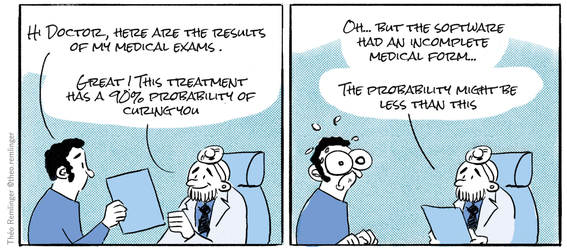
Conformal Prediction with Missing Values

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

Paper ⟶
5' talk ⟶
Code ⟶



Setting and goal

Regression with missing values: impute-then-regress

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n} \in \left(\mathbb{R}^d \times \{0,1\}^d \times \mathbb{R}\right)^n$

Y	X_1	X_2	X_3	Mask $M = (M_1$	M_2	$M_3)$
10	-1	-10	6	0	0	0
4	4	-2	NA	0	0	1
8	NA	NA	2	1	1	0

$\Longrightarrow$ $\mathbf{2^d}$ potential **masks**.

A possible missing mechanism: Missing Completely At Random (**MCAR**)

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

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

				$\xrightarrow{\phi}$			
$x^{(1)}$	-1	-10	6		-1	-10	6
$x^{(2)}$	4	-2	NA		4	-2	4
$x^{(3)}$	NA	NA	2		1.5	-6	2

Lemma: exchangeability after imputation

Let $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ be exchangeable. For any missing mechanism, for almost all ϕ: $\left(\phi\left(X^{(k)}, M^{(k)}\right), M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

Predictive uncertainty quantification with NA

Goal: predict $Y^{(n+1)}$ with confidence $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$, s.t.:

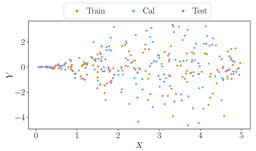
1. Marginal Validity (MV)

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

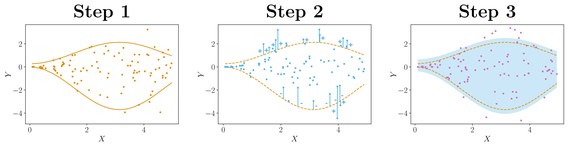
2. Mask-Conditional-Validity (MCV)

$$\forall m \in \{0,1\}^d : \mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, m\right) | M^{(n+1)} = m\right\} \geq 1-\alpha.$$

Conformalized Quantile Regression [CQR, 1]



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



▶ Learn (or get) $\widehat{\mathrm{QR}}_{\mathrm{low}}$ and $\widehat{\mathrm{QR}}_{\mathrm{up}}$.

▶ Predict with $\widehat{\mathrm{QR}}_{\mathrm{low}}$ and $\widehat{\mathrm{QR}}_{\mathrm{up}}$.

▶ Get the set of scores $\mathcal{S} = \left\{S^{(k)}\right\}_{\mathrm{Cal}} \cup \{+\infty\}$, and its empirical $1-\alpha$ quantile: $q_{1-\alpha}(\mathcal{S})$.

▶ Predict with $\widehat{\mathrm{QR}}_{\mathrm{low}}$ and $\widehat{\mathrm{QR}}_{\mathrm{up}}$.

▶ Build $\widehat{\mathcal{C}}_\alpha(x)$: $\left[\widehat{\mathrm{QR}}_{\mathrm{low}}(x) - q_{1-\alpha}(\mathcal{S}); \widehat{\mathrm{QR}}_{\mathrm{up}}(x) + q_{1-\alpha}(\mathcal{S})\right]$

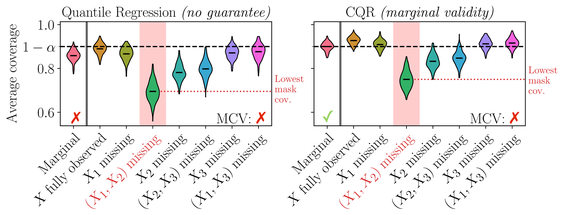
If $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are **exchangeable (or i.i.d.)**, CQR intervals achieve Marginal Validity, i.e. $\mathbb{P}\left\{Y^{(n+1)} \in \widehat{\mathcal{C}}_\alpha\left(X^{(n+1)}\right)\right\} \geq 1-\alpha$.

Impute-then-CQR

Proposition (informal)

Impute-then-CQR is Marginally Valid (MV).

Illustrative data: Gaussian linear model, X_2 has more predictive power.



Conclusion: missing values induce heteroskedasticity.

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 $X|(M = m) \sim \mathcal{N}(\mu_m, \Sigma_m)$.

Proposition: oracle intervals

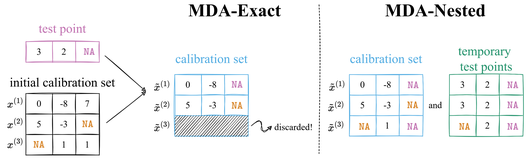
Under the Gaussian linear model, for any $m \in \{0,1\}^d$, the oracle length is:

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

The oracle intervals depend on the mask in a non-linear fashion.

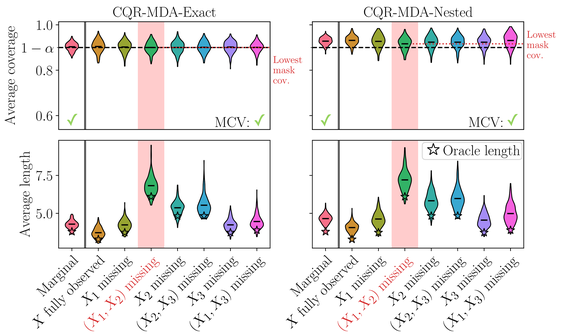
Missing Data Augmentation (MDA)

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



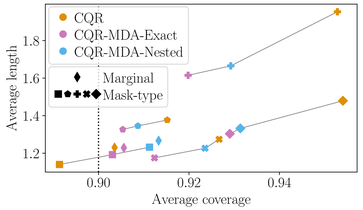
Theorem (informal): CQR-MDA is MCV

If $M \perp\!\!\!\perp (X, Y)$ and the data is exchangeable, for almost all imputation function, CQR-MDA is Mask-Conditionally-Valid (MCV).



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%.



- MDA **recovers mask-validity** on CQR's under-covered masks (■).
- MDA **improves efficiency** on over-covered masks.

Asymptotic regime

Define $g_{\delta,\phi}^\star \in \underset{g:\mathbb{R}^d \to \mathbb{R}}{\mathrm{argmin}}\ \mathbb{E}\left[\rho_\delta\left(Y - g \circ \phi(X, M)\right)\right]$, where ρ_δ is the **pinball loss** associated to the quantile of level δ.

Proposition: Bayes-consistency of impute-then-QR

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

This result is an extension of [2].

A universally consistent learner trained on deterministically imputed data set will be Bayes optimal.
$\Rightarrow$ it will reach individualized conditional coverage.

Open directions

1. Going beyond the MCAR assumption?
2. Impact of the imputation on QR with finite sample.

Main references

[1] Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized Quantile Regression. *NeurIPS*, 2019.

[2] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gael Varoquaux. What's a good imputation to predict with missing values? *NeurIPS*, 2021.