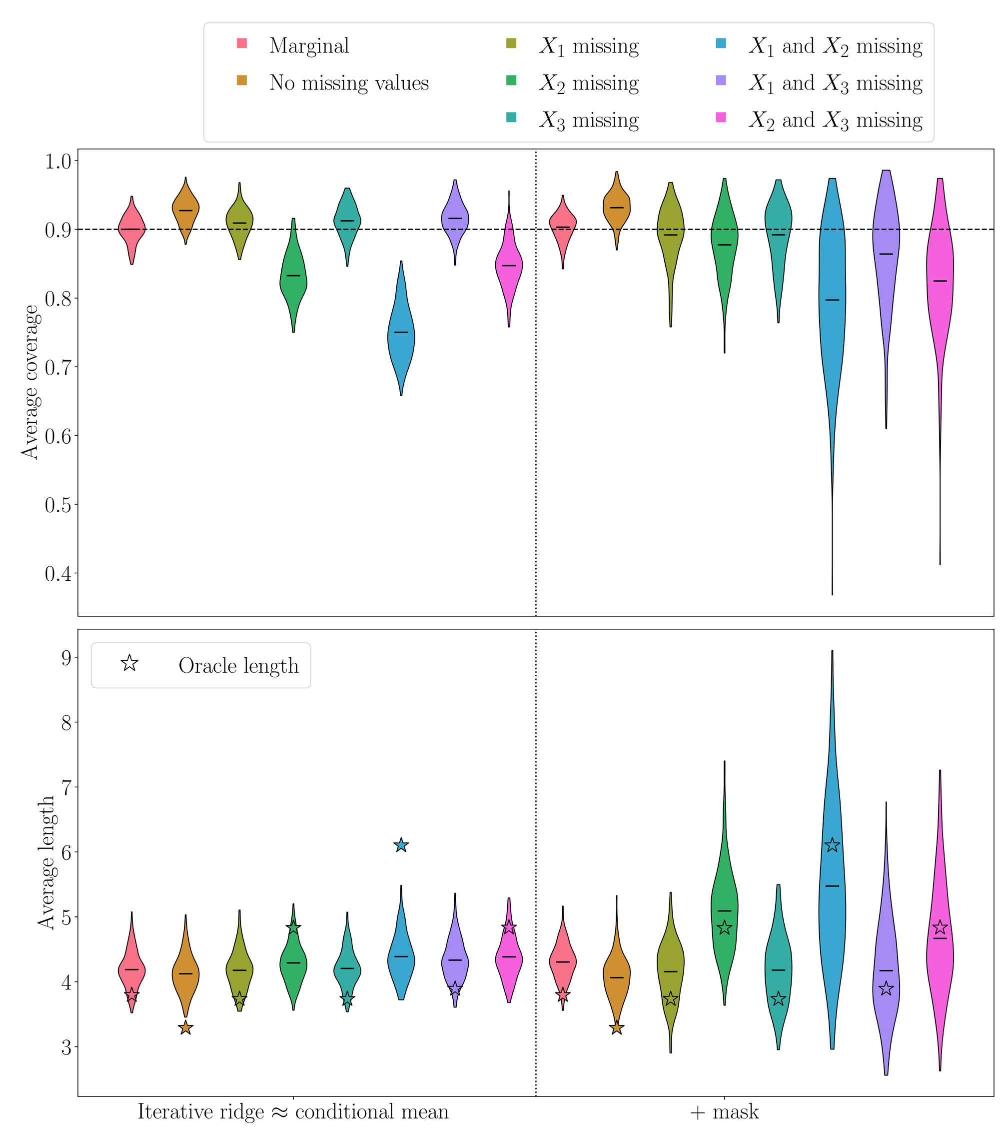
$$\hookrightarrow e^{(k)} := \max\left\{\hat{q}_{\text{inf}}\left(x^{(k)}\right) - y^{(k)}, y^{(k)} - \hat{q}_{\text{sup}}\left(x^{(k)}\right)\right\}$$

Step 3



▶ Predict with $\hat{q}_{\text{inf}}$ and $\hat{q}_{\text{sup}}$
▶ Build $\hat{C}_\alpha(x)$:
$[\hat{q}_{\text{inf}}(x) - Q_{1-\tilde{\alpha}}(e), \hat{q}_{\text{sup}}(x) + Q_{1-\tilde{\alpha}}(e)]$

How conditional coverage fails



- $Y = \beta^T X + \varepsilon$
  - ○ $X \sim \mathcal{N}\left(\begin{pmatrix}1\\1\\1\end{pmatrix}, \begin{pmatrix}1 & 0.8 & 0.8\\0.8 & 1 & 0.8\\0.8 & 0.8 & 1\end{pmatrix}\right)$
  - ○ $\beta = (1, 2, -1)^T$ ○ $\varepsilon \sim \mathcal{N}(0, 1)$
- 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.

Insights from the Gaussian linear model

- $Y = \beta^T X + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon) \perp\!\!\!\perp X$, 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:

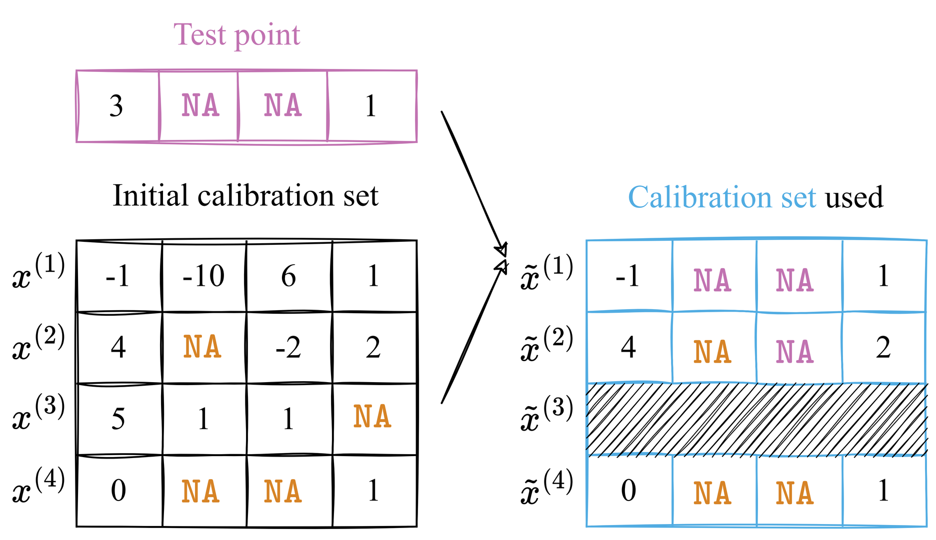
$$\mathcal{L}^*_\alpha(m) = 2 \times q^{\mathcal{N}(0,1)}_{1-\alpha/2} \times \sqrt{\beta_{\text{mis}(m)} \Sigma_{\text{mis}(m)|\text{obs}(m)} \beta^T_{\text{mis}(m)} + \sigma^2_\varepsilon},$$

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.
  $\hookrightarrow$ even under MCAR data, it is useful to add the mask as feature.

CQR-MDA-Exact: recovering mask-conditional-coverage

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



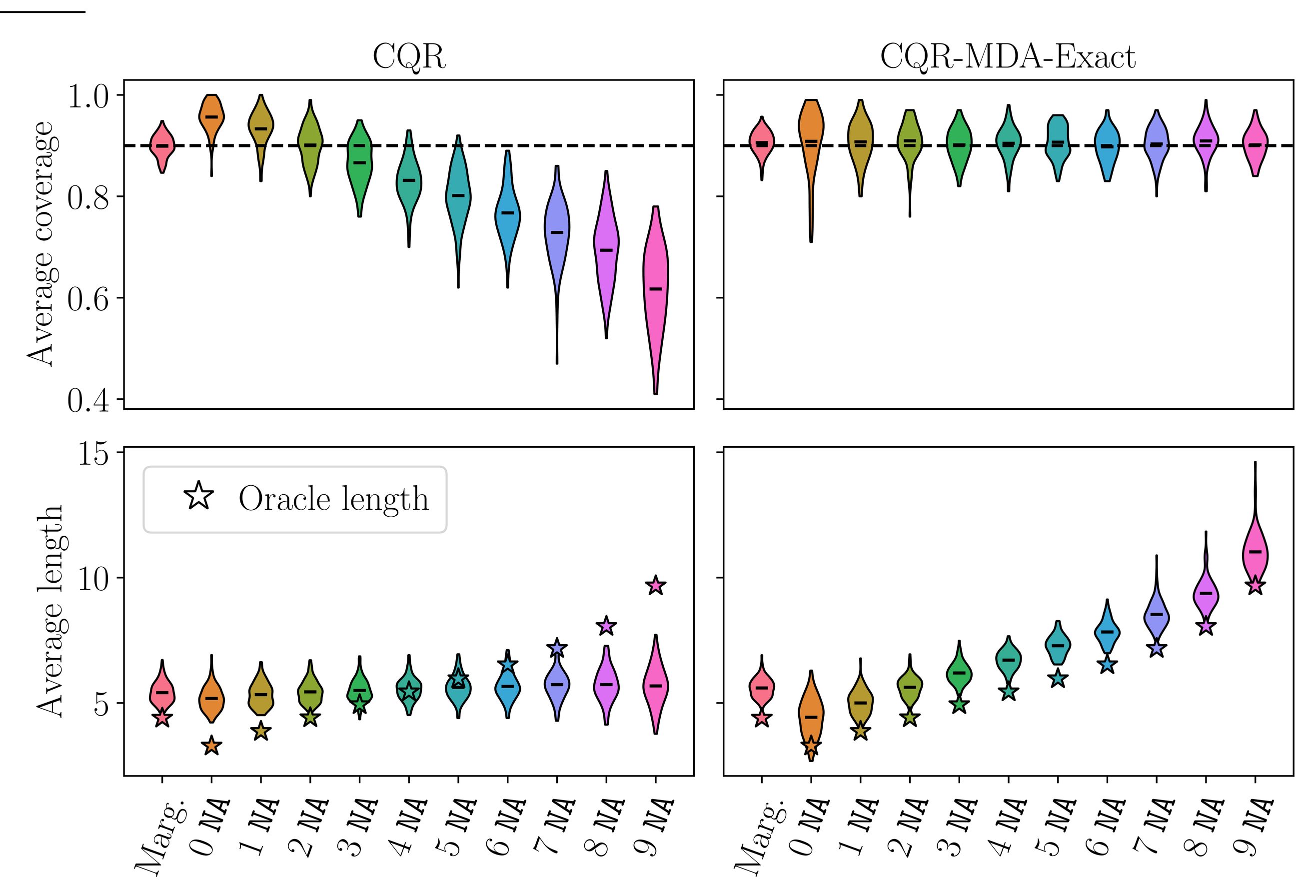
$$e^{(k)} = \max\left\{\hat{q}_{\text{inf}}\left(\tilde{x}^{(k)}\right) - y^{(k)}, y^{(k)} - \hat{q}_{\text{sup}}\left(\tilde{x}^{(k)}\right)\right\}$$

Theorem

Assume:
- ○ exchangeable data,
- ○ MCAR mechanism.

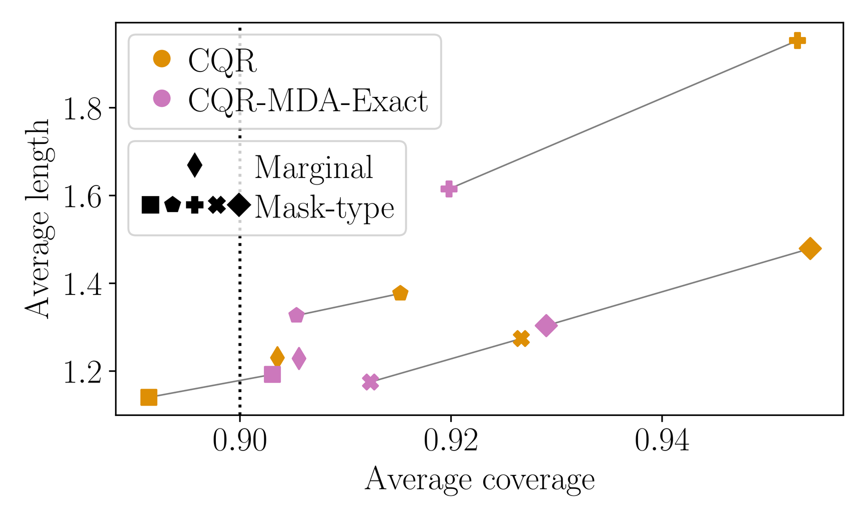
Then, CQR-MDA-Exact's predictive intervals satisfy for any $m \in \{0,1\}^d$:
$\mathbb{P}(Y \in \hat{C}_\alpha(X, M)|M = m) \geq 1 - \alpha$.

⋆ Sanity check: Gaussian linear data with $d = 10$.



TraumaBase®: critical care medicine

- Predict the levels of blood platelets upon arrival at the hospital;
- 7 explanatory variables selected by medical doctors;
- Missing values vary from 0% to 24% by features, with a total average of 7%.



Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. (2021). What's a good imputation to predict with missing values? *NeurIPS*.
Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized Quantile Regression. *NeurIPS*.

Paper ⟶
Poster ⟶
Code ⟶